Electronic Journal of Statistics
Vol. 14 (2020) 1302–1347
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1686

Rate optimal Chernoff bound and application to community detection in the stochastic block models

Zhixin Zhou and Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th St, Bellevue, WA 98004, USA
e-mail: zhixin0825@gmail.com; pingli98@gmail.com

Abstract: The Chernoff coefficient is known to be an upper bound of Bayes error probability in classification problem. In this paper, we will develop a rate optimal Chernoff bound on the Bayes error probability. The new bound is not only an upper bound but also a lower bound of Bayes error probability up to a constant factor. Moreover, we will apply this result to community detection in the stochastic block models. As a clustering problem, the optimal misclassification rate of community detection problem can be characterized by our rate optimal Chernoff bound. This can be formalized by deriving a minimax error rate over certain parameter space of stochastic block models, then achieving such an error rate by a feasible algorithm employing multiple steps of EM type updates.

MSC 2010 subject classifications: Primary 62F03; secondary 60G05.
Keywords and phrases: Chernoff information, Bayes error probability, hypothesis testing, community detection, stochastic block models.

Received June 2019.

1. Introduction

Many classification and clustering problems in statistical literature can be reduced to *symmetric hypothesis testing*. In a classical setting, given two hypotheses H_0 and H_1, where H_i assumes that observing data from a measurable space with distribution P_i, one discriminates between them according to certain decision rule. Type-I error occurs if one accepts H_0 while the data are generated from distribution P_1, and vice versa on type-II error. Symmetric hypothesis testing indicates that the hypotheses are equiprobable, and the loss function weighs type-I error and type-II error equally. Therefore, we would like to focus on the *Bayes error probability*, which averages two kinds of error probabilities.

The asymptotic behavior of the Bayes error probability becomes an essential problem in symmetric hypothesis testing. Given probability density functions (PDFs) or probability mass functions (PMFs) φ_0 and φ_1 of distribution P_0 and P_1 respectively, the *Chernoff information*, defined as

$$D_{\alpha^*}(\varphi_0\|\varphi_1) = -\inf_{\alpha\in(0,1)} \log \int \varphi_0^{1-\alpha}\varphi_1^{\alpha} d\mu \tag{1.1}$$

is known as the best exponent of Bayes error probability [1]. A Chernoff type lower bound was investigated in [2, 3]. It is a powerful tool in many researches, such as community detection [4, 5, 6] and quantum information theory [7, 8]. However, the ratio between *Chernoff coefficient*, defined as $\exp(-D_{\alpha^*}(\varphi_0\|\varphi_1))$, and the Bayes error probability has not been investigated in previous literature. In this paper, we will propose a rate optimal Chernoff bound for Bayes error probability. Observing i.i.d. samples with distribution either φ_0 or φ_1, we will show that the Bayes error probability is asymptotically equivalent to

$$\frac{1}{\sqrt{n}\alpha^*(1-\alpha^*)}e^{-nD_{\alpha^*}(\varphi_0\|\varphi_1)} \tag{1.2}$$

up to a constant factor. This result can be generalized to the non-i.i.d. case. Although a comparable second-order asymptotics for *asymmetric hypothesis testing* was investigated in [9, 10], there is no direct application to our situation.

This paper will also apply the main result of the rate optimal Chernoff upper and lower bound to one of popular clustering problems in statistics, namely community detection. Particularly, we will focus on the *stochastic block models* (SBM). Many effective algorithms and related theories have been proposed for solving community detection in SBMs, including global approaches such as spectral clustering [11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23] and convex relaxations via semidefinite programs (SDPs) [24, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Global approaches usually involve a single optimization step (either spectral clustering or SDP after convex relaxation) and do not require good initialization. However, these algorithms are usually not optimal on their own, because both spectral clustering and SDP lose the block structure in SBM. The pseudo-likelihood approach [34] fills in the gap with local refinement and makes optimal clustering possible. The general idea was concluded as "Good Initialization followed by Fast Local Updates" (GI-FLU) by [35]. Since the minimax error rate proposed in [36], algorithms in the manner of GI-FLU are developed in [37, 35, 4, 6]. However, as the rate optimal Chernoff upper and lower bounds were not used in these papers, the minimax rate is not sufficiently accurate and very few algorithm have been proved to be optimal. Details can be found in the following table. Here, n is the number of nodes and d is the average degree of a node in the network. K denotes number of communities. I indicates Chernoff information in (1.1). The specific one for community detection will be defined in (3.7). $o(1)$ is some eventually positive sequences converging to 0.

TABLE 1
Comparison with existing results.

paper	density	symmetry	minimax error	algorithmic error
[37]	not needed	yes	not derived	$\exp(-CnI)$, $(C<1)$
[4]	$\Theta(\log n)$	yes	not derived	$o(1/n)$
[35]	not needed	yes	$\exp(-(1+o(1))nI)$	$\exp(-(1-o(1))nI)$
[6]	$O(\sqrt{n})$	no	$\Omega(\exp(-nI)/d^{K/2})$	$O(\exp(-nI)/\sqrt{d})$
This paper	not needed	yes	$\Omega(\exp(-nI)/\sqrt{d})$	$O(\exp(-nI)/\sqrt{d})$

Some features or assumptions of the problem are described as follows. *Density* indicates the average degree of a node. *Symmetry* means the paper assumes that the network is an undirected graph. Community detection on symmetric network is usually more difficult since half edges are duplicated. *Minimax error rate* can be considered as fundamental limit of community detection problem. *Algorithmic error rate* are the theoretical guarantees of feasible algorithms.

Block partitioning skills introduced in [37] generate enough independence between different steps of their algorithm. However, the last local update can only be applied on half of dataset, so the error rate is much higher than $\exp(-nI)$. Algorithm derived in [35] has error rate similar to the minimax error rate in [36], but the ratio between upper and lower bound has order $\exp(o(1)nI)$, which can be arbitrary divergent sequence. The analysis in [4] focuses on the density regime $\Theta(\log n)$, but it cannot generalize to other densities. To achieve an optimal error rate, the algorithm in [6] allows twice local update. However, their approach cannot extend to undirected network. We will combine different existing techniques and propose a new algorithm that achieves the minimax error rate (up to a constant).

We summarize the contributions of this paper as follows:

1. We investigate the rate optimal Chernoff upper and lower bound for Bayes error probability.
2. Considering certain parameter space, we propose the second-order asymptotics for minimax lower bound for community detection in SBM.
3. We provide a feasible algorithm which guarantees to achieve the minimax lower bound up to a constant factor.

The rest of the paper will be organized as follows. We introduce the Chernoff type upper and lower bound in Section 2, then we present our minimax lower bounds and the provable community detection algorithm with its analysis in Section 3. Simulations will appear in Section 5. Proofs of Theorems and corollaries in Section 2 will appear in Section 6. Proofs about minimax error rate and consistency of community detection can be found in Section 7.

Here, we briefly introduce the notations will be used in this paper. $[n]$ is the set of integers from 1 to n, i.e., $[n] = \{1, 2, \ldots, n\}$. A random variable $X \sim f$ means X has probability mass function or density function f. $a_n \lesssim b_n$ or equivalently $a_n = O(b_n)$ holds if there exists a constant C such that $a_n \leq Cb_n$ for sufficiently large n. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$, then $a_n \asymp b_n$. $a_n = o(1)$ means a_n converges to 0 and nonnegative for sufficiently large n. Furthermore, we use $a \vee b$ and $a \wedge b$ to denote $\max(a, b)$ and $\min(a, b)$ respectively.

2. Rate optimal Chernoff upper and lower bounds

We will introduce a fundamental testing problem under a Bayes setting, then present a new Chernoff type upper and lower bound of Bayes error probability. We will also introduce its application exponential families.

2.1. Symmetric hypothesis testing

We will define a symmetric hypothesis testing problem and its Bayes error probability. Let φ_{0j} and φ_{1j} for $j \in [n]$ be two sequences of measurable PDFs for one-dimensional real random variables. Same results hold if they are PMFs, but we only consider PDFs for brevity. We assume for every $j \in [n]$, φ_{0j} and φ_{1j} are defined on the same measure space $(\Omega_j, \Sigma_j, \mu_j)$. Let us consider the product measure space (Ω, Σ, μ) defined by

$$\Omega := \Omega_1 \times \cdots \times \Omega_n, \quad \Sigma = \Sigma_1 \otimes \cdots \otimes \Sigma_n \quad \text{and} \quad \mu = \mu_1 \times \cdots \times \mu_n \tag{2.1}$$

and measurable PDFs

$$\varphi_z(x) := \varphi_z(x_1, \ldots, x_n) := \prod_{j=1}^{n} \varphi_{zj}(x_j) \text{ for } z \in \{0,1\}. \tag{2.2}$$

Furthermore, we denote the *Kullback–Leibler divergence* of φ_1 from φ_0 by

$$D_{\mathrm{KL}}(\varphi_0 \| \varphi_1) = \int_\Omega \varphi_0 \log \frac{\varphi_0}{\varphi_1} d\mu.$$

We assume both $D_{\mathrm{KL}}(\varphi_0\|\varphi_1)$ and $D_{\mathrm{KL}}(\varphi_1\|\varphi_0)$ exist, which implies $\int_\Omega (\varphi_0 + \varphi_1)\left|\log \frac{\varphi_1}{\varphi_0}\right| d\mu < \infty$. In particular, it requires φ_0 and φ_1 to have the same support, and take different values on a set with non-zero measure. For a pair of density functions satisfying these conditions, we say

$$(\varphi_0, \varphi_1) \in \mathcal{F}(\Omega, \Sigma, \mu, n). \tag{2.3}$$

Now we randomly draw a number $z \in \{0,1\}$ with equal prior probability $1/2$. We note that the following arguments about the rate optimal Chernoff bound still hold as long as the prior probability is nondegenerate, but we assume equiprobable for simplicity. Then we draw a random sample $X = (X_1, \ldots, X_n)$ where $X_j \sim \varphi_{zj}$ independently. We aim to recover the label z given the observation $X = x = (x_1, \ldots, x_n)$. For any estimator $\hat{z} := \hat{z}(x)$ of z, we define the *Bayes error probability*, also called *Bayes risk*, given by

$$R(\hat{z}, z) := \frac{1}{2} \sum_{z \in \{0,1\}} \mathbb{P}(\hat{z} \neq z)$$

Due to the Neyman-Pearson lemma, the Bayes estimator, which is known to be the best estimator in the measurement of Bayes risk, is given by

$$\hat{z} := \arg \max_{z \in \{0,1\}} \varphi_z(x).$$

In the rest of this section, we let $\hat{z}$ be the Bayes estimator. The Bayes error probability is closely related to *total variation affinity* between φ_0 and φ_1, denoted as $\eta(\varphi_0, \varphi_1)$, which will be defined as follows:

$$\begin{aligned} \eta(\varphi_0, \varphi_1) &:= \int_\Omega \min(\varphi_0, \varphi_1) d\mu \\ &= \int_\Omega \varphi_0 1\{\varphi_0 \leq \varphi_1\} d\mu + \int_\Omega \varphi_1 1\{\varphi_1 < \varphi_0\} d\mu = 2R(\hat{z}, z). \end{aligned} \tag{2.4}$$

The naming of total variation affinity comes from the fact that

$$\eta(\varphi_0, \varphi_1) = 1 - D_{\mathrm{TV}}(\varphi_0 \| \varphi_1), \quad \text{where} \quad D_{\mathrm{TV}}(\varphi_0 \| \varphi_1) := \sup_{A \in \Sigma} \Big| \int_A \varphi_0 - \varphi_1 d\mu \Big|.$$

Here, $D_{\mathrm{TV}}(\varphi_0 \| \varphi_1)$ is the total variation distance. Now we can focus on the total variation affinity and express it as

$$\eta(\varphi_0, \varphi_1) = \int_\Omega \min(\varphi_0, \varphi_1) d\mu = \int_\Omega \varphi_0^{1-\alpha} \varphi_1^{\alpha} \min(l^\alpha, l^{\alpha-1}) d\mu, \tag{2.5}$$

where $l = \varphi_0/\varphi_1$ is the likelihood ratio defined pointwisely on Ω. This ratio is well defined since we assume that φ_0 and φ_1 have the same support. We observe that $\varphi_0^{1-\alpha} \varphi_1^{\alpha}$ is a PDF on Ω up to a normalizer and $\min(l^\alpha, l^{\alpha-1})$ is a real valued function on Ω, so it would be convenient to express $\eta(\varphi_0, \varphi_1)$ as an expectation. For $\alpha \in (0, 1)$, we define PDF

$$\begin{gathered} \varphi_\alpha(x) := \varphi_0(x)^{1-\alpha} \varphi_1(x)^\alpha e^{D_\alpha(\varphi_0 \| \varphi_1)}, \\ \text{where} \quad D_\alpha(\varphi_0 \| \varphi_1) := -\log \int_\Omega \varphi_0^{1-\alpha} \varphi_1^\alpha d\mu. \end{gathered} \tag{2.6}$$

We call $D_\alpha(\varphi_0 \| \varphi_1)$ the *Chernoff α-divergence* between φ_0 and φ_1. We also define a real valued function, which will play an important role in the analysis of the higher-order term:

$$g_\alpha : \mathbb{R} \to \mathbb{R}, \quad g_\alpha(x) := \exp[\min(\alpha x, (\alpha - 1)x)] = \min(e^{\alpha x}, e^{(\alpha-1)x}). \tag{2.7}$$

Then by direct calculation from (2.5), we have

$$\begin{aligned} \eta(\varphi_0, \varphi_1) &= e^{-D_\alpha(\varphi_0 \| \varphi_1)} \int_\Omega \varphi_\alpha \min(l^\alpha, l^{\alpha-1}) d\mu \\ &= e^{-D_\alpha(\varphi_0 \| \varphi_1)} \mathbb{E}_{Y \sim \varphi_\alpha}[g_\alpha(\log l(Y))]. \end{aligned} \tag{2.8}$$

We note that since $g_\alpha(x) \leq 1$, we always have $\mathbb{E}_{Y \sim \varphi_\alpha}[g_\alpha(\log l(Y))] \leq 1$, which implies $e^{-D_\alpha(\varphi_0 \| \varphi_1)}$ is an upper bound of $\eta(\varphi_0, \varphi_1)$. In (2.8), $Y := (Y_1, \ldots, Y_n)$ is a random vector with independent elements on the product space Ω, and one can observe that

$$\begin{gathered} Y_j \sim \varphi_{\alpha j} := \varphi_{0j}^{1-\alpha} \varphi_{1j}^\alpha e^{D_\alpha(\varphi_{0j} \| \varphi_{1j})}, \\ \text{where} \quad D_\alpha(\varphi_{0j} \| \varphi_{1j}) := -\log \int_{\Omega_j} \varphi_{0j}^{1-\alpha} \varphi_{1j}^\alpha d\mu_j. \end{gathered} \tag{2.9}$$

Let $l_j = \varphi_{0j}/\varphi_{1j}$ and $Z_j := \log l_j(Y_j)$, then we can decompose $\log l(Y)$ as

$$\log l(Y) = \sum_{j=1}^{n} \log l_j(Y_j) = \sum_{j=1}^{n} Z_j. \tag{2.10}$$

$\log l(Y)$ is indeed the sum of independent random variables, so it is approximately normally distributed under some regularization condition, which will be specified in the following theorem. Obtaining normal approximation from the Berry-Esseen theorem (Theorem 6.1), we have the following result.

Theorem 2.1. *We consider the PDFs or PMFs $(\varphi_0, \varphi_1) \in \mathcal{F}(\Omega, \Sigma, \mu, n)$ defined in (2.3), and recall the definitions of φ_α in (2.6), g_α in (2.7), $\varphi_{\alpha j}$ in (2.9), Y_j, $l_j = \varphi_{0j}/\varphi_{1j}$ and $Z_j = \log l_j(Y_j)$. Let*

$$\alpha^* := \arg\max_{\alpha\in(0,1)} D_\alpha(\varphi_0\|\varphi_1), \quad Y_j \sim \varphi_{\alpha^* j}, \ \textit{and} \quad \bar{\sigma}_n := \Big(\frac{1}{n}\sum_{j=1}^{n} Var[Z_j]\Big)^{1/2}.$$

If $\sum_{j=1}^{n} \mathbb{E}|Z_j|^3 \le C_1 n \bar{\sigma}_n^2$, then there exists constant C_2 which only depends on C_1 such that

$$\mathbb{E}_{Y\sim\varphi_{\alpha^*}}[g_{\alpha^*}(\log l(Y))] \le \frac{C_2}{\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^*}.$$

Furthermore, there exist positive constants C_3 and C_4 which only depend on C_1, such that, if $\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^* \ge C_3$, then

$$\mathbb{E}_{Y\sim\varphi_{\alpha^*}}[g_{\alpha^*}(\log l(Y))] \ge \frac{C_4}{\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^*}.$$

As a direct consequence of (2.8),

$$\frac{C_4}{\sqrt{n}\bar{\sigma}_n\alpha^*(1-\alpha^*)} e^{-D_{\alpha^*}(\varphi_0\|\varphi_1)} \le \eta(\varphi_0,\varphi_1) \le \frac{C_2}{\sqrt{n}\bar{\sigma}_n\alpha^*(1-\alpha^*)} e^{-D_{\alpha^*}(\varphi_0\|\varphi_1)},$$

where $D_{\alpha^*}(\varphi_0\|\varphi_1)$ is the Chernoff information defined in (1.1). By (2.4), same upper and lower bound hold for $2R(\hat{z}, z)$.

Remark 1. *A possible (but not necessarily optimal) choice of C_2, C_3 and C_4 can be $C_2 = 1 + 0.28C_1$, $C_3 = 2 \vee [2(0.56C_1)^{3/2}\exp(-\sqrt{2\pi}C_1)]$ and $C_4 = \exp(-2(0.56)\sqrt{2\pi}C_1)/30$. The gap between C_2 and C_4 vanishes when we observe samples with normal distribution, but a positive gap exists in general, e.g., for Bernoulli distribution. We will be demonstrate this fact empirically by simulation in Section 5.*

Remark 2. *The names of Chernoff information/coefficient/divergence in this paper are according to the survey [38]. The Chernoff information indicates the Chernoff α-divergence with $\alpha = \alpha^*$ which maximize $D_\alpha(\varphi_0\|\varphi_1)$. We are not going to calculate the exact value of α^* in this work. $D_{\alpha^*}(\,\cdot\,\|\,\cdot\,)$ represents the Chernoff information, and α^* might be different for variant inputs. α^* in Theorem 2.1 is unique since we assume φ_0 and φ_1 are different on a set with positive measure.*

To gain better understanding of Theorem 2.1, we will introduce a corollary of the i.i.d. case. Under the assumption of i.i.d. sampling, some quantities in the theorem become constants. Since many existing results only consider i.i.d. cases, the following corollary will be helpful for comparison.

Corollary 1 (i.i.d. case)**.** *Let $(\varphi_0^{(n)}, \varphi_1^{(n)}) \in \mathcal{F}(\Omega, \Sigma, \mu, n)$ be sequences of PMFs or PDFs satisfying for $z \in \{0,1\}$, $\varphi_z^{(n)}(x) = \prod_{j=1}^n \bar{\varphi}_z(x_j)$ for fixed $\bar{\varphi}_z$, then*

$$\eta(\varphi_0^{(n)}, \varphi_1^{(n)}) \asymp \frac{1}{\sqrt{n}} e^{-nD_{\alpha^*}(\bar{\varphi}_0 \| \bar{\varphi}_1)}.$$

Proof. Under the assumptions, $\mathbb{E}[|Z_j|^3]$, $\bar{\sigma}_n$ and α^* are constants only depending on $\bar{\varphi}_0$ and $\bar{\varphi}_1$, so the assumption $\sum_{j=1}^n \mathbb{E}|Z_j|^3 \lesssim n\bar{\sigma}_n^2$ is satisfied. Moreover, we have $D_{\alpha^*}(\varphi_0^{(n)} \| \varphi_1^{(n)}) = nD_{\alpha^*}(\bar{\varphi}_0 \| \bar{\varphi}_1)$ and $\frac{1}{\sqrt{n}\bar{\sigma}_n \alpha^*(1-\alpha^*)} \asymp \frac{1}{\sqrt{n}}$. $\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^*$ is sufficiently large as n increases. Hence the result holds by Theorem 2.1. $\square$

Comparison with existing results The Chernoff type lower bound can be traced back to early literature. [2, Theorem 5] produced the following lower bound for Bayes risk, namely

$$R(\hat{z}, z) \geq \frac{1}{4} \min(e^{-\alpha^* \sqrt{n}\bar{\sigma}_n}, e^{-(1-\alpha^*)\sqrt{n}\bar{\sigma}_n}) e^{-D_{\alpha^*}(\varphi_0 \| \varphi_1)}. \tag{2.11}$$

Since $\min(e^{-\alpha^* \sqrt{n}\bar{\sigma}_n}, e^{-(1-\alpha^*)\sqrt{n}\bar{\sigma}_n}) \ll \frac{1}{\sqrt{n}\bar{\sigma}_n \alpha^*(1-\alpha^*)}$, (2.11) is strictly weaker than the result in Theorem 2.1. This lower bound has been applied to the Bayes risk of quantum hypothesis testing, such as [8]. For the i.i.d. case, $D_{\alpha^*}(\varphi_0 \| \varphi_1)$ has been shown to be the best achievable exponent [39, Theorem 11.9.1] and it is restated in [7, Theorem 2.1] as follows:

$$\lim_{n \to \infty} \frac{1}{n} \log \eta(\varphi_0^{(n)}, \varphi_1^{(n)}) = -D_{\alpha^*}(\bar{\varphi}_0 \| \bar{\varphi}_1),$$

under the same conditions as Corollary 1. This result can be obtained from Corollary 1 since the exponent of $\eta(\bar{\varphi}_0, \bar{\varphi}_1)$ has the form $\log \eta(\varphi_0^{(n)}, \varphi_1^{(n)}) = -nD_{\alpha^*}(\bar{\varphi}_0 \| \bar{\varphi}_1) - \frac{1}{2}\log n + O(1)$. The $\log n$ term was investigated in [40], and applied to hypothesis testing problem in [4, Lemma 11]. However, the result can only be applied to Poisson distribution when the samples are i.i.d. in a fixed setting $\bar{\sigma}_n \asymp \frac{\log n}{n}$. If the samples are not identical, their lower bound is not valid. The authors of [6] generalize the result to other setting; however, their bounds cannot be applied to the case when observed data are not i.i.d. Poisson. Therefore, neither of them proposed a minimax lower bound that matches their algorithmic error rate in community detection problems.

2.2. Application to exponential families

With a concrete expressions of φ_0 and φ_1, we can write the Chernoff type bound in Theorem 2.1 with a closed form up to a constant factor. In this section, we are interested in exponential families with PMFs or PDFs of the form

$$\varphi(x;\theta) = h(x)\exp[\theta^\top T(x) - A(\theta)], \tag{2.12}$$

for $x \in \Omega$ and $\theta \in \Theta$. We assume the parameter space Θ is a convex subset of Euclidean space and A is a smooth function on Θ. Let $\{\varphi_{zj} : z \in \{0,1\}, j \in [n]\}$ belongs to an exponential family. To be more specific, we assume that there exist parameters $\theta_{zj} \in \Theta$, $z \in \{0,1\}, j \in [n]$ such that

$$\varphi_{zj}(x_j) = p(x;\theta_{zj}) = h(x_j)\exp[\theta_{zj}^\top T(x_j) - A(\theta_{zj})]. \tag{2.13}$$

We still define φ_0 and φ_1 as in (2.2) on some measure space such that $(\varphi_0, \varphi_1) \in \mathcal{F}(\Omega, \Sigma, \mu, n)$ (see (2.3)). Let us define $\theta_{\alpha j} := (1-\alpha)\theta_{0j} + \alpha\theta_{1j}$, then $\theta_{\alpha j}$ is a valid parameter since we assume the parameter space Θ is convex. The Chernoff α-divergence has a close form:

$$D_\alpha(\varphi_{0j}\|\varphi_{1j}) = (1-\alpha)A(\theta_{0j}) + \alpha A(\theta_{1j}) - A(\theta_{\alpha j}). \tag{2.14}$$

See Section 6.2 for derivation of the last equation. Suppose $Y_j \sim \varphi_{\alpha^* j}$, then using the definition of Z_j in (2.10),

$$Z_j = \log l_j(Y_j) = (\theta_{0j} - \theta_{1j})^\top T(Y_j) - A(\theta_{0j}) + A(\theta_{1j}). \tag{2.15}$$

We have $\mathrm{Var}[Z_j] = (\theta_{0j} - \theta_{1j})^\top \mathbf{H}(A(\theta_{\alpha^* j}))(\theta_{0j} - \theta_{1j})$ where $\mathbf{H}(A(\theta))$ is the Hessian matrix of A evaluated at θ. Now we can establish a corollary when φ_{zj}'s belong to exponential families.

Corollary 2 (exponential family)**.** *Under same assumptions in Theorem 2.1, we assume φ_{zj}'s have the form* (2.13)*, and let*

$$\bar{\sigma}_n := \Big(\frac{1}{n}\sum_{j=1}^n \mathit{Var}[Z_j]\Big)^{1/2} = \Big(\frac{1}{n}\sum_{j=1}^n (\theta_{0j} - \theta_{1j})^\top \mathbf{H}(A(\theta_{\alpha^* j}))(\theta_{0j} - \theta_{1j})\Big)^{1/2}.$$

Suppose $\sum_{j=1}^n \mathbb{E}|Z_j|^3 \le C_1 n\bar{\sigma}_n^2$, using the same constants C_2, C_3 and C_4 in Theorem 2.1, then

$$\eta(\varphi_0,\varphi_1) \le \Big(\frac{C_2}{\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^*} e^{-\sum_{j=1}^n[(1-\alpha^*)A(\theta_{0j}) + \alpha^* A(\theta_{1j}) - A(\theta_{\alpha^* j})]}\Big).$$

If $\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^* \ge C_3$, then we have

$$\eta(\varphi_0,\varphi_1) \ge \Big(\frac{C_4}{\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^*} e^{-\sum_{j=1}^n[(1-\alpha^*)A(\theta_{0j}) + \alpha^* A(\theta_{1j}) - A(\theta_{\alpha^* j})]}\Big).$$

2.3. Application to Bernoulli distribution

We are going to investigate a specific exponential family. Let $p_{zj} \in (0,1)$ and $\theta_{zj} = \log\big(\frac{p_{zj}}{1-p_{zj}}\big)$ for $z \in \{0,1\}, j \in [n]$, and define PMFs of $\mathrm{Bern}(p_{zj})$:

$$\varphi_{zj}(x) := \varphi(x;\theta_{zj}) := \begin{cases} p_{zj}, & \text{if } x = 1, \\ 1 - p_{zj}, & \text{if } x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

It coincides with (2.13) if we let $A(\theta) = \log(1+e^\theta)$, $T(x) = x$ and $h(x) = 1_{\{0,1\}}(x)$, i.e., $\varphi_{zj}(x) = h(x)\exp[\theta_{zj}^\top T(x) - A(\theta_{zj})]$. Let us briefly recall the testing problem in Section 2.1. We randomly draw a number $z \in \{0,1\}$ with equal probability 1/2, and draw a random sample $X = \{X_1, \dots, X_n\}$ where $X_j \sim \text{Bern}(p_{zj})$ independently. As usual, we want to recover z given $X = x \in \{0,1\}^n$. Then we have

$$e^{-D_\alpha(\varphi_0\|\varphi_1)} = \prod_{j=1}^n [p_{0j}^{1-\alpha}p_{1j}^\alpha + (1-p_{0j})^{1-\alpha}(1-p_{1j})^\alpha]. \tag{2.16}$$

Let $p_{\alpha j} = \frac{p_{0j}^{1-\alpha}p_{1j}^\alpha}{p_{0j}^{1-\alpha}p_{1j}^\alpha + (1-p_{0j})^{1-\alpha}(1-p_{1j})^\alpha}$ and recall the definition of α^* and $\bar{\sigma}_n$ from Theorem 2.1, then we have

$$n\bar{\sigma}_n^2 = \sum_{j=1}^n \left[\log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})}\right]^2 p_{\alpha^* j}(1-p_{\alpha^* j}). \tag{2.17}$$

Now let us apply Theorem 2.1. Suppose $\max_{j\in[n]}\left|\log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})}\right| \le C_1$, then there exists constants C_2, C_3 and C_4 which only depend on C_1, such that if $\sqrt{n}\bar{\sigma}_n\alpha^*(1-\alpha^*) \ge C_2$, then the upper and lower bound of $\eta(\varphi_0,\varphi_1) = 2R(\hat{z},z)$ in the theorem holds.

Finally, it is worth mentioning a special case when $p_{z1} = \cdots = p_{zn} := \bar{p}_z$ for $z \in \{0,1\}$. Let $\psi_z(x)$ be the PMF of $\text{Bin}(n,\bar{p}_z)$. Then one can check that

$$\eta(\psi_0,\psi_1) = \eta(\varphi_0,\varphi_1). \tag{2.18}$$

This is due to the fact that, given the observed data x, the optimal test only relies on the minimal sufficient statistic $\sum_{j=1}^n x_j$. This observation can generalize Corollary 1 to the cases when only the sufficient statistic, which is the sum of i.i.d. random variables, are observed. For example, one can apply Corollary 1 to Bayes error probability of Poisson parameter testing by the fact that a Poisson variable is the sum of multiple i.i.d. Poisson variables.

Calculation details in this section will appear in Section 6.4 to 6.7.

3. Community detection in the stochastic block models

The results in Section 2 can apply to many clustering and classification problem in statistics. A typical example is community detection in the stochastic block models (SBM). Given some good estimates of the parameters, community detection is indeed a classification problem. Hence the clustering error rate of the label estimates heavily depends on the Bayes error probability.

3.1. Background of stochastic block models

We will focus on a network which can be represented by a symmetric adjacency matrix $A \in \{0,1\}^{n\times n}$, where the nodes are indexed by $[n]$. We assume that there are K communities on $[n]$, and the membership of the nodes are given by $z \in [K]^n$. Thus $z_i = k$ if node $i \in [n]$ belongs to community $k \in [K]$. We let $n_k := |\{i : z_i = k\}|$ be the size of kth community. Under the assumptions of SBM, given a symmetric connectivity matrix $P \in [0,1]^{K\times K}$,

$$A_{ij} = A_{ji} \sim \mathrm{Bern}(P_{z_i z_j}) \text{ for all } i > j \text{ independently,} \tag{3.1}$$

and $A_{ii} = 0$ for all $i \in [n]$. That is, the connectivity of nodes only depends on their memberships, and there are no self-loops. A fundamental task of community detection on SBM is to recover z given A and K. For consistency of notation with Section 2.3, we define

$$p_{kj} := P_{kz_j} \quad \text{for } k \in [K], j \in [n]. \tag{3.2}$$

In other words, if $z_i = k$, $\mathbb{E}[A_{ij}] = p_{kj}$ whenever $j \neq i$. Thus, the vectors p_{k*} and $\mathbb{E}[A_{i*}]$ are the same at all entries but the ith one. When n is large, the effect of one entry is merely a constant factor. Here $p_{k*} = (p_{k1}, \dots, p_{kn})$, and similarly A_{i*} is the ith row of A. This notation will be used in the rest of this paper. We will consider the parameters satisfying

$$n_k \in \big[\frac{n}{\beta K}, \frac{\beta n}{K}\big], \quad \max_{k,\ell} P_{k\ell} := p^* \leq 1-\varepsilon, \quad \text{and} \quad \frac{\max_{k,\ell} P_{k\ell}}{\min_{k',\ell'} P_{k'\ell'}} \leq \omega \tag{3.3}$$

for some fixed constants K, $\beta > 1, \varepsilon \in (0,1)$, and $\omega > 1$. β controls the balance between different communities. There are no too small or too large communities. All connectivity probabilities are bounded above by $1-\varepsilon$, which is a mild sparsity assumption.

For estimate $\hat{z}$ of z, we are interested in the misclassification rate defined as

$$\mathrm{Mis}(\hat{z}, z) = \min_{\pi \in \mathfrak{S}_K} \frac{1}{n} \sum_{i=1}^{n} 1\{\pi(\hat{z}_i) \neq z_i\} \tag{3.4}$$

where $\mathfrak{S}_K$ is the symmetric group which contains all permutations of $[K]$ and the permutation π will apply entrywisely on the label vector $\hat{z}$.

3.2. Fundamental limit

Let us first consider a simplified symmetric hypothesis testing problem in SBM. In the community detection problem described in the previous section, only the adjacency matrix A and number of community K is given. Now suppose additionally, we know z_{-i}, i.e., all the labels but the ith one, and the connectivity matrix P, our goal is to recover z_i. To further simplify the problem, we assume $z_i \in \{k, \ell\}$, then the hypothesis problem becomes comparison between the parameters p_{k*} and $p_{\ell*}$ defined in (3.2). Since the distribution of Bernoulli vector

A_{i*} can be characterized by p_{k*} if $z_i = k$, we will write

$$\begin{aligned} D_\alpha(p_{k*}\|p_{\ell*}) &:= D_\alpha\Big(\bigotimes_{j=1}^n \mathrm{Bern}(p_{kj})\|\bigotimes_{j=1}^n \mathrm{Bern}(p_{\ell j})\Big), \\ \text{and} \quad \eta(p_{k*}, p_{\ell*}) &:= \eta\Big(\bigotimes_{j=1}^n \mathrm{Bern}(p_{kj}), \bigotimes_{j=1}^n \mathrm{Bern}(p_{\ell j})\Big). \end{aligned} \tag{3.5}$$

which also denote the same quantities if we input the corresponding PMF's. Substituting p_{0*} and p_{1*} with p_{k*} and $p_{\ell*}$ in Section 2.3, and using the assumptions about SBM in Section 3.1, we have the following lemma.

Lemma 1. *Given adjacency matrix A and parameters K, z_{-i} and P, and knowing that $z_i = k$ or ℓ with probability $1/2$, then Bayes estimator $\hat{z}_i$*

$$\hat{z}_i := \arg\max_{r\in\{k,\ell\}} \sum_{j\neq i} A_{ij}\log p_{rj} + (1 - A_{ij})\log(1 - p_{rj})$$

satisfies $\mathbb{P}(\hat{z}_i \neq z_i) = \frac{1}{2}\eta(p_{k*}, p_{\ell*})$. Assuming (3.3), there exists a constant C only depending on β, ε, K and ω such that, if $np^* \leq C(D_{\alpha^*}(p_{k*}\|p_{\ell*}))^2$, then

$$\mathbb{P}(\hat{z}_i \neq z_i) \asymp \Big(\sqrt{np^*}\max_{j\in[n]}\Big|\log\frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})}\Big|\Big)^{-1}\exp(-D_{\alpha^*}(p_{k*}\|p_{\ell*})).$$

Now we will derive a minimax lower bound of community detection problem. We will consider the following parameter space:

$$\begin{aligned} \Theta(n, K, p, q) := \Big\{&(z, P) : P \in (0,1)^{K\times K}, P = P^\top, P_{kk} \geq p, \\ &P_{k\ell} \leq q \text{ if } k \neq \ell, p_{kj} = P_{kz_j}, \text{ and (3.3) is satisfied}\Big\}. \end{aligned} \tag{3.6}$$

Theorem 3.1 (minimax lower bound)**.** *We define*

$$I_K = \begin{cases} -\frac{2}{K}\log[\sqrt{pq} + \sqrt{(1-p)(1-q)}], & \text{if } K = 2; \\ -\frac{2}{\beta K}\log[\sqrt{pq} + \sqrt{(1-p)(1-q)}], & \text{if } K \geq 3. \end{cases} \tag{3.7}$$

If $p > q$ and $nI_K^2 \geq Cp$ for some C only depending on β, ε, K and ω, then

$$\inf_{\hat{z}} \sup_{\Theta(n,K,p,q)} \mathbb{E}[\mathrm{Mis}(\hat{z}, z)] \gtrsim \Big(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\Big)^{-1}\exp(-nI_K). \tag{3.8}$$

The proof will appear in Section 7.3, which is inspired by [36]. Compared with the minimax lower bound in [36], which states

$$\inf_{\hat{z}} \sup_{\Theta(n,K,p,q)} \mathbb{E}[\mathrm{Mis}(\hat{z}, z)] \geq \exp(-(1 + o(1))nI_K),$$

Theorem 3.1 specifies that the $o(1)$ term is of the form

$$\frac{1}{nI_K}\Big(\frac{1}{2}\log(np) + \log\log\frac{p(1-q)}{q(1-p)}\Big). \tag{3.9}$$

Considering the case $p/q \to c > 1$ and $np \to \infty$, the higher order term $\left(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\right)^{-1} \asymp (np)^{-1/2}$ converges to 0 as average degree increases. It requires extra effort to find an algorithm achieving this sharp lower bound.

Remark 3. *The condition $nI_K^2 \geq Cp$ is not required in [36], but it is needed in this theorem due to a technical reason. Essentially, the lower bound in Theorem 2.1 requires the condition $\sqrt{n}\bar{\sigma}_n(1-\alpha^*)\alpha^* \geq C_3$ for some sufficiently large C_3. The corresponding condition $nI_K^2 \geq Cp$ in community detection scenario is needed due to Lemma 4. In other words, the prefactor $\left(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\right)^{-1}$ in (3.8) is valid only if it is small enough.*

3.3. Algorithm achieving the minimax lower bound

Our algorithm is inspired by the pseudo-likelihood approach in [34]. We define an operator to estimate P given adjacency matrix A and estimated labels $\tilde{z}$:

$$\mathcal{B}(A,\tilde{z}) := (\hat{P}_{k\ell}) \in [0,1]^{K\times K}, \quad \hat{P}_{k\ell} := \frac{\sum_{i>j} A_{ij}1\{\tilde{z}_i = k, \tilde{z}_j = \ell\}}{\sum_{i>j} 1\{\tilde{z}_i = k, \tilde{z}_j = \ell\}}. \tag{3.10}$$

We will also use likelihood ratio classifier defined as follows:

$$\begin{aligned}\mathcal{L}(A,\hat{P},\tilde{z}) &:= (\hat{z}_i) \in [K]^n,\\ \hat{z}_i = \arg\max_{k\in[K]} &\sum_{j\neq i} A_{ij}\log\hat{P}_{k\hat{z}_j} + (1-A_{ij})\log(1-\hat{P}_{k\hat{z}_j}).\end{aligned} \tag{3.11}$$

Note that we can apply these two operators on submatrices of A with the corresponding indices if needed. This is an EM-type algorithm if we repeat (3.10) and (3.11) iteratively, i.e., (3.10) is the expectation step and (3.11) is the maximization step. As pointed out in [6], it requires at least two iterations of EM-type update to achieve the optimal error rate up to a constant. To generate enough independence between iterations, we combine the block partition method in [37] and "leave-one-out" trick in [35]. It is worth noting that besides the dependence between A and $\tilde{z}$ in $\mathcal{L}(A,\hat{P},\tilde{z})$, other dependence can be handled by uniform bounds. Details about Algorithm 1 will be describe as follows:

Step 3 to 4: We apply spectral clustering on the whole adjacency matrix. However, we will only use its output in the matching step (step 9) and approximate an initial estimate $\tilde{P}$ of P. The dependence between $\tilde{P}$ and A can be handled by uniform bounds.

Step 5: This is the block partitioning trick. Data in different blocks will be used in different steps to acquire independence.

Step 6 to 7: This is the "leave-one-out" trick. In each iteration, we only use the data of the jth node in step 12, so the last likelihood ratio classifier will be independent with other steps in the for loop.

Step 8 to 9: We apply spectral clustering on two of the subblocks. Although the labels from spectral clustering have consistent misclassification, but the corresponding optimal permutations in (3.4) are not necessarily the same in general. This issue can be solved by step 9. After the matching step, the new label

Algorithm 1 Community detection

1: **Input:** Adjacency matrix A, number of communities K.
2: **Output:** Estimated labels $\hat{z}$.
3: $\tilde{z} \leftarrow \mathrm{SC}(A, K)$.
4: $\tilde{P} \leftarrow \mathcal{B}(A, \tilde{z})$.
5: Let $I \subset [n]$ with $|I| = \lfloor n/2 \rfloor$ be a random subset of indices. Let $J = [n] \backslash I$.
6: **for** $j = 1$ to n **do**
7: $\quad I' \leftarrow I \backslash \{j\}$, $J' \leftarrow J \backslash \{j\}$.
8: $\quad \tilde{z}'_{I'} \leftarrow \mathrm{SC}(A_{I' \times I'}, K)$, $\tilde{z}'_{J'} \leftarrow \mathrm{SC}(A_{J' \times J'}, K)$.
9: $\quad \tilde{z}'_{I'} \leftarrow \textsc{Match}(\tilde{z}_{I'}, \tilde{z}'_{I'})$, $\tilde{z}'_{J'} \leftarrow \textsc{Match}(\tilde{z}_{J'}, \tilde{z}'_{J'})$.
10: $\quad \tilde{z}'_{I'} \leftarrow \mathcal{L}(A_{I' \times J'}, \tilde{P}, \tilde{z}'_{J'})$, $\tilde{z}'_{J'} \leftarrow \mathcal{L}(A_{J' \times I'}, \tilde{P}, \tilde{z}'_{I'})$.
11: $\quad \tilde{z}' \leftarrow (\tilde{z}_i, \tilde{z}'_{I'}, \tilde{z}'_{J'})$, $\hat{P} \leftarrow \mathcal{B}(A, \tilde{z}')$.
12: $\quad \hat{z}_j \leftarrow \mathcal{L}(A_{j*}, \hat{P}, \tilde{z}')$.
13: **end for**
14: **function** $\mathrm{SC}(A, K)$
15: $\quad$ Apply degree-truncation to A to obtain A_{re}.
16: $\quad$ Apply SVD on A_{re} so that $A_{\mathrm{re}} = U \Sigma U^T$. Let $\hat{\Sigma}$ contains top K singular values on the diagonal and $\hat{U}$ contains corresponding singular vectors.
17: $\quad$ Output the K-means clustering result on the rows of $\hat{U}\hat{\Sigma}$.
18: **end function**
19: **function** $\textsc{Match}\ (\tilde{z}, z)$
20: $\quad \tilde{z} \leftarrow \arg\min_{\pi(\tilde{z}): \pi \in \mathfrak{S}_K} \sum_{i=1}^{n} 1\{z_i \neq \pi(z_i)\}$.
21: $\quad$ Output $\tilde{z}$.
22: **end function**

vector $\tilde{z}'$ has the same permutation as $\tilde{z}$ when computing the misclassification rate. This fact will be clarified in the proof. Note that although $\tilde{z}$ depends on A, $\tilde{z}'_{I'}$ and $\tilde{z}'_{J'}$ only depend on the corresponding subblocks as long as the spectral clustering algorithm outputs good enough labels.

Step 10: We apply the first likelihood ratio classifier on a different subblock using estimated connectivity matrix from step 4 and labels from step 9.

Step 11: We obtain the updated labels and estimate the connectivity matrix by $\hat{P}$ according to the new labels..

Step 12: We update the label again according to the new $\hat{P}$ and $\tilde{z}'$ obtained in step 11 by likelihood ratio classifier.

Step 14 to 18: A spectral clustering algorithm proposed in [22]. Details of the degree-truncation step appears in Section 8.

Step 19 to 22: A matching algorithm finding the optimal permutation between labels. A linear assignment algorithm with computational complexity $O(K^3)$ can find the exact solution of $\tilde{z}$ [41].

The following block matrix might help understand the partitioning of adjacency matrix A in the algorithm.

$$
A = \left[\begin{array}{c|cc} 0 & \multicolumn{2}{c}{A_{j \times (I' \cup J')}} \\ \hline \vdots & A_{I' \times I'} & A_{I' \times J'} \\ \cline{2-3} & A_{J' \times I'} & A_{J' \times J'} \end{array}\right], \quad \left[\begin{array}{c|c|c} 0 & \multicolumn{2}{c}{\text{2nd LR (step 12)}} \\ \hline \vdots & \begin{array}{c}\text{2nd SC}\\ \text{(step 8)}\end{array} & \begin{array}{c}\text{1st LR}\\ \text{(step 10)}\end{array} \\ \cline{2-3} & \begin{array}{c}\text{1st LR}\\ \text{(step 10)}\end{array} & \begin{array}{c}\text{2nd SC}\\ \text{(step 8)}\end{array} \end{array}\right]
$$

Note that " $\vdots$ " represents the block $A^{\top}_{j\times(I'\cup J')}$. We can see that the second spectral clustering and both likelihood ratio tests are applied on different blocks of the adjacency matrix, so we do not need to worry about dependence between steps. Now we present the theoretical guarantees of the output of Algorithm 1.

Theorem 3.2. *Let us assume* (3.3) *for some fixed constants* β, ω, ε, *and* K. *We also briefly denote* $D^* := \min_{k\neq\ell} D_{\alpha^*}(p_{k*}\|p_{\ell*})$ *and* $\eta^* = \max_{k\neq\ell}\eta(p_{k*}, p_{\ell*})$. *There exists constant* C *only depending on* β, ε, K *and* ω *such that, if* $Cnp^* \leq (D^*)^2$, *then the output* $\hat{z}$ *from Algorithm 1 satisfies:*

- *(a) If* $D^* \leq 2\log n$, *then* $\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] = O(\eta^*)$.
- *(b) If* $n\eta^* = o(1)$, *then* $\hat{z}$ *achieves exact recovery with high probability, i.e.,* $\mathbb{P}(\hat{z} = z) \geq 1 - o(1)$.

Moreover, η^* *in (a) can be replaced by*

$$\max_{k\neq\ell}\Big(\sqrt{np^*}\max_{j\in[n]}\Big|\log\frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})}\Big|\Big)^{-1}\exp(-D_{\alpha^*}(p_{k*}\|p_{\ell*})).$$

Case (a) and case (b) in the theorem have described all situations in the model. Case (a) assumes $D^* \leq 2\log n$, and if $D^* > 2\log n$, it is easy to check that $n\eta^* = o(1)$ with the help of Lemma 1.

Theorem 3.2 immediately implies the error rate on the parameter space $\Theta(n, K, p, q)$ defined in (3.6) by considering the least favorable submodel. It shows that the misclassification error of our algorithm is rate optimal in this parameter space. We summarize this result in the following corollary.

Corollary 3. *Recall* I_K *in* (3.7) *and suppose it satisfies* $\sqrt{Cnp^*} \leq D^* \leq 2\log n$ *for the constant* C *in Theorem 3.2, then the output* $\hat{z}$ *from Algorithm 1 satisfies*

$$\sup_{\Theta(n,K,p,q)}\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] \lesssim \Big(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\Big)^{-1}\exp(-nI_K).$$

Remark 4 (Comparison with existing results.)**.** *We have already compared some results in literature. Here, we will summarize the novelty in details.*

1. *Existing papers either consider the asymptotic behavior of optimal community detection in general undirected or bipartite SBM [6] or symmetric assortative SBM [37, 35, 42]. We extend the algorithms to general SBM.*
2. *We apply twice local updates (likelihood ratio tests) on symmetric adjacency matrix in our algorithm. It is also possible to apply multiple times by partitioning more blocks. Although multiple steps of local updates are allowed in [43] by variational inference, data splitting method for initialization is required and lacking in their algorithm.*
3. *By the new Chernoff bound introduced in Theorem 2.1, we provide sharpened minimax error rate and tight misclassification rate for our algorithm. In particular, we replace the uncertain term* $\exp(o(1)nI_K)$ *in [36] by an explicit expression. Although a high order term is also discovered in [42], their result only applies to assortative SBM with* $K = 2, P_{11} = P_{12} = p > P_{12} = P_{21} = q$ *and* $p \asymp q \asymp \frac{\log n}{n}$. *Our algorithm applies to general SBM.*

4. *In a more general setting than [42], the authors of [21, 32, 33] consider* $K = 2, P_{11} = P_{22} = p > P_{12} = P_{21} = q$ *and their algorithms can achieve error rate* $\exp(-(1-o(1))nI_2)$*, where* I_2 *is defined in* (3.7)*. In particular, the term* $o(1)$ *can behave like* $O((nI_2)^{-1/2})$ *in [32, 33]. The error rates in these work is not as sharp as the result in Corollary 3 because* $D^* = nI_2$ *in this setting and our theorem shows the error rate can be as sharp as* $\exp(-(1+o(1))nI_2)$*, where the* $o(1)$ *term is of the form* (3.9)*.*
5. *When considering the error rate in Corollary 3, the condition* $Cnp^* \leq (D^*)^2$ *is not required in [21, 32, 33]. However, this is due to the fact that they consider a simpler model. Under the same setting, if we use the error rate proposed in these works for our initialization steps, e.g., step 8 in Algorithm 1, then the condition* $Cnp^* \leq (D^*)^2$ *can be removed in Corollary 3. However, we have not found the generalization of the results in [21, 32, 33] to the cases when* $K > 2$ *with sharp enough error rates, so we still require the condition* $Cnp^* \leq (D^*)^2$ *in our theorem.*

4. Discussion

We discuss some possible extensions and future works in this section.

4.1. Rate optimal Chernoff bound for quantum hypothesis testing

The Quantum Chernoff bound has been shown to be the upper [44] and lower [7] bound of symmetric quantum hypothesis testing. In the proof of Chernoff lower bound, the authors reduce the problem from quantum setting to classical probability space, and apply classical Chernoff bound. Hence, the second-order term in Theorem 2.1 can apply to lower bound immediately. However, its application to Chernoff upper bound is more technical.

4.2. Simplified feasible algorithm for community detection

It was pointed out in [35] that the "leave-one-out" trick is not necessary in practice, so it only requires a single spectral clustering. See Algorithm 3 in their paper. However, they cannot provide theoretical guarantee for this simplified algorithm. To the best of our knowledge, considering general SBM, there is no algorithm with finite number of global method (either spectral clustering or semidefinite programming) can achieve the error rate in Theorem 3.2. The idea of "leave-one-out" can be generalized to leave more than one out [42], but minimum number of global methods still grows as the number of nodes increases.

With the assortativity assumption, a concurrent work [32] shows that in the simplest SBM setting, i.e., $K = 2$, $P_{11} = P_{22} = p > P_{12} = P_{21} = q$, a semidefinite programming approach can achieve a sharp error rate of the form $\exp((1-o(1))I_2)$, where I_2 is defined in (3.7). With such a tight error rate in the initialization step, it only requires one step of global update the achieve the minimax lower bound in Theorem 3.1. However, the "leave-one-out" step is still required, and their method does not apply to general SBM, e.g., when $q > p$.

5. Simulation

We will show that, in some asymptotic setting, the Bayes error probability converges with a rate expected in Theorem 2.1 by simulation. Since the exponent of the error rate has been well known, we will focus on the second-order asymptotics in our experiments. Let us consider Bernoulli distributions analyzed in Section 2.3. Let $p_{01} = p_{02} = \cdots = p_{0n} = p_{1(n+1)} = p_{1(n+2)} = \cdots = p_{1(2n)} = 0.55$, and $p_{11} = p_{12} = \cdots = p_{1n} = p_{0(n+1)} = p_{0(n+2)} = \cdots = p_{0(2n)} = 0.45$, i.e.,

$$p_{0*} = (\underbrace{0.55, 0.55, \ldots, 0.55}_{n \text{ times}}, \underbrace{0.45, 0.45, \ldots, 0.45}_{n \text{ times}});$$
$$p_{1*} = (\underbrace{0.45, 0.45, \ldots, 0.45}_{n \text{ times}}, \underbrace{0.55, 0.55, \ldots, 0.55}_{n \text{ times}}).$$

Now we consider the Chernoff α-divergence. The optimal α^* in Theorem 2.1 is 1/2 by symmetry. Using the notation in (3.5), by (2.16), we have

$$e^{-D_{\alpha^*}(p_{0*}\|p_{1*})} = (2\sqrt{0.55 \cdot 0.45})^{2n}.$$

By (2.17) with some details in Section 6.4, we have

$$n\bar{\sigma}_n^2 = \sum_{j=1}^{2n} \Big(\log \frac{0.55^2}{0.45^2}\Big)^2 0.5(1-0.5) \asymp n.$$

By Theorem 2.1, we expect

$$\eta(p_{0*}, p_{1*}) \asymp \frac{1}{\sqrt{n}} (2\sqrt{0.55 \cdot 0.45})^{2n}.$$

Or equivalently,

$$a_n := \log \eta(p_{0*}, p_{1*}) - 2n \log(2\sqrt{0.55 \cdot 0.45})$$

asymptotically behaves like $-\frac{1}{2}\log n + C$. We can also think of p_{0*} and p_{1*} as the parameters in (3.2) associated with SBM with community sizes $n_0 = n_1 = n$ and connectivity matrix

$$P = \begin{bmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{bmatrix}.$$

By Theorem 3.2, we expect

$$b_n := \log(2 \cdot \text{Misclassification rate}) - 2n \log(2\sqrt{0.55 \cdot 0.45})$$

also tends to $-\frac{1}{2}\log n + C$. Note that 2 comes from the fact that $\eta(p_{0*}, p_{1*}) = 2 \cdot (\text{Bayes error probability})$ would help the simulation scale better. We will use the true Bernoulli PMF to compute $\eta(p_{0*}, p_{1*})$, then find the misclassification rate of Algorithm 1 and compute b_n.

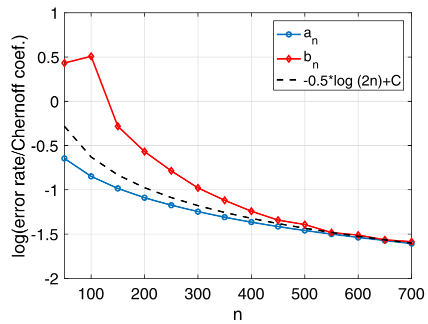

FIG 1. *Second-order asymptotics of a_n and b_n.*

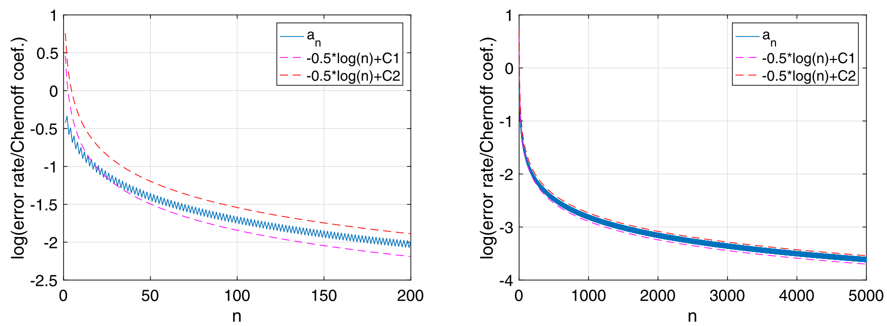

FIG 2. *Second-order asymptotics of Bayes probability error.*

From Figure 1, we observe that n increases, both a_n and b_n behave like $-\frac{1}{2}\log n + C$ for the same constant C. For smaller n, the misclassification rate is large since initialization in Algorithm 1 is not accurate enough; however, b_n becomes stable when n gets larger.

Another interesting empirical result we want to show by simulation is that, there is a gap between the constants C_2 and C_4 in Theorem 2.1 in general. We let $p_{0*} = 0.3 \cdot \mathbf{1}_n$ and $p_{1*} = 0.7 \cdot \mathbf{1}_n$, i.e.,

$$p_{0*} = (\underbrace{0.3, 0.3, \dots, 0.3}_{n \text{ times}}) \quad \text{and} \quad p_{1*} = (\underbrace{0.7, 0.7, \dots, 0.7}_{n \text{ times}}).$$

By symmetry, we have $\alpha^* = 1/2$, so

$$\eta(p_{0*}, p_{1*}) \asymp \frac{1}{\sqrt{n}}(2\sqrt{0.3 \cdot 0.7})^n.$$

Again, we let $a_n = \log \eta(p_{0*}, p_{1*}) - n \log(2\sqrt{0.3 \cdot 0.7})$. The following plots shows the behavior of a_n. They are plots of a_n in different ranges of n. Although a_n asymptotically behaves like $-\frac{1}{2}\log n$, it oscillates up and down until infinity. This simulation result empirically shows that, $a_n + \frac{1}{2}\log n$ does not converge to any constant for such p_{0*} and p_{1*}.

6. Proofs of Section 2

6.1. Proof of Theorem 2.1

Lemma 2. *We recall $\alpha^* = \arg\max_{\alpha\in(0,1)} D_\alpha(\varphi_0\|\varphi_1)$ and $Y \sim \varphi_{\alpha^*}$ from the assumption of the theorem, then $\mathbb{E}[\log l(Y)] = 0$.*

Proof. By definition of Y, we have

$$\mathbb{E}[\log l(Y)] = \int_\Omega \varphi_{\alpha^*} \log l(y) d\mu(y) = e^{-D_{\alpha^*}(\varphi_0\|\varphi_1)} \int_\Omega \varphi_0^{1-\alpha_*}\varphi_1^{\alpha_*} \log \frac{\varphi_0}{\varphi_1} d\mu$$

We recall that we assume the Kullback–Leibler divergences $D_{\mathrm{KL}}(\varphi_0\|\varphi_1)$ and $D_{\mathrm{KL}}(\varphi_0\|\varphi_1)$ exist, so for $\alpha \in (0,1), \left|\varphi_0^{1-\alpha}\varphi_1^\alpha \log \frac{\varphi_0}{\varphi_1}\right| \le \left|(\varphi_0+\varphi_1)\log\frac{\varphi_1}{\varphi_0}\right|$ is integrable. By the mean value theorem and the dominated convergence theorem,

$$\int_\Omega \varphi_0^{1-\alpha}\varphi_1^\alpha \log\frac{\varphi_0}{\varphi_1} d\mu = -\int_\Omega \frac{d}{d\alpha}\varphi_0^{1-\alpha}\varphi_1^\alpha d\mu = -\frac{d}{d\alpha}\int_\Omega \varphi_0^{1-\alpha}\varphi_1^\alpha d\mu. \tag{6.1}$$

Since $\alpha \mapsto -\varphi_0(x)^{1-\alpha}\varphi_1(x)^\alpha$ is convex for $x\in\Omega$, $\alpha \mapsto -\int_\Omega \varphi_0^{1-\alpha}\varphi_1^\alpha d\mu$ is also convex, and it is indeed strictly convex if $\varphi_0 \ne \varphi_1$ on a set with nonzero measure. Therefore, $D_\alpha(\varphi_0\|\varphi_1)$ achieves maximum if and only if $\frac{d}{d\alpha}\int_\Omega \varphi_0^{1-\alpha}\varphi_1^\alpha d\mu = 0$, which is true if we evaluate at $\alpha = \alpha^*$. Hence $\mathbb{E}[\log l(Y)] = 0$. □

Proposition 1. *Let Φ be the cumulative distribution function of standard normal distribution, then for $x > 0$,*

$$\frac{1}{x} - \frac{1}{x^3} \le \sqrt{2\pi} e^{x^2/2}\Phi(-x) \le \frac{1}{x}. \tag{6.2}$$

and

$$\sqrt{2\pi} e^{x^2/2}\Big(\Phi(x) - \frac{1}{2}\Big) \ge x + \frac{x^3}{3}. \tag{6.3}$$

In particular,

$$\Phi(x) \ge \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\Big(x - \frac{2x^3}{3}\Big). \tag{6.4}$$

Proof. For $x > 0$, we have divergent series expanded at ∞:

$$\sqrt{2\pi}e^{x^2/2}\Phi(-x) = \frac{1}{x} + \sum_{i=1}^\infty \frac{(-1)^i(2i-1)!}{2^{i-1}(i-1)!}\frac{1}{x^{2i+1}} = \frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} - \frac{15}{x^7} + \dots$$

which implies (6.2). We also have the power series expanded at 0. Let $m!! = 1\cdot 3\cdot\dots\cdot m$ for odd integer m, we have

$$\sqrt{2\pi}e^{x^2/2}\Big(\Phi(x) - \frac{1}{2}\Big) = \sum_{i=0}^\infty \frac{x^{2i+1}}{(2i+1)!!} = x + \frac{x^3}{3} + \frac{x^5}{3\cdot 5} + \frac{x^7}{3\cdot 5\cdot 7} + \dots$$

which implies (6.3). Then we expand $e^{-x^2/2}\left(x+\frac{x^3}{3}\right)$ and have

$$e^{-x^2/2}\left(x+\frac{x^3}{3}\right)=x-\frac{2x^3}{3}+\frac{7x^5}{30}-\dots$$

Then we obtain (6.4). $\square$

Lemma 3. *Recall from* (2.7) *that* $g_\alpha(x)=\exp(\min(\alpha x,(\alpha-1)x))$. *Suppose* $Z\sim\mathcal{N}(0,\sigma^2)$, *then*

$$\frac{1}{\sqrt{2\pi}\sigma\alpha(1-\alpha)}-\frac{1}{\sqrt{2\pi}\sigma^3\alpha^3(1-\alpha)^3}\le\mathbb{E}[g_\alpha(Z)]\le\frac{1}{\sqrt{2\pi}\sigma\alpha(1-\alpha)}.$$

Proof. We have

$$\begin{aligned}\mathbb{E}[g_\alpha(Z)]&=\mathbb{E}[\exp(\min(\alpha Z,(\alpha-1)Z))]\\&=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{0}e^{\alpha x}e^{-\frac{x^2}{2\sigma^2}}dx+\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{(\alpha-1)x}e^{-\frac{x^2}{2\sigma^2}}dx\\&=\frac{e^{\sigma^2\alpha^2/2}}{\sqrt{2\pi}\sigma}\int_{-\infty}^{0}e^{-\frac{(x-\sigma^2\alpha)^2}{2\sigma^2}}dx+\frac{e^{\sigma^2(1-\alpha)^2/2}}{\sqrt{2\pi}\sigma}\int_0^\infty e^{-\frac{(x-\sigma^2(1-\alpha))^2}{2\sigma^2}}dx\\&=\exp\left(\frac{\sigma^2\alpha^2}{2}\right)\Phi(-\sigma\alpha)+\exp\left(\frac{\sigma^2(1-\alpha)^2}{2}\right)\Phi(-\sigma(1-\alpha))\end{aligned}$$

By (6.2), we have

$$\mathbb{E}[g_\alpha(Z)]\le\frac{1}{\sqrt{2\pi}\sigma\alpha}+\frac{1}{\sqrt{2\pi}\sigma(1-\alpha)}=\frac{1}{\sqrt{2\pi}\sigma\alpha(1-\alpha)},$$

and

$$\begin{aligned}\mathbb{E}[g_\alpha(Z)]&\ge\frac{1}{\sqrt{2\pi}\sigma\alpha}-\frac{1}{\sqrt{2\pi}\sigma^3\alpha^3}+\frac{1}{\sqrt{2\pi}\sigma(1-\alpha)}-\frac{1}{\sqrt{2\pi}\sigma^3(1-\alpha)^3}\\&\ge\frac{1}{\sqrt{2\pi}\sigma\alpha(1-\alpha)}-\frac{1}{\sqrt{2\pi}\sigma^3\alpha^3(1-\alpha)^3}.\square\end{aligned}$$

Theorem 6.1 (Berry-Esseen theorem)**.** *Let* $\{Z_i\}_{i=1}^n$ *be independent random variables with zero means and* $\mathbb{E}[\sum_{i=1}^n Z_i^2]=\sigma^2$. *Let* F *be the distribution function of* $\sum_{i=1}^n Z_i/\sigma$, *then there exists an absolute constant* $C_0\le 0.56$ *such that for every* $x\in\mathbb{R}$,

$$|F(x)-\Phi(x)|\le\frac{C_0\sum_{i=1}^n\mathbb{E}[|Z_i|^3]}{\sigma^3}. \tag{6.5}$$

Proof. The best upper bound of C_0 so far is given by [45]. We would also like to refer readers to see a proof by Stein's method in [46]. $\square$

Proof of Theorem 2.1. Let $\alpha=\alpha^*$. By Lemma 2, $\mathbb{E}[\sum_{j=1}^n Z_j]=\mathbb{E}[\log l(Y)]=0$. Let us define $\sigma^2=\mathbb{E}[\sum_{j=1}^n Z_j^2]$ as in Theorem 6.1. Note that $\sigma=\sqrt{n}\bar{\sigma}_n$. By assumption $\sum_{j=1}^n\mathbb{E}|Z_j|^3\le C_1 n\bar{\sigma}_n^2$, the distribution function F of $\sum_{j=1}^n Z_j/\sigma$

satisfies for $x \in \mathbb{R}$, $|F(x) - \Phi(x)| \leq \frac{C}{\sigma}$ where $C := 0.56C_1$. Recall that $\log l(Y) = \sum_{j=1}^n Z_j$, we have

$$\begin{aligned}
\mathbb{E}[g_\alpha(\log l(Y))] &= \int_0^1 \mathbb{P}(g_\alpha(\log l(Y)) > x)dx \\
&= \int_0^1 \mathbb{P}\Big(\frac{\log x}{\alpha\sigma} < \frac{\log l(Y)}{\sigma} < \frac{\log x}{(\alpha-1)\sigma}\Big)dx \\
&= \int_0^1 F\Big(\frac{\log x}{(\alpha-1)\sigma}\Big) - F\Big(\frac{\log x}{\alpha\sigma}\Big)dx \\
&\leq \int_0^1 \Phi\Big(\frac{\log x}{(\alpha-1)\sigma}\Big) - \Phi\Big(\frac{\log x}{\alpha\sigma}\Big)dx + \frac{2C}{\sigma} \\
&= \frac{1}{\sqrt{2\pi}\sigma\alpha(1-\alpha)} + \frac{2C}{\sigma} \leq \frac{1 + C/2}{\sigma\alpha(1-\alpha)}.
\end{aligned}$$

On the other hand, for any $t \in [0, 1]$,

$$\begin{aligned}
\mathbb{E}[g_\alpha(\log l(Y))] &= \int_0^1 F\Big(\frac{\log x}{(\alpha-1)\sigma}\Big) - F\Big(\frac{\log x}{\alpha\sigma}\Big)dx \\
&\geq \int_0^t \Phi\Big(\frac{\log x}{(\alpha-1)\sigma}\Big) - \Phi\Big(\frac{\log x}{\alpha\sigma}\Big) - \frac{2C}{\sigma}dx \\
&= \int_0^t \Phi\Big(\frac{\log x}{(\alpha-1)\sigma}\Big) - \Phi\Big(\frac{\log x}{\alpha\sigma}\Big)dx - \frac{2tC}{\sigma}.
\end{aligned}$$

By Fubini's theorem,

$$\begin{aligned}
&\int_0^t \Phi\Big(\frac{\log x}{\sigma(\alpha-1)}\Big) - \Phi\Big(\frac{\log x}{\sigma\alpha}\Big)dx = \int_0^t \frac{1}{\sqrt{2\pi}} \int_{\frac{\log x}{\sigma\alpha}}^{\frac{\log x}{\sigma(\alpha-1)}} e^{-y^2/2}dydx \\
=&\frac{1}{\sqrt{2\pi}}\Big[\int_{-\infty}^{\frac{\log t}{\sigma\alpha}} e^{\sigma\alpha y - \frac{y^2}{2}}dy + \int_{\frac{\log t}{\sigma(\alpha-1)}}^{\infty} e^{\sigma(\alpha-1)x - \frac{y^2}{2}}dy\Big] \qquad (6.6) \\
&+ t\Big[\Phi\Big(\frac{\log t}{\sigma(\alpha-1)}\Big) - \Phi\Big(\frac{\log t}{\sigma\alpha}\Big)\Big].
\end{aligned}$$

The first integral in the last step can be evaluated as

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log t}{\sigma\alpha}} e^{\sigma\alpha y - \frac{y^2}{2}}dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\log t}{\sigma\alpha}} e^{\frac{\sigma^2\alpha^2}{2} - \frac{(y+\sigma\alpha)^2}{2}}dy = e^{\frac{\sigma^2\alpha^2}{2}}\Phi\Big(\frac{\log t}{\sigma\alpha} - \sigma\alpha\Big).$$

For the second integral in the last step of (6.6), we similarly have

$$\frac{1}{\sqrt{2\pi}} \int_{\frac{\log t}{\sigma(\alpha-1)}}^{\infty} e^{\sigma(\alpha-1)x - \frac{y^2}{2}}dy = e^{\frac{\sigma^2(1-\alpha)^2}{2}}\Phi\Big(\frac{\log t}{\sigma(1-\alpha)} - \sigma(1-\alpha)\Big).$$

Assuming $\sigma\alpha(1-\alpha) \geq \sqrt{2\pi}C \vee 2$ and letting $t = \exp[-2\sqrt{2\pi}C(1-\alpha)\alpha]$, using $\alpha(1-\alpha) \leq 1/4$, we have

$$-\frac{\log t}{\sigma\alpha} \leq \frac{2\sqrt{2\pi}C\alpha(1-\alpha)}{\sqrt{2\pi}C} \leq \frac{1}{2} \quad \text{and} \quad \sigma\alpha - \frac{\log t}{\sigma\alpha} \leq \sigma\alpha + \frac{1}{2} \leq \frac{5\sigma\alpha}{4}.$$

By (6.2) and the fact that the function $\frac{1}{x} - \frac{1}{x^3}$ is decreasing on $[\sqrt{3}, \infty]$, we have

$$\begin{aligned}
&e^{\frac{\sigma^2\alpha^2}{2}}\Phi\Big(\frac{\log t}{\sigma\alpha} - \sigma\alpha\Big)\\
&\geq \frac{1}{\sqrt{2\pi}}\exp\bigg(\frac{\sigma^2\alpha^2 - \big(\frac{\log t}{\sigma\alpha} - \sigma\alpha\big)^2}{2}\bigg)\bigg[\frac{1}{\sigma\alpha - \frac{\log t}{\sigma\alpha}} - \frac{1}{\Big(\sigma\alpha - \frac{\log t}{\sigma\alpha}\Big)^3}\bigg]\\
&\geq \frac{1}{\sqrt{2\pi}}\exp\bigg(\log t - \frac{1}{2}\Big(\frac{\log t}{\sigma\alpha}\Big)^2\bigg)\Big[\frac{4}{5\sigma\alpha} - \frac{64}{125\sigma^3\alpha^3}\Big]\\
&\geq \frac{1}{\sqrt{2\pi}}\exp(-2\sqrt{2\pi}C(1-\alpha)\alpha - 1/8)\Big[\frac{4}{5\sigma\alpha} - \frac{64}{125(4\sigma\alpha)}\Big]\\
&\geq \frac{\exp(-2\sqrt{2\pi}C)}{5\sigma\alpha}.
\end{aligned}$$

Similarly, we have

$$e^{\frac{\sigma^2(1-\alpha)^2}{2}}\Phi\Big(\frac{\log t}{\sigma(1-\alpha)} - \sigma(1-\alpha)\Big) \geq \frac{\exp(-2\sqrt{2\pi}C)}{5\sigma(1-\alpha)}.$$

Hence the integral in (6.6) has lower bound

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\Big[\int_{-\infty}^{\frac{\log t}{\sigma\alpha}} e^{\sigma\alpha y - \frac{y^2}{2}}\,dy + \int_{\frac{\log t}{\sigma(\alpha-1)}}^{\infty} e^{\sigma(\alpha-1)x - \frac{y^2}{2}}\,dy\Big]\\
&\qquad \geq \frac{\exp(-2\sqrt{2\pi}C)}{5\sigma\alpha} + \frac{\exp(-2\sqrt{2\pi}C)}{5\sigma(1-\alpha)} = \frac{\exp(-2\sqrt{2\pi}C)}{5\sigma\alpha(1-\alpha)}.
\end{aligned} \tag{6.7}$$

Now we consider another term $\Phi\Big(\frac{\log t}{\sigma(\alpha-1)}\Big)$ in (6.6). By (6.4), we have

$$\begin{aligned}
\Phi\Big(\frac{\log t}{\sigma(\alpha-1)}\Big) &\geq \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\Big(\frac{\log t}{\sigma(\alpha-1)} - \frac{2}{3}\Big(\frac{\log t}{\sigma(\alpha-1)}\Big)^3\Big)\\
&= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\Big(\frac{-2\sqrt{2\pi}C\alpha(1-\alpha)}{\sigma(\alpha-1)} - \frac{2}{3}\Big(\frac{-2\sqrt{2\pi}C\alpha(1-\alpha)}{\sigma(\alpha-1)}\Big)^3\Big)\\
&= \frac{1}{2} + \frac{2C\alpha}{\sigma} - \frac{32\pi C^3\alpha^3}{3\sigma^3}
\end{aligned}$$

and similarly the term $-\Phi\Big(\frac{\log t}{\sigma\alpha}\Big)$ has lower bound

$$-\Phi\Big(\frac{\log t}{\sigma\alpha}\Big) = \Phi\Big(-\frac{\log t}{\sigma\alpha}\Big) - 1 \geq \frac{1}{2} + \frac{2C(1-\alpha)}{\sigma} - \frac{32\pi C^3(1-\alpha)^3}{3\sigma^3} - 1.$$

Therefore,

$$\begin{aligned}
&t\Big[\Phi\Big(\frac{\log t}{(\alpha-1)\sigma}\Big) - \Phi\Big(\frac{\log t}{\alpha\sigma}\Big)\Big] - \frac{2tC}{\sigma}\\
&\geq \frac{2tC}{\sigma} - \frac{32\pi C^3\alpha^3}{3\sigma^3} - \frac{32\pi C^3(1-\alpha)^3}{3\sigma^3} - \frac{2tC}{\sigma} \geq -\frac{32\pi C^3}{3\sigma^3}.
\end{aligned}$$

Let us assume that $\sigma\alpha(1-\alpha) \geq 2C^{3/2}\exp(\sqrt{2\pi}C)$, then we have $\sigma^2\alpha^2(1-\alpha)^2 \geq 4C^3\exp(2\sqrt{2\pi}C)$, so we have

$$\begin{aligned}\frac{32\pi C^3}{3\sigma^3} &\leq \frac{32\pi C^3\alpha^3(1-\alpha)^3}{3\sigma^3\alpha^3(1-\alpha)^3} \leq \frac{\pi C^3}{6\sigma^3\alpha^3(1-\alpha)^3} \\ &\leq \frac{\pi\exp(-2\sqrt{2\pi}C)}{24\sigma\alpha(1-\alpha)} \leq \frac{\exp(-2\sqrt{2\pi}C)}{6\sigma\alpha(1-\alpha)}.\end{aligned} \tag{6.8}$$

We combine (6.7) and (6.8) and have,

$$\mathbb{E}[g_\alpha(\log l(Y))] \geq \frac{\exp(-2\sqrt{2\pi}C)}{5\sigma\alpha(1-\alpha)} - \frac{\exp(-2\sqrt{2\pi}C)}{6\sigma\alpha(1-\alpha)} = \frac{\exp(-2\sqrt{2\pi}C)}{30\sigma\alpha(1-\alpha)}. \qquad \square$$

6.2. Proof of (2.14)

$$\begin{aligned}&D_\alpha(\varphi_{0j}\|\varphi_{1j}) = -\log\int_{\Omega_j}\varphi_{0j}^{1-\alpha}\,\varphi_{1j}^{\alpha}d\mu \\ &= -\log\int_{\Omega_j} h(x)\exp\{[(1-\alpha)\theta_{0j}+\alpha\theta_{1j}]^\top T(x) - (1-\alpha)A(\theta_{0j}) - \alpha A(\theta_{1j})\} \\ &= -\log\left\{\exp[-(1-\alpha)A(\theta_{0j}) - \alpha A(\theta_{1j}) + A(\theta_{\alpha j}))]\int_{\Omega_j}\varphi(x;\theta_{\alpha j})dx\right\} \\ &= (1-\alpha)A(\theta_{0j}) + \alpha A(\theta_{1j}) - A(\theta_{\alpha j}).\end{aligned}$$

6.3. Proof of variance of (2.15)

The variance of Z_j can be directly derived from the following proposition. Its proof is skipped for brevity.

Proposition 2. *A random variable $X \sim \varphi(x;\theta)$ in (2.12) satisfies:*

(a) The moment generating function of $T(X)$, $M_{T(X)}(t) = \exp[A(\theta+t)-A(\theta)]$ if it exists.
(b) $\mathbb{E}[T(X)] = \nabla A(\theta)$ and $Var[T(X)] = \mathbf{H}(A(\theta))$ where $\mathbf{H}(A(\theta))$ is the Hessian matrix of A evaluated at θ.

6.4. Proof of (2.16)

For $\alpha \in (0,1)$, we recall $\theta_{\alpha j} = (1-\alpha)\theta_{0j} + \alpha\theta_{1j}$ and define

$$\begin{aligned}p_{\alpha j} &:= \frac{1}{1+e^{-\theta_{\alpha j}}} = \frac{1}{1+e^{-(1-\alpha)\theta_{0j}-\alpha\theta_{1j}}} \\ &= \frac{1}{1+\left(\frac{1-p_{0j}}{p_{0j}}\right)^{1-\alpha}\left(\frac{1-p_{1j}}{p_{1j}}\right)^{\alpha}} = \frac{p_{0j}^{1-\alpha}p_{1j}^{\alpha}}{p_{0j}^{1-\alpha}p_{1j}^{\alpha} + (1-p_{0j})^{1-\alpha}(1-p_{1j})^{\alpha}}.\end{aligned}$$

It is worth noting that the identity still holds when $\alpha = 0$ or 1, including the next one.

$$\exp(A(\theta_{\alpha j})) = 1 + e^{\theta_{\alpha j}} = 1 + \frac{p_{\alpha j}}{1 - p_{\alpha j}} = \frac{1}{1 - p_{\alpha j}}$$
$$= \frac{p_{0j}^{1-\alpha} p_{1j}^{\alpha} + (1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha}}{(1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha}}.$$

By (2.14), we can write the Chernoff coefficient for p_{0j} and p_{1j} in terms of $p_{\alpha j}$'s:

$$\begin{aligned} e^{-D_\alpha(\varphi_{0j}\|\varphi_{1j})} &= e^{-(1-\alpha)A(\theta_{0j}) - \alpha A(\theta_{1j}) + A(\theta_{\alpha j})} \\ &= (1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha} \frac{p_{0j}^{1-\alpha} p_{1j}^{\alpha} + (1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha}}{(1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha}} \\ &= p_{0j}^{1-\alpha} p_{1j}^{\alpha} + (1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha}. \end{aligned}$$

Hence,

$$e^{-D_\alpha(\varphi_0\|\varphi_1)} = e^{-\sum_{j=1}^{n} D_\alpha(p_{0j}\|p_{1j})} = \prod_{j=1}^{n} [p_{0j}^{1-\alpha} p_{1j}^{\alpha} + (1 - p_{0j})^{1-\alpha}(1 - p_{1j})^{\alpha}].$$

6.5. *Proof of* (2.17)

We recall the definition of Y_j from (2.9), and have $Y_j \sim \mathrm{Bern}(p_{\alpha j})$. Then by (2.15), and letting $\alpha = \alpha^*$,

$$Z_j - \mathbb{E}[Z_j] = (Y_j - p_{\alpha^* j}) \log \frac{p_{0j}(1 - p_{1j})}{p_{1j}(1 - p_{0j})}$$

with variance

$$\mathrm{Var}[Z_j] = (\theta_{0j} - \theta_{1j})^2 \mathrm{Var}(Y_j) = \left[\log \frac{p_{0j}(1 - p_{1j})}{p_{1j}(1 - p_{0j})}\right]^2 p_{\alpha j}(1 - p_{\alpha j}).$$

Summing over $j \in [n]$ of $\mathrm{Var}[Z_j]$ gives (2.17).

6.6. *The rest of proof of Bernoulli example in Section 2.3*

To show the upper and lower bound, it remains to show that

$$\max_{j\in[n]} \left| \log \frac{p_{0j}(1 - p_{1j})}{p_{1j}(1 - p_{0j})} \right| \le C_1$$

is a sufficient condition of Theorem 2.1. We have

$$\mathbb{E}[|Z_j|^3] = \left| \log \frac{p_{0j}(1 - p_{1j})}{p_{1j}(1 - p_{0j})} \right|^3 [(1 - p_{\alpha j}) p_{\alpha j}^3 + p_{\alpha j}(1 - p_{\alpha j})^3]$$

$$\leq \Big| \log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})} \Big|^3 p_{\alpha j}(1-p_{\alpha j}).$$

Suppose $\max_{j\in[n]} \Big| \log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})} \Big| \leq C_1$, then

$$\begin{aligned} \frac{\sum_{j=1}^n \mathbb{E}[|Z_j|^3]}{\sum_{j=1}^n \mathrm{Var}[Z_j]} &= \frac{\sum_{j=1}^n \Big| \log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})} \Big|^3 p_{\alpha j}(1-p_{\alpha j})}{\sum_{j=1}^n \Big[\log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})} \Big]^2 p_{\alpha j}(1-p_{\alpha j})} \\ &\leq \max_{j\in[n]} \Big| \log \frac{p_{0j}(1-p_{1j})}{p_{1j}(1-p_{0j})} \Big| \leq C_1. \end{aligned}$$

This implies $\sum_{j=1}^n \mathbb{E}|Z_j|^3 \leq C_1 n\bar{\sigma}_n^2$.

6.7. Proof of (2.18)

We recall the definition of total variation affinity η from (2.4), and have

$$\begin{aligned} &\eta(\varphi_0,\varphi_1) = \sum_{x\in\{0,1\}^n} \min\Big(\prod_{j=1}^n \bar{p}_0^{x_j}(1-\bar{p}_0)^{1-x_j}, \prod_{j=1}^n \bar{p}_1^{x_j}(1-\bar{p}_1)^{1-x_j} \Big) \\ &= \sum_{x\in\{0,1\}^n} \min\Big(\bar{p}_0^{\sum_{j=1}^n x_j}(1-\bar{p}_0)^{n-\sum_{j=1}^n x_j}, \bar{p}_1^{\sum_{j=1}^n x_j}(1-\bar{p}_1)^{n-\sum_{j=1}^n x_j} \Big) \\ &= \sum_{y=0}^n \binom{n}{y} \min(\bar{p}_0^y(1-\bar{p}_0)^{n-y}, \bar{p}_1^y(1-\bar{p}_1)^{n-y}) \\ &= \eta(\psi_0,\psi_1). \end{aligned}$$

7. Proofs of Section 3

7.1. Proof of Lemma 1

It is required to check the assumptions in Theorem 2.1 are satisfied. Firstly, we need to check that $\sum_{i=1}^n \mathbb{E}[|Z_j|^3] \lesssim \sum_{i=1}^n \mathrm{Var}[Z_j]$. We use the notation in Section 2.3, and replace p_{0*} and p_{1*} by p_{k*} and $p_{\ell*}$, under (3.3), we have

$$\frac{\sum_{i=1}^n \mathbb{E}[|Z_j|^3]}{\sum_{i=1}^n \mathrm{Var}[Z_j]} \leq \max_{j\in[n]} \Big| \log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big| \leq \Big| \log \frac{\omega}{\varepsilon} \Big|.$$

Secondly, by Lemma 5, we can remove $\alpha^*(1-\alpha^*)$ since it is bounded below by constant and bounded above by $1/4$. Thirdly, by Lemma 4, $n\bar{\sigma}_n^2$ in (2.17) is sufficiently large under the assumption $np^* \leq C(D_{\alpha^*}(p_{k*}\|p_{\ell*}))^2$. Furthermore, we need to check that $n\bar{\sigma}_n^2$ can be replaced by

$$\max_{j\in[n]} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 np^*$$

up to a constant factor. For all k, ℓ and j, $p_{\alpha^* j} = p_{kj}^{1-\alpha^*} p_{\ell j}^{\alpha^*} \le p^*$, so we have

$$\sum_{j=1}^{n} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 p_{\alpha^* j}(1-p_{\alpha^* j}) \le \max_{j\in[n]} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 np^*. \quad (7.1)$$

Under the assumption of the block structure and (3.3), for at least $\frac{n}{\beta K}$ many $i \in [n]$,

$$\Big[\log \frac{p_{ki}(1-p_{\ell i})}{p_{\ell i}(1-p_{ki})} \Big]^2 = \max_{j\in[n]} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2.$$

Combining with $p_{\alpha^* j}(1-p_{\alpha^* j}) \ge \frac{p^*\varepsilon}{\omega}$, we have

$$\sum_{j=1}^{n} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 p_{\alpha^* j}(1-p_{\alpha^* j}) \ge \max_{j\in[n]} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 \frac{np^*\varepsilon}{\omega\beta K}. \quad (7.2)$$

Now we combine (7.1) and (7.2) and obtain

$$\sum_{j=1}^{n} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 p_{\alpha^* j}(1-p_{\alpha^* j}) \asymp \max_{j\in[n]} \Big[\log \frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})} \Big]^2 np^*.$$

Finally, by Lemma 5, we can remove $\alpha^*(1-\alpha^*)$ since it is bounded below by constant and bounded above by 1/4.

7.2. Auxiliary lemmas for proof of Lemma 1

Lemma 4. *Under the assumption (3.3), for any C_1, there exists C_2 only depends on β, ε, K and ω such that if $D_{\alpha^*}(p_{k*}\|p_{\ell *})^2 \ge C_2 np^*$, then $\sqrt{n}\bar{\sigma}_n \alpha^*(1-\alpha^*) \ge C_1$, where $\bar{\sigma}_n$ is defined in (2.17). In particular, we can choose large enough C_2 so that C_1 is also sufficiently large.*

Proof. We briefly write $\alpha = \alpha^*$ in this proof. We recall that

$$D_\alpha(p_{k*}\|p_{\ell *}) = \sum_{j=1}^{n} -\log(p_{kj}^{1-\alpha} p_{\ell j}^{\alpha} + (1-p_{kj})^{1-\alpha}(1-p_{\ell j})^{\alpha}),$$

so there exists $j \in [n]$ such that

$$-\log(p_{kj}^{1-\alpha} p_{\ell j}^{\alpha} + (1-p_{kj})^{1-\alpha}(1-p_{\ell j})^{\alpha}) \ge \frac{D_\alpha(p_{k*}\|p_{\ell *})}{n}.$$

In this proof, we briefly denote $p_0 := p_{kj}$, $p_1 := p_{\ell j} = p_1$ and $a := \frac{D_\alpha(p_{k*}\|p_{\ell *})}{n}$. Without loss of generality, we assume $p_1 \ge p_0$. Then the inequality above implies

$$e^{-a} \ge p_0^{1-\alpha} p_1^{\alpha} + (1-p_0)^{1-\alpha}(1-p_1)^{\alpha} \ge p_1 \Big(\frac{p_0}{p_1}\Big)^{1-\alpha} + 1 - p_1.$$

By straightforward rearrangement, we have

$$(1-\alpha)\log\frac{p_1}{p_0} \geq \log p_1 - \log(p_1 - (1-e^{-a})) \geq \frac{1-e^{-a}}{p_1} \geq \frac{a - a^2/2}{p_1}$$

where the second inequality is due to the derivative of $\log(p_1 + x)$ is at least $1/p_1$ on $(-p_1, 0]$, and the last inequality uses the fact that $1 - e^{-x} \geq x - \frac{x^2}{2}$ for $x \geq 0$. Therefore,

$$(1-\alpha)\log\frac{p_1}{p_0} \geq \begin{cases} \frac{1-e^{-a}}{p_1} \geq \frac{1-e^{-1}}{p_1}, & \text{if } a \geq 1; \\ \frac{a-a^2/2}{p_1} \geq \frac{a}{2p_1}, & \text{if } a < 1. \end{cases}$$

We recall that we assume $p_1 \geq p_0$, so $\log\frac{1-p_0}{1-p_1} \geq 0$. By (7.2),

$$\begin{aligned} n\bar{\sigma}_n^2 &\gtrsim \Big[\log\frac{p_1(1-p_0)}{p_0(1-p_1)}\Big]^2 np^* \\ &= \Big(\log\frac{p_1}{p_0} + \log\frac{1-p_0}{1-p_1}\Big)^2 np^* \\ &\geq \frac{1}{(1-\alpha)^2}\Big(\frac{1-e^{-1}}{p_1} \wedge \frac{a}{2p_1}\Big)^2 np^*. \end{aligned}$$

Since we choose $\alpha = \alpha^*$ to be optimal, then by Lemma 5, $\frac{1}{(1-\alpha)^2}$ is bounded below by constant. Now it suffices to show both $\frac{(1-e^{-1})^2 np^*}{p_1^2}$ and $\frac{a^2 np^*}{4p_1^2}$ are bounded below by constant. Under the assumption $D_{\alpha^*}(p_{k*}\|p_{\ell*})^2 \geq C_2 np^*$, by Lemma 10 and using C_ε in that lemma,

$$np^* \geq \frac{D_{\alpha^*}(p_{k*}\|p_{\ell*})}{C_\varepsilon} \geq \frac{\sqrt{C_2 np^*}}{C_\varepsilon},$$

which implies $np^* \geq C_2/C_\varepsilon^2$. Now we consider the first term and have

$$\frac{(1-e^{-1})^2 np^*}{p_1^2} \geq (1-e^{-1})^2 np^* \geq \frac{C_2(1-e^{-1})^2}{C_\varepsilon^2}.$$

By choosing sufficiently large C_2, the RHS is sufficiently large. For the second term, we recall $a = \frac{D_\alpha(p_{k*}\|p_{\ell*})}{n}$ and the assumption that $p_1 \leq p^*$, so

$$\frac{a^2 np^*}{4p_1^2} = \frac{D_\alpha(p_{k*}\|p_{\ell*})^2 np^*}{4n^2 p_1^2} \geq \frac{D_\alpha(p_{k*}\|p_{\ell*})^2}{4np^*} \geq \frac{C_2}{4}.$$

This term is also sufficiently large by chosen a big enough C_2. □

Lemma 5 (Bounds of α^*)**.** *Under the setting in Section 2.3, suppose*

$$\max_{j\in[n]}\Big(\frac{p_{0j}}{p_{1j}} \vee \frac{p_{1j}}{p_{0j}}\Big) \leq \omega, \max_{j\in[n]}(p_{0j} \vee p_{1j}) \leq 1-\varepsilon \text{ and } \alpha^* = \arg\max_{\alpha\in[0,1]} D_\alpha(p_{0*}\|p_{1*})$$

for $\omega > 1$ and $\varepsilon \in (0,1)$, then there exists $\delta \in (0, 1/2)$ which only depends on ε and ω such that $\alpha^* \in [\delta, 1-\delta]$.

Proof. We first consider the case $n = 1$ and briefly denote $p_{01} := p$ and $p_{11} := q$ (in this proof only). Let $f(\alpha) = p^{1-\alpha}q^{\alpha} + (1-p)^{1-\alpha}(1-q)^{\alpha}$, then

$$f'(\alpha) = p^{1-\alpha}q^{\alpha}\log\frac{q}{p} + (1-p)^{1-\alpha}(1-q)^{\alpha}\log\frac{1-q}{1-p}.$$

Since f is smooth and convex, α^* minimize $f(\alpha)$ if and only if $f'(\alpha^*) = 0$. Let us define $x := \log\frac{q}{p}$ and $y := \log\frac{1-q}{1-p}$, then $p = \frac{1-e^y}{e^x-e^y}$ and $1-p = \frac{e^x-1}{e^x-e^y}$. Without loss of generality, we assume $p > q$, so $x < 0$ and $y > 0$. Hence

$$f'(\alpha) = \frac{1-e^y}{e^x-e^y}xe^{\alpha x} + \frac{e^x-1}{e^x-e^y}ye^{\alpha y}.$$

$f'(\alpha^*) = 0$ implies

$$\alpha^* = \frac{\log\frac{e^y-1}{y} - \log\frac{e^x-1}{x}}{y-x}.$$

Let

$$g(z) = \begin{cases} 0, & \text{if } z = 0; \\ \log\frac{e^z-1}{z} & \text{otherwise.} \end{cases}$$

We can observe that g is a strictly increasing smooth function on $\mathbb{R}$, and $g' \in (0,1)$. α^* is the slope of a secant line that intersects the function g at x and y, so α^* can only take value $g'(z)$ for some $z \in [x,y]$. Since $x \in [-\log\omega,\ \log\omega]$ and $y \in [1-\varepsilon, \frac{1}{1-\varepsilon}]$, there exists δ which only depends on ω and ε such that $\alpha^* \in [\delta, 1-\delta]$. Now we can generalize the conclusion to $n > 1$. Let

$$f(\alpha) := \prod_{i=1}^{n} f_j(\alpha) := \prod_{i=1}^{n}[p_{0j}^{1-\alpha}p_{1j}^{\alpha} + (1-p_{0j})^{1-\alpha}(1-p_{1j})^{\alpha}]$$

Since each positive convex function f_j is decreasing on $[0,\delta]$ and increasing on $[1-\delta, 1]$, so is their product pointwise f. Therefore, f achieves minimum on $[\delta, 1-\delta]$. $\square$

7.3. Proof of Theorem 3.1

It suffices to replace Lemma 5.2 in [36] with the Chernoff lower bound in Section 2.3. Let $n' = \lfloor n/K \rfloor$,

$$p_{0*} = (\underbrace{p,p,\dots,p}_{n'\text{ times}},\underbrace{q,q,\dots,q}_{n'\text{ times}}), \quad \text{and} \quad p_{1*} = (\underbrace{q,q,\dots,q}_{n'\text{ times}},\underbrace{p,p,\dots,p}_{n'\text{ times}}).$$

Let $X_i \sim \mathrm{Bern}(p)$ and $Y_i \sim \mathrm{Bern}(q)$ for $i \in [n']$. For sufficiently large np,

$$\mathbb{P}\Big(\sum_{i=1}^{n'} Y_i - X_i \geq 0\Big) = \mathbb{P}(\varphi((X,Y);p_{0*}) \leq \varphi((X,Y);p_{1*}))$$

$$
\begin{aligned}
&= \sum_{x\in\{0,1\}^{2n'}} \varphi(x;p_{0*})1\{\varphi(x;p_{0*}) \le \varphi(x;p_{1*})\} \\
&\ge \frac{1}{2} \sum_{x\in\{0,1\}^{2n'}} \min(\varphi(x;p_{0*}),\varphi(x;p_{1*})) \\
&\gtrsim \Big(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\Big)^{-1} \exp(-D_{\alpha^*}(p_{0*}\|p_{1*})) \\
&= \Big(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\Big)^{-1} (\sqrt{pq}+\sqrt{(1-p)(1-q)})^{2n'} \\
&\ge \Big(\sqrt{np}\log\frac{p(1-q)}{q(1-p)}\Big)^{-1} (\sqrt{pq}+\sqrt{(1-p)(1-q)})^{2n/K}.
\end{aligned}
$$

This inequality provides a lower bound of $B_\tau(\hat\sigma(1))$ in [36]. The rest of proofs follow from the arguments in [36].

7.4. Auxiliary lemmas for proof of Theorem 3.2

In this section, we will use the following concentration inequality [47, p. 118]:

Proposition 3 (Prokhorov)**.** *Let $S=\sum_i X_i$ for independent centered variables $\{X_i\}$, each bounded by $c<\infty$ in absolute value a.s. and suppose $v \ge \sum_i \mathbb{E}X_i^2$, then for $t>0$,*

$$
\mathbb{P}\big(S > vt\big) \le \exp[-vh_c(t)], \quad \text{where } h_c(t) := \frac{3}{4c}t\log\big(1+\frac{2c}{3}t\big). \tag{7.3}
$$

Same bound holds for $\mathbb{P}(S<-vt)$.

Lemma 6 (Uniform Parameter Estimation)**.** *For $\hat P$ obtained from the operation $\mathcal{B}(A,\tilde z)$, and assuming* $\mathrm{Mis}(\tilde z, z) \le \gamma$ *for $\frac{1}{n} \le \gamma \le \frac{1}{2\beta K}$ with optimal permutation $\pi^* = id$, we have*

$$
\begin{aligned}
\mathbb{P}\Big(\sup\{\|\hat P - P\|_\infty : \sum_{i=1}^n 1\{\tilde z_i \ne z_i\} \le n\gamma\} > C(8\beta K\gamma+\tau)p^*\Big) \\
\le \exp\Big[-\frac{n^2p^*h_1(\tau)}{8\beta^2K^2} - 2n\gamma\log\gamma\Big].
\end{aligned}
$$

for every $\tau>0$. If $\gamma<\frac{1}{n}$, we can replace $n\gamma\log\gamma$ by 0.

Proof. We only consider the case $k=\ell$. If $k\ne\ell$, the arguments will similarly follow. Let $E=\{(i,j):\tilde z_i = z_i = \tilde z_j = z_j = k\}$, $F=\{(i,j):\tilde z_i=\tilde z_j = k,$ but $z_i \ne k$ or $z_j\ne k\}$. Let $\hat n_k = |\{i\in[n]:\tilde z_i = k\}|$. According to assumptions,

$$
(\hat n_k - n\gamma)^2 \le |E| \le \hat n_k^2 \quad \text{and} \quad |F| \le 2n\gamma\hat n_k.
$$

Hence by definition of $\hat P_{k\ell}$ from (3.10), we have upper bound

$$
\hat n_k^2\mathbb{E}[\hat P_{k\ell}] \le |E|P_{k\ell} + |F|p^* \le \hat n_k^2 P_{k\ell} + 2n\gamma\hat n_k p^*
$$

Since $\hat{n}_k \ge n_k - n\gamma \ge \frac{n}{\beta K} - \frac{n}{2\beta K} = \frac{n}{2\beta K}$, so

$$\mathbb{E}[\hat{P}_{k\ell}] \le \frac{\hat{n}_k^2 P_{k\ell} + 2n\gamma \hat{n}_k p^*}{\hat{n}_k^2} \le P_{k\ell} + 4\beta K \gamma p^*.$$

For lower bound, we have

$$\begin{aligned}\hat{n}_k^2 \mathbb{E}[\hat{P}_{k\ell}] &\ge (\hat{n}_k - n\gamma)(\hat{n}_k - n\gamma - 1) P_{k\ell} \ge \hat{n}_k \hat{n}_\ell P_{k\ell} - 2n\gamma(\hat{n}_k + 1) p^* \\ &\ge \hat{n}_k \hat{n}_\ell P_{k\ell} - 4n\gamma \hat{n}_k p^*.\end{aligned}$$

Therefore, using $\hat{n}_k \ge \frac{n}{2\beta K}$ again, we have

$$\mathbb{E}[\hat{P}_{k\ell}] \ge \frac{\hat{n}_k^2 P_{k\ell} - 4n\gamma \hat{n}_k p^*}{\hat{n}_k^2} \ge P_{k\ell} - 8\beta K \gamma p^*.$$

Thus $|P_{k\ell} - \mathbb{E}[\hat{P}_{k\ell}]| \le 8\beta K \gamma p^*$. By Proposition 3,

$$\begin{aligned}\mathbb{P}(|\hat{P}_{k\ell} - \mathbb{E}[\hat{P}_{k\ell}]| \ge \tau p^*) &= \mathbb{P}\Big(\frac{\hat{n}_k(\hat{n}_k - 1)}{2}(\hat{P}_{k\ell} - \mathbb{E}[\hat{P}_{k\ell}]) \ge \frac{\hat{n}_k(\hat{n}_k - 1)}{2}\tau p^*\Big) \\ &\le 2\exp\Big[-\frac{\hat{n}_k(\hat{n}_k - 1)}{2} p^* h_1(\tau)\Big] \le 2\exp\Big[-\frac{n^2 p^* h_1(\tau)}{8\beta^2 K^2}\Big].\end{aligned}$$

There are at most

$$\sum_{i=0}^{\lfloor n\gamma \rfloor} \binom{n}{i} \le \binom{n}{\lfloor n\gamma \rfloor} \sum_{j=0}^{\infty} \Big(\frac{\lfloor n\gamma \rfloor}{n - \lfloor n\gamma \rfloor + 1}\Big)^j.$$

many different $\tilde{z}$ with error rate at most γ. We consider the function $f(x) = \left(\frac{a}{x}\right)^x$ which is increasing when $x \le a/e$. By the fact that $\lfloor n\gamma \rfloor \le n\gamma \le n$, the first term on the RHS has upper bound

$$\binom{n}{\lfloor n\gamma \rfloor} \le \Big(\frac{en}{\lfloor n\gamma \rfloor}\Big)^{\lfloor n\gamma \rfloor} \le \Big(\frac{en}{n\gamma}\Big)^{n\gamma} = \exp(-n\gamma \log \gamma).$$

The second term on the RHS is a geometric sum, using $1 \le \lfloor n\gamma \rfloor \le \frac{n}{2\beta K} \le \frac{n}{4}$, we have

$$\begin{aligned}\sum_{j=0}^{\infty} \Big(\frac{\lfloor n\gamma \rfloor}{n - \lfloor n\gamma \rfloor + 1}\Big)^j &= \Big(1 - \frac{\lfloor n\gamma \rfloor}{n - \lfloor n\gamma \rfloor + 1}\Big)^{-1} \\ &= \frac{n - \lfloor n\gamma \rfloor + 1}{n - 2\lfloor n\gamma \rfloor + 1} \le \frac{n}{n - 2\lfloor n\gamma \rfloor} \le 2.\end{aligned}$$

If $\gamma < \frac{1}{n}$, then $\tilde{z} = z$ is unique. Taking the union bound, we obtain the desired probability. $\square$

Let $\varphi(x; p)$ be the PMF evaluated at x of a Poisson-Binomial variable with parameters $p = (p_1, \ldots, p_n)$. In particular, if $p = \bar{p}\mathbf{1}_n$, then $\varphi(x, p)$ is the PMF of a binomial distribution with parameters n and $\bar{p}$.

Lemma 7 (Binomial Perturbation). *If* $|\bar{p}_1 - \bar{p}_2| \le \delta \max(\bar{p}_1, \bar{p}_2) := \delta p^*$, $p^*/\omega \le \bar{p}_1, \bar{p}_2 \le 1 - \varepsilon$, *then*

$$\frac{\varphi(x; \bar{p}_1 \mathbf{1}_n)}{\varphi(x; \bar{p}_2 \mathbf{1}_n)} \le \exp\Big(\delta\omega x + \frac{\delta n p^*}{\varepsilon}\Big) \quad \forall x \in \mathbb{Z}_+.$$

Proof. Using $1 + x \le e^x$ several times, we have

$$\begin{aligned}
\frac{\varphi(x; \bar{p}_1 \mathbf{1}_n)}{\varphi(x; \bar{p}_2 \mathbf{1}_n)} &= \Big(\frac{\bar{p}_1}{\bar{p}_2}\Big)^x \frac{(1-\bar{p}_1)^{n-x}}{(1-\bar{p}_2)^{n-x}} \\
&\le \Big(\frac{\bar{p}_2 + \delta p^*}{\bar{p}_2}\Big)^x \Big(\frac{1 - \bar{p}_2 + \delta p^*}{1 - \bar{p}_2}\Big)^{n-x} \\
&\le (1 + \delta\omega)^x \Big(1 + \frac{\delta p^*}{\varepsilon}\Big)^{n-x} \\
&\le \exp\Big(\delta\omega x + \frac{\delta n p^*}{\varepsilon}\Big).
\end{aligned}$$

as desired. □

Lemma 8 (Poisson-Binomial Approximation). *Let* $p = (p_1, \dots, p_n)$ *be parameter of a Poisson binomial distribution. Let* $p^* := \max_{i \in [n]} p_i$. *We assume at least* $n(1-\gamma)$ *entries of* p *are exactly* $\bar{p}$, *and* $p^* \le \min(1 - \varepsilon, \omega\bar{p})$. *Then,*

$$\frac{\varphi(x; p)}{\varphi(x; \bar{p}\mathbf{1}_n)} \le \exp\Big(\frac{\gamma}{\varepsilon}(n\bar{p} + \omega x)\Big), \quad \forall x \in \mathbb{Z}_+.$$

Proof. Let $\mathcal{S}(x) = \{S \subset [n] : |S| = x\}$ for $x \in \mathbb{Z}_+$, then

$$\varphi(x; p) = \prod_{i=1}^n (1 - p_i) \sum_{S \in \mathcal{S}(x)} \prod_{j \in S} \frac{p_j}{1 - p_j}$$

By Maclaurin's inequality,

$$\sum_{S \in \mathcal{S}(x)} \prod_{j \in S} \frac{p_j}{1 - p_j} \le \binom{n}{x} \frac{1}{n^x} \Big(\sum_{i=1}^n \frac{p_i}{1 - p_i}\Big)^x,$$

so we have

$$\begin{aligned}
\frac{\varphi(x; p)}{\varphi(x; \bar{p}\mathbf{1}_n)} &\le \frac{\prod_{i=1}^n (1-p_i) \sum_{S \in \mathcal{S}(x)} \prod_{j \in S} \frac{p_j}{1-p_j}}{\binom{n}{x} \bar{p}^x (1-\bar{p})^x} \\
&\le \frac{\prod_{i=1}^n (1-p_i) \frac{1}{n^x} \Big(\sum_{i=1}^n \frac{p_i}{1-p_i}\Big)^x}{\bar{p}^x (1-\bar{p})^{n-x}}.
\end{aligned}$$

Without loss of generality, we assume $p_{\lfloor n\gamma \rfloor + 1} = \cdots = p_n = \bar{p}$. We have

$$\frac{\prod_{i=1}^n (1-p_i) \frac{1}{n^x} \Big(\sum_{i=1}^n \frac{p_i}{1-p_i}\Big)^x}{\bar{p}^x (1-\bar{p})^{n-x}} = \Big(\prod_{i=1}^n \frac{1-p_i}{1-\bar{p}}\Big) \Big(\frac{1}{n} \sum_{i=1}^n \frac{p_i(1-\bar{p})}{\bar{p}(1-p_i)}\Big)^x$$

$$= \Big(\prod_{i=1}^{\lfloor n\gamma\rfloor}\frac{1-p_i}{1-\bar{p}}\Big)\Big(1-\frac{\lfloor n\gamma\rfloor}{n}+\frac{1}{n}\sum_{i=1}^{\lfloor n\gamma\rfloor}\frac{p_i(1-\bar{p})}{\bar{p}(1-p_i)}\Big)^x$$
$$\le \frac{1}{(1-\bar{p})^{n\gamma}}\Big(1+\frac{1}{n}\sum_{i=1}^{\lfloor n\gamma\rfloor}\frac{p_i}{\bar{p}(1-p_i)}\Big)^x.$$

By the inequality $\frac{1}{1-x}\le \exp\left(\frac{x}{1-x}\right)$ for $x\in(0,1)$, we have

$$\frac{1}{(1-\bar{p})^{n\gamma}}\le \exp\left(\frac{n\gamma\bar{p}}{1-\bar{p}}\right)\le \exp\left(\frac{n\gamma\bar{p}}{\varepsilon}\right).$$

For the other term, $1+x\le e^x$ implies

$$\Big(1+\frac{1}{n}\sum_{i=1}^{\lfloor n\gamma\rfloor}\frac{p_i(1-\bar{p})}{\bar{p}(1-p_i)}\Big)^x\le \exp\left(\frac{x\gamma p^*}{\bar{p}(1-p^*)}\right)\le \exp\left(\frac{\gamma\omega x}{\varepsilon}\right).$$

Therefore, we have

$$\frac{\varphi(x;p)}{\varphi(x;\bar{p}\mathbf{1}_n)}\le \frac{\prod_{i=1}^n(1-p_i)\frac{1}{n^x}\left(\sum_{i=1}^n\frac{p_i}{1-p_i}\right)^x}{\bar{p}^x(1-\bar{p})^{n-x}}\le \exp\left(\frac{\gamma}{\varepsilon}(n\bar{p}+\omega x)\right). \qquad \square$$

Lemma 9 (Degree Truncation)**.** *For fixed $i\in[n]$, let $b_{i+}=\sum_{r\in[K]}b_{ir}=\sum_{j=1}^n A_{ij}$ be the degree of node i, where A is the adjacency matrix in SBM (see (3.1)), and assuming $\max_{j\in[n]}\mathbb{E}[A_{ij}]\le p^*\le 1-\varepsilon$. Then there exists $C_\varepsilon>0$, which only depends on ε such that*

$$\mathbb{P}(b_{i+}>C_\varepsilon np^*)\le \exp(-np^*(1+\log\varepsilon^{-1})).$$

Proof of Lemma 9. We choose large enough C_ε such that

$$(C_\varepsilon-1)\log\left(1+\frac{2(C_\varepsilon-1)}{3}\right)\ge 1+\log\varepsilon^{-1}.$$

Now we want to find the upper bound of the following probability:

$$\mathbb{P}\big(b_{i+}>C_\varepsilon np^*\big)\le \mathbb{P}\big(b_{i+}-\mathbb{E}[b_{i+}]>(C_\varepsilon-1)np^*\big).$$

For fixed i, let $p_j=A_{ij}$, and $v=\sum_{j=1}^n p_j(1-p_j)$, $vt=(C_\varepsilon-1)np^*$, so $t\ge C_\varepsilon-1$. By Proposition 3, we have

$$\begin{aligned}\mathbb{P}\big(b_{i+}-\mathbb{E}[b_{i+}]>(C_\varepsilon-1)np^*\big)&\le \exp\left[-\frac{3}{4}vt\log\left(1+\frac{2t}{3}\right)\right]\\&\le \exp\left[-(C_\varepsilon-1)\log\left(1+\frac{2(C_\varepsilon-1)}{3}\right)np^*\right]\\&\le \exp(-np^*(\log\varepsilon^{-1})),\end{aligned}$$

where the last inequality holds by the choice of C_ε. $\square$

Lemma 10. *Recall the Chernoff information $D_\alpha(p_{k*}\|p_{\ell*})$ between Bernoulli distribution from (2.16), assuming $\max_i \max(p_{ki}, p_{\ell i}) = p^* \le 1-\varepsilon$, then we have $D_\alpha(p_{k*}\|p_{\ell*}) \le C_\varepsilon np^*$ where C_ε only depends on ε.*

Proof. For $\alpha \in [0,1]$, we have

$$\begin{aligned}
D_\alpha(p_{k*}\|p_{\ell*}) &= \sum_{j=1}^n -\log[p_{kj}^{1-\alpha}p_{\ell j}^{\alpha} + (1-p_{kj})^{1-\alpha}(1-p_{\ell j})^{\alpha}] \\
&\le \sum_{j=1}^n -\log[(1-p_{kj})^{1-\alpha}(1-p_{\ell j})^{\alpha}] \\
&= \sum_{j=1}^n -(1-\alpha)\log(1-p_{kj}) - \alpha\log(1-p_{\ell j}) \\
&\le -n\log(1-p^*) \\
&\le np^*(\log\varepsilon^{-1}),
\end{aligned}$$

where the last inequality uses $p^* \le 1-\varepsilon$. □

Lemma 11 (Perturbed Likelihood Ratio Test)**.** *We assuming the parameters satisfies (3.3), given the ith row of the adjacency matrix A_{i*} and $\hat{z}_j$ for $j\in[n]$ such that $\sum_{j=1}^n 1\{\hat{z}_j \ne z_j\} \le n\gamma \le \frac{n}{2\beta K}$, and let $B(\rho) = \{\tilde{P} : \|P-\tilde{P}\|_\infty \le \rho\}$. Assuming $\rho \le \varepsilon/2$ and $D_\alpha(p_{k*}\|p_{\ell*})^2 \ge C_1 np^*$ for some sufficiently large C_1, then the likelihood ratio test variable*

$$Y_{ik\ell} := Y_{ik\ell}(\hat{P},\hat{z}) := \sum_{j\ne i} A_{ij}\log\frac{\hat{P}_{\ell\hat{z}_j}}{\hat{P}_{k\hat{z}_j}} + (1-A_{ij})\log\frac{1-\hat{P}_{\ell\hat{z}_j}}{1-\hat{P}_{k\hat{z}_j}} \tag{7.4}$$

satisfies

$$\mathbb{P}(\exists\hat{P}\in B(\rho), Y_{ik\ell}(\hat{P},\hat{z})\ge 0) \lesssim \exp\Big(C_2\Big(\delta+\gamma+\frac{1}{n}\Big)np^*\Big)\eta(p_{k*},p_{\ell*}).$$

Proof. Firstly, we define the following probability mass functions:

$$\begin{aligned}
&\bar{\psi}_0 \sim \bigotimes_{r=1}^K \mathrm{Bin}(\hat{n}_r, P_{kr}), \quad \hat{\psi}_0 \sim \bigotimes_{r=1}^K \mathrm{Bin}(\hat{n}_r, P_{kr}+\rho), \\
&\hat{\psi}_0 \sim \bigotimes_{r=1}^K \mathrm{Bin}(\hat{n}_r, P_{kr}-\rho) \quad \text{and} \quad \hat{\psi}_1 \sim \bigotimes_{r=1}^K \mathrm{Bin}(\hat{n}_r, P_{\ell r}+\rho),
\end{aligned} \tag{7.5}$$

where $\hat{n}_r = \sum_{j\ne i}^n 1\{z_j = r\}$. Let $b_{ir} := \sum_{j=1}^n A_{ij}1\{\hat{z}_j = r\}$, then we have

$$\begin{aligned}
\sup_{\hat{P}\in B(\rho)} Y_{ik\ell} &= \sup_{\hat{P}\in B(\rho)} \sum_{r=1}^K b_{ir}\log\frac{\hat{P}_{\ell r}}{\hat{P}_{kr}} + (\hat{n}_r - b_{ir})\log\frac{1-\hat{P}_{\ell r}}{1-\hat{P}_{kr}} \\
&\le \sum_{r=1}^K b_{ir}\log\frac{P_{\ell r}+\rho}{P_{kr}-\rho} + (\hat{n}_r - b_{ir})\log\frac{1-P_{\ell r}+\rho}{1-P_{kr}-\rho}
\end{aligned}$$

We have $P_{kr} \le 1-\varepsilon$, so $1-P_{kr} \ge \varepsilon$. Let $\varepsilon \le a, b \le 1$, then for $x \in [-\varepsilon/2, \varepsilon/2]$,

$$\frac{d}{dx}\log\frac{a+x}{b-x} = \frac{1}{a+x} + \frac{1}{b+x} \le \frac{2}{\varepsilon} + \frac{2}{\varepsilon} = \frac{4}{\varepsilon}.$$

Therefore, $\log\frac{a+x}{b-x}$ is a $\frac{4}{\varepsilon}$-Lipschitz function. As a result,

$$\log\frac{1-P_{\ell r}+\rho}{1-P_{kr}-\rho} \le \log\frac{1-P_{\ell r}-\rho}{1-P_{kr}+\rho} + \frac{4\rho}{\varepsilon}.$$

Hence, using $\sum_{r=1}^K(\hat{n}_r - b_{ir}) \le n$, we have

$$\begin{aligned}\sup_{\hat{P}\in B(\rho)} Y_{ik\ell} &\le \frac{4n\rho}{\varepsilon} + \sum_{r=1}^K b_{ir}\log\frac{P_{\ell r}+\rho}{P_{kr}-\rho} + (\hat{n}_r - b_{ir})\log\frac{1-P_{\ell r}-\rho}{1-P_{kr}+\rho} \\ &= \frac{4n\rho}{\varepsilon} + \log\frac{\hat{\psi}_1(x)}{\hat{\psi}_0(x)}.\end{aligned} \tag{7.6}$$

Now we consider the tail bound of $Y_{ik\ell}$. The only random variable on the RHS of the equation above is b_{ir} for $r \in [K]$. Let $\tilde{\psi}_0$ be the probability mass function of $(b_{ir}) \in \mathbb{Z}_+^K$. If $z_i \ne r$, then b_{ir} follows a Poisson binomial distribution with at least $\hat{n}_r - n\gamma$ parameters equal to P_{kr}. If $z_i = r$, then $\hat{n}_r - n\gamma$ need to be replaced by $\hat{n}_r - n\gamma - 1$ in the previous sentence. Since $\gamma \le \frac{1}{2\beta K}$, so $\hat{n}_r \ge \frac{n}{\beta K} - \frac{n}{2\beta K} = \frac{n}{2\beta K}$. The proportion of parameters different from P_{kr} is at most

$$\gamma_r := \frac{n\gamma+1}{\hat{n}_r} \le (n\gamma+1)\cdot\frac{2\beta K}{n} \le 2\beta K\Big(\gamma + \frac{1}{n}\Big).$$

Since $\tilde{\psi}_0$ is the joint probability mass function of a Poisson binomial distribution, by Lemma 8, we have

$$\begin{aligned}\frac{\tilde{\psi}_0(x)}{\bar{\psi}_0(x)} &\le \prod_{r=1}^K \exp\Big(\frac{2\beta K}{\varepsilon}\Big(\gamma+\frac{1}{n}\Big)(\hat{n}_r P_{kr} + \omega x_r)\Big) \\ &\le \exp\Big(\frac{2\beta K}{\varepsilon}\Big(\gamma+\frac{1}{n}\Big)\Big(np^* + \omega\sum_{r=1}^K x_r\Big)\Big).\end{aligned} \tag{7.7}$$

By Lemma 7, let $\delta := \rho/p^*$, and by the assumption $\max_r P_{kr} + \rho \le 1 - \varepsilon/2$, we obtain

$$\frac{\bar{\psi}_0(x)}{\hat{\psi}_0(x)} \le \prod_{r=1}^K \exp\Big(\delta\omega x_r + \frac{2\delta\hat{n}_r p^*}{\varepsilon}\Big) = \exp\Big(\delta\omega\sum_{r=1}^K x_r + \frac{2\delta n p^*}{\varepsilon}\Big). \tag{7.8}$$

We define subset of Z_+^K:

$$E = \{x \in \mathbb{Z}_+^K : \sum_{i=1}^K x_r \le C_\varepsilon n p^*\},$$

where C_ε is chosen to be the one in Lemma 9. Then for $x \in E$, $\sum_{r=1}^K x_r \le C_\varepsilon np^*$. We combine (7.7) and (7.8), and have

$$\frac{\tilde{\psi}_0(x)}{\hat{\psi}_0(x)} \le \exp\Big(C_3\Big(\delta+\gamma+\frac{1}{n}\Big)np^*\Big), \quad \forall x \in E, \tag{7.9}$$

where C_3 only depends on β, K, ε, and ω. Hence we have

$$\begin{aligned}
&\mathbb{P}\Big(\sup_{\hat{P}\in B(\rho)} Y_{ik\ell} \ge 0\Big) \le \sum_{x\in\mathbb{Z}_+^K} \tilde{\psi}_0(x) 1\Big\{\log\frac{\hat{\psi}_1(x)}{\hat{\psi}_0(x)} \ge -\frac{4n\rho}{\varepsilon}\Big\} \\
&= \sum_{x\in\mathbb{Z}_+^K} \tilde{\psi}_0(x) 1\Big\{\frac{e^{4n\rho/\varepsilon}\hat{\psi}_1(x)}{\hat{\psi}_0(x)} \ge 1\Big\} \\
&\le \sum_{x\in E} \exp\Big(C_3\Big(\delta+\gamma+\frac{1}{n}\Big)np^*\Big)\hat{\psi}_0(x) 1\Big\{\frac{e^{4n\rho/\varepsilon}\hat{\psi}_1(x)}{\hat{\psi}_0(x)} \ge 1\Big\} + \sum_{x\notin E}\tilde{\psi}_0(x) \\
&\le \exp\Big(C_3\Big(\delta+\gamma+\frac{1}{n}\Big)np^* + \frac{4n\rho}{\varepsilon}\Big) \sum_{x\in E}\min(\hat{\psi}_0(x),\hat{\psi}_1(x)) + \sum_{x\notin E}\tilde{\psi}_0(x) \\
&\le \exp\Big(C_4\Big(\delta+\gamma+\frac{1}{n}\Big)np^*\Big) \sum_{x\in E}\min(\hat{\psi}_0(x),\hat{\psi}_1(x)) + \sum_{x\notin E}\tilde{\psi}_0(x).
\end{aligned} \tag{7.10}$$

Again, by Lemma 7, for $x \in E$, we have

$$\frac{\hat{\psi}_0(x)}{\bar{\psi}_0(x)} \le \prod_{r=1}^K \exp\Big(\delta\omega x_r + \frac{2\delta\hat{n}_r p^*}{\varepsilon}\Big) = \exp\Big(\delta\omega\sum_{r=1}^K x_r + \frac{2\delta np^*}{\varepsilon}\Big) \le \exp(C_5 np^*).$$

Therefore, by Lemma 1, for all $\alpha \in (0,1)$,

$$\begin{aligned}
&\sum_{x\in E}\min(\hat{\psi}_0(x),\hat{\psi}_1(x)) \le \exp(C_5np^*)\sum_{x\in E}\min(\bar{\psi}_0(x),\bar{\psi}_1(x)) \\
&\lesssim \Big(\sqrt{np^*}\max_{r\in[K]}\Big|\log\frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})}\Big|\Big)^{-1}\exp(-D_\alpha(\bar{\psi}_0\|\bar{\psi}_1)).
\end{aligned} \tag{7.11}$$

Now we consider the perturbation of the Chernoff information. Let

$$\psi_0 \sim \bigotimes_{r=1}^K \mathrm{Bin}(n_r, P_{kr}) \quad \text{and} \quad \psi_1 \sim \bigotimes_{r=1}^K \mathrm{Bin}(n_r, P_{\ell r}),$$

where $n_r = \sum_{i=1}^n 1\{z_i = r\}$. By (2.16), we have

$$\frac{\exp(-D_\alpha(\bar{\psi}_0\|\bar{\psi}_1))}{\exp(-D_\alpha(\psi_0\|\psi_1))} = \prod_{r=1}^K [P_{kr}^{1-\alpha}P_{\ell r}^\alpha + (1-P_{kr})^{1-\alpha}(1-P_{\ell r})^\alpha]^{\hat{n}_r - n_r}. \tag{7.12}$$

Under the definition of $\hat{n}_r$ and assumption on $\hat{z}_j$, we have

$$\sum_{r=1}^K |\hat{n}_r - n_r| \le \sum_{j=1}^n 1\{z_j \ne \hat{z}_j\} \le n\gamma.$$

By assumption (3.3), we have $P_{kr}, P_{\ell r} \le p^*$ for $r \in [K]$. Thus

$$P_{kr}^{1-\alpha} P_{\ell r}^{\alpha} + (1-P_{kr})^{1-\alpha}(1-P_{\ell r})^{\alpha} \ge 1 - p^*.$$

Therefore,

$$\prod_{r=1}^{K} [P_{kr}^{1-\alpha} P_{\ell r}^{\alpha} + (1-P_{kr})^{1-\alpha}(1-P_{\ell r})^{\alpha}]^{\hat{n}_r - n_r} \le (1-p^*)^{n\gamma} \le e^{np^*\gamma}.$$

Hence, $\exp(-D_\alpha(\bar{\psi}_0 \| \bar{\psi}_1)) \le \exp(np^*\gamma - D_\alpha(\psi_0 \| \psi_1))$. Applying this bound to (7.11), and by Lemma 1, we have

$$\begin{aligned}
&\sum_{x \in E} \min(\hat{\psi}_0(x), \hat{\psi}_1(x)) \\
\lesssim &\Big(\sqrt{np^*} \max_{r \in [K]} \Big| \log \frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})} \Big| \Big)^{-1} \exp(-D_{\alpha^*}(\bar{\psi}_0 \| \bar{\psi}_1)) \\
\le &\Big(\sqrt{np^*} \max_{r \in [K]} \Big| \log \frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})} \Big| \Big)^{-1} \exp(np^*\gamma - D_{\alpha^*}(\psi_0 \| \psi_1)) \\
\lesssim &\exp(np^*\gamma)\eta(p_{k*}, p_{\ell *}),
\end{aligned}$$

where $\alpha^* = \arg\max_{\alpha \in (0,1)} D_{\alpha^*}(\psi_0 \| \psi_1)$. Thus,

$$\begin{aligned}
&\exp\Big(C_4\Big(\delta + \gamma + \frac{1}{n}\Big)np^*\Big) \sum_{x \in E} \min(\hat{\psi}_0(x), \hat{\psi}_1(x)) \\
\le &C_5 \exp\Big(C_2\Big(\delta + \gamma + \frac{1}{n}\Big)np^*\Big)\eta(p_{k*}, p_{\ell *}).
\end{aligned} \tag{7.13}$$

By Lemma 9 and Lemma 10,

$$\sum_{x \notin E} \tilde{\psi}_0(x) \le \exp(-np^*(1 + \log \varepsilon^{-1})) \le \exp(-np^* - D_{\alpha^*}(p_{k*} \| p_{\ell *})).$$

Under the assumption on P_{kr} and $P_{\ell r}$, we have

$$\frac{\varepsilon}{\omega} \le \frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})} \le \frac{\omega}{\varepsilon},$$

so $\Big| \log \frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})} \Big|^{-1} \ge \log \frac{\omega}{\varepsilon}$. Under the assumption $D_{\alpha^*}(p_{k*} \| p_{\ell *})^2 \ge C_1 np^*$, by Lemma 10 and using C_ε in that lemma,

$$np^* \ge \frac{D_{\alpha^*}(p_{k*} \| p_{\ell *})}{C_\varepsilon} \ge \frac{\sqrt{C_1 np^*}}{C_\varepsilon},$$

which implies $np^* \ge C_1 / C_\varepsilon^2$. Hence, for sufficiently large C_1, we have

$$\Big(\sqrt{np^*} \max_{r \in [K]} \Big| \log \frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})} \Big| \Big)^{-1} \ge \exp(-np^*).$$

Therefore, by Lemma 1,

$$\begin{aligned}&\exp(-np^* - D_{\alpha^*}(p_{k*}\|p_{\ell*}))\\ \le&\Big(\sqrt{np^*}\max_{r\in[K]}\Big|\log\frac{P_{kr}(1-P_{\ell r})}{P_{\ell r}(1-P_{kr})}\Big|\Big)^{-1}\exp(-D_{\alpha^*}(\psi_0\|\psi_1))\lesssim\eta(p_{k*},p_{\ell*}).\end{aligned}$$

This implies $\sum_{x\neq E}\tilde{\psi}_0(x)\lesssim\eta(p_{k*},p_{\ell*})$. Applying this bound and (7.13) to (7.11), we have

$$\mathbb{P}(\exists\hat{P}\in B(\rho),Y_{ik\ell}(\hat{P},\hat{z})\ge 0)\lesssim\exp\Big(C_2\Big(\delta+\gamma+\frac{1}{n}\Big)np^*\Big)\eta(p_{k*},p_{\ell*}),$$

as desired. □

Lemma 12 (Random Partitioning)**.** *Let I be a random subset of $[n]$ with $|I|=\lfloor n/2\rfloor$ in Algorithm 1 and $n_k^I=|\{i\in I:z_i=k\}|$, then for any $\zeta>0$ and sufficiently large n,*

$$\max_{k\in[K]}\Big|n_k^I-\frac{n_k}{2}\Big|\le n\xi$$

holds with probability at least $1-2K\exp\big(-n\xi^2/3\big)$.

Proof of Lemma 12. We have $n_k^I\sim\text{Hypergeometric}(\lfloor n/2\rfloor,n_k,n)$. For any fixed $k\in[K]$, the concentration of hypergeometric distribution [48] gives $\Big|n_k^I-\frac{n_k}{2}\Big|\le n\xi$ with probability at least $1-2\exp(-n\xi^2/3)$ when n is sufficiently large. Taking the union bound over all $k\in[K]$ gives the desired result. □

Lemma 13. *Suppose $D^*:=\min_{k\neq\ell}D_{\alpha^*}(p_{k*}\|p_{\ell*})$ is sufficiently large, then under SBM defined in Section 3.1, for any $r>0$ the misclassification rate of $\tilde{z}$ by spectral clustering in step 14 of Algorithm 1 satisfies $\text{Mis}(\tilde{z},z)\le Cr^{3/2}(D^*)^{-1}$ with probability at least $1-n^{-r}$, where C only depends on β,K,ω and r.*

Proof. The result follows from [6, Corollary 5]. □

7.5. Proof of Theorem 3.2

Under the assumption $(D^*)^2\ge C_2np^*$ for sufficiently large C_2, by Lemma 4, we have $np^*\gtrsim\sqrt{n}\bar{\sigma}_n\alpha^*(1-\alpha^*)\ge C_1$ where $\sqrt{n}\bar{\sigma}_n$ is defined in (2.17). Choosing sufficiently large C_2, C_1 is also large enough. To simplify the notation, we assume $np^*\ge C_1$. We will analyze the algorithm step by step. Each step fails with some probability, which will be summed up before calculating the error rate.

Spectral clustering and matching Assuming $D^*:=\min_{k\neq\ell}D_{\alpha^*}(p_k\|p_\ell)$ is sufficiently large, and let $r=4$, by Lemma 13, we have $\text{Mis}(\tilde{z},z)\le C_3/D^*\le\frac{1}{8\beta K}$ with probability at least $1-n^{-4}$, because β is fixed and $K=O(1)$. Without loss of generality, we assume the optimal permutation between $\tilde{z}$ and z is identity, that is, $n\,\mathrm{Mis}(\tilde{z}, z) = \sum_{i=1}^{n} 1\{\tilde{z}_i \neq z_i\}$. Now we consider spectral clustering in the for loop. Using Lemma 12 and let $\xi = \frac{1}{6\beta K}$, when n is sufficiently large,

$$\frac{n_k}{3} \leq \frac{n_k}{2} - \frac{n}{6\beta K} \leq n_k^I := |\{i \in I : z_i = k\}|.$$

For sufficiently large n, we have $n/(4\beta K) \leq n_k/4 \leq n_k^{I'}$. Similar bound holds for $n_k^{J'}$, i.e., $n_k/4 \leq n_k^{J'}$. This guarantees the community size in each partitioned subgraph is sufficiently large. Hence for $\alpha \in (0,1)$,

$$\begin{aligned} D_\alpha(p_{kI'} \| p_{\ell I'}) &= -\sum_{j \in I'} \log(p_{kj}^{1-\alpha} p_{\ell j}^{\alpha} + p_{kj}^{1-\alpha} p_{\ell j}^{\alpha}) \\ &\geq -\sum_{r=1}^{K} \frac{n_k}{4} \log(P_{kr}^{1-\alpha} P_{\ell r}^{\alpha} + P_{kr}^{1-\alpha} P_{\ell r}^{\alpha}) \\ &= \frac{D_\alpha(p_{k*} \| p_{\ell *})}{4}. \end{aligned} \tag{7.14}$$

Let $\alpha^* = \arg\max_{\alpha \in (0,1)} D_\alpha(p_{k*} \| p_{\ell *})$, then

$$D_{\alpha^*}(p_{kI'} \| p_{\ell I'}) \geq \frac{D_{\alpha^*}(p_{k*} \| p_{\ell *})}{4} \geq \frac{D^*}{4}.$$

Then the output $\tilde{z}'_{I'}$ of first spectral clustering in step 8 satisfies

$$\gamma_1 := \mathrm{Mis}(\tilde{z}'_{I'}, z_{I'}) \leq C_3 (D^*/4)^{-1} \leq \frac{1}{8\beta K}$$

when D^* is sufficiently large with probability at least $1 - (n/2 - 1)^{-4} \geq 1 - (n/3)^{-4}$. Now we consider the first matching algorithm in step 9. Let

$$\pi^* = \arg\max_{\pi \in S^K} \sum_{i \in I'} 1\{\tilde{z}_i \neq \pi(\tilde{z}'_i)\},$$

then $\sum_{i \in I'} 1\{\tilde{z}_i \neq \pi^*(\tilde{z}'_i)\} \leq \frac{|I'|}{8\beta K} \leq \frac{n}{16\beta K}$. On the other hand, since $n_k^{I'} \geq \frac{n}{4\beta K}$, we must have $|\{i \in I' : \tilde{z}'_i = k\}| \geq \frac{n}{4\beta K} - \frac{n}{16\beta K} = \frac{3n}{16\beta K}$. Hence for every $k \in [K]$, $|\{i \in I' : \tilde{z}_i = \pi^*(\tilde{z}'_i) = k\}| \geq \frac{3n}{16\beta K} - \frac{n}{16\beta K} = \frac{n}{8\beta K}$. On the other hand, for any $\pi \neq \pi^*$, $\sum_{i \in I'} 1\{\tilde{z}_i \neq \pi(\tilde{z}'_i)\} \geq 2 \cdot \frac{n}{8\beta K} = \frac{n}{4\beta K}$ because at least two labels have been permuted and at least $\frac{n}{4\beta K}$ of them match $\tilde{z}$ under the permutation π^*. Then by triangle inequality of the hamming distance, we have

$$\begin{aligned} \sum_{i \in I'} 1\{z_i \neq \pi(\tilde{z}'_i)\} &\geq \sum_{i \in I'} 1\{\tilde{z}_i \neq \pi(\tilde{z}'_i)\} - \sum_{i \in I'} 1\{z_i \neq \tilde{z}_i\} \\ &\geq \frac{n}{4\beta K} - \frac{n}{16\beta K} = \frac{3n}{16\beta K}. \end{aligned}$$

Therefore, π^* is the unique permutation such that $\sum_{i \in I'} 1\{z_i \neq \pi^*(\tilde{z}'_i)\} \leq \frac{n}{16\beta K}$. In other words, the matching algorithm succeed to find the optimal permutation between $\tilde{z}'_{I'}$ and $z_{I'}$. The second matching algorithm will similarly work. Therefore, the updated $\tilde{z}_{I'}$ and $\tilde{z}_{J'}$ are consistent with z.

First estimated parameters We will apply Lemma 6 to find the bound for $\tilde{P}$ in step 4. Recall from the spectral clustering step that we have $\gamma_1 \le 4C_3(D^*)^{-1}$ and we let $\tau_1 = C_4 D^*/(np^*) \le 1$ for some sufficiently small C_4. By concavity of $\log\left(1+\frac{2}{3}t\right)$ and $0 = \log 1 \le \frac{1}{6} \le \log\frac{5}{2} = \log\left(1+\frac{2}{3}\right)$, we have $\log\left(1+\frac{2}{3}t\right) \ge \frac{t}{6}$ on $[0,1]$.

$$h_1(t) = \frac{3}{4}t\log\Big(1+\frac{2}{3}t\Big) \ge \frac{3}{4}t\Big(\frac{t}{6}\Big) = \frac{t^2}{8} \quad \text{for} \quad t\in[0,1].$$

Thus, we have

$$\begin{aligned}\frac{n^2p^*h_1(\tau_1)}{4\beta^2K^2} &\ge \frac{n^2p^*}{32\beta^2K^2}\Big(\frac{C_4D^*}{np^*}\Big)^2 = \frac{C_4^2n(D^*)^2}{32\beta^2K^2np^*} \ge \frac{C_2C_4^2n}{32\beta^2K^2}\\ &\ge \frac{4n\log(D^*/(4C_3))}{D^*/(4C_3)} \ge -4n\gamma_1\log\gamma_1\end{aligned}$$

where the second and the third inequalities hold when D^* is sufficiently large, and the last inequality is due to the fact that $-x\log x$ is increasing on $[0,1/e]$. Therefore, with probability at most

$$\begin{aligned}\exp\Big[-\frac{n^2p^*h_1(\tau_1)}{4\beta^2K^2} - 2n\gamma\log\gamma\Big] &\le \exp\Big[-\frac{n^2p^*h_1(\tau_1)}{8\beta^2K^2}\Big]\\ &\le \exp\Big[-\frac{n(D^*)^2}{32C_2^2\beta^2K^2np^*}\Big] \le \exp(-2D^*),\end{aligned} \tag{7.15}$$

we have $\|\tilde{P}-P\|_\infty \le C_5(8\beta K\gamma_1+\tau_1)p^*$ fails, where C_5 corresponds to constants in Lemma 6, and the last inequality holds for sufficiently large D^*.

First likelihood ratio test In step 10, we apply likelihood ratio test on $A_{I'\times J'}$. We recall the definition of $Y_{ik\ell}$ in (7.4) from Lemma 11. The updated $\tilde{z}'_{I'}$ satisfies $\tilde{z}'_i = z_i$ if $Y_{iz_i\ell} < 0$ for every $\ell \ne z_i$. For $\tilde{P} \in B(\rho) := \{\hat{P} : \|\hat{P}-P\|_\infty \le \rho\}$, the probability that the classification error rate on nodes I' is at least γ_2 after the first likelihood ratio test is

$$\begin{aligned}&\mathbb{P}\Big(\sum_{i\in I'} 1\{\max_{\ell\ne z_i} Y_{iz_i\ell}(\tilde{P},\tilde{z}'_{I'}) \ge 0\} \ge |I'|\gamma_2\Big)\\ \le&\mathbb{P}\Big(\sum_{i\in I'} 1\{\exists\tilde{P}\in B(\rho), \max_{\ell\ne z_i} Y_{iz_i\ell}(\tilde{P},\tilde{z}'_{I'}) \ge 0\} \ge |I'|\gamma_2\Big).\end{aligned} \tag{7.16}$$

Let us define random variable

$$Z_i = 1\{\exists\tilde{P}\in B(\rho), \max_{\ell\ne z_i} Y_{iz_i\ell}(\tilde{P},\tilde{z}'_{I'}) \ge 0\}, \quad \text{for} \quad i\in I'.$$

Since $\tilde{z}'_{I'}$ only depends on $A_{I'\times J'}$, which is independent with $A_{I'\times I'}$. $\sum_{i\in I'} Z_i$ is a sum of independent variables. We can assume $\tilde{z}'_{I'}$ is fixed and satisfying $\mathrm{Mis}(\tilde{z}'_{I'}z_{I'}) \le 4C_3((D^*)^{-1}) := \gamma_1$. We apply Lemma 11 with $\rho := \rho_1 :=$

$C_5(8\beta K\gamma_1+\tau_1)p^*$, and let C_6 be the constant in the lemma,

$$\begin{aligned}\mathbb{P}(Z_i=1) &\le K\exp(C_6 n(\rho_1+p^*(\gamma_1+1/n)))\min_{\ell\neq z_i}\exp(-D_{\alpha^*}(p_{z_iJ'}\|p_{\ell J'}))\\ &\le K\exp\Big(C_6np^*\Big(C_5(8\beta K\gamma_1+\tau_1)+\gamma_1+\frac{1}{n}\Big)\Big)e^{-D^*/4}\\ &\le K\exp\Big(8C_5C_6\beta Knp^*\frac{4C_3}{D^*}+C_5C_6np^*\frac{C_4D^*}{np^*}+C_6p^*\Big)e^{-D^*/4}\\ &= K\exp\Big(8C_5C_6\beta Knp^*\frac{4C_3}{D^*}+C_4C_5C_6D^*+C_6p^*\Big)e^{-D^*/4}.\end{aligned}$$

Since $C_2np^*\le(D^*)^2$ for sufficiently large C_2, choosing sufficiently small C_4, we have

$$8C_5C_6\beta Knp^*\frac{4C_3}{D^*}+C_4C_5C_6D^*+C_6p^*\le\frac{D^*}{16}.$$

Since D^* is sufficiently large and $K=O(1)$, we have $\log K\le D^*/24$. Thus,

$$\mathbb{P}(Z_i=1)\le\exp(-D^*/4+D^*/16+D^*/16)=\exp(-D^*/8).$$

Applying Proposition 3 to (7.16), let $v=|I'|\exp(-D^*/8)$, $vt:=|I'|\gamma_2:=3|I'|e^{-D^*/16}+64$, then $t=e^{D^*/16}+\frac{64}{|I'|}e^{D^*/8}$, so the failing probability

$$\begin{aligned}&\mathbb{P}\Big(\sum_{i\in I'}Z_i\ge3|I'|e^{-D^*/16}+64\Big)\\ =&\mathbb{P}\Big(\sum_{i\in I'}Z_i-|I'|\mathbb{P}(Z_i=1)\ge3|I'|e^{-D^*/16}+64-|I'|\mathbb{P}(Z_i=1)\Big)\\ \le&\mathbb{P}\Big(\sum_{i\in I'}Z_i-|I'|\mathbb{P}(Z_i=1)\ge2|I'|e^{-D^*/16}+64\Big)\\ \le&\exp\Big(-(2|I'|e^{-D^*/16}+64)\frac{3}{4}\log\Big(1+\frac{4e^{D^*/16}}{3}\Big)\Big)\\ \le&\exp(-2D^*)\end{aligned}\tag{7.17}$$

The same error rate holds for $\tilde z'_{J'}$. Therefore, the updated $\tilde z'$ satisfies

$$\mathrm{Mis}(\tilde z',z)\le\frac{1}{n}(3|I'|e^{-D^*/16}+64+3|J'|e^{-D^*/16}+64+1)\le3e^{-D^*/16}+\frac{129}{n}.$$

Second estimated parameters As we have obtained labels $\tilde z'$ with higher accuracy, we would like to update $\tilde P$ as well. The proof is similar as the first estimated parameter, but with τ_2 and γ_2 different from τ_1 and γ_1. Let $\tau_2:=\frac{16\beta K(1\vee\sqrt{p^*D^*})}{np^*}$. Since np^* is sufficiently large and by Lemma 10, $\frac{\sqrt{D^*}}{np^*}\lesssim\frac{1}{\sqrt{D^*}}$, τ_2 is arbitrarily small. Using $h_1(t)\ge t^2/8$ for $t\in[0,1]$ again, we have

$$\frac{n^2p^*h_1(\tau_2)}{4\beta^2K^2}\ge\frac{256n^2p^*(1\vee p^*D^*)}{32(np^*)^2}=8D^*\vee\frac{8}{p^*}.$$

Let $\gamma_2 := 3e^{-D^*/16} + 129/n \leq 6e^{-D^*/16} \vee 258/n$. Since $-x \log x$ is increasing on $[0, 1/e]$, we have

$$\begin{aligned} -n\gamma_2 \log \gamma_2 &\leq (-n(6e^{-D^*/16}) \log(6e^{-D^*/16})) \vee (-258 \log \Big(\frac{258}{n}\Big)) \\ &= 6ne^{-D^*/16}\Big(\frac{D^*}{16} - \log 6\Big) \vee (258 \log n - 258 \log 258) \\ &\leq ne^{-D^*/32} \vee (258 \log n). \end{aligned}$$

Again, similar as in the step of first estimated parameter, we want to show that $\frac{n^2 p^* h_1(\tau_2)}{8\beta^2 K^2} \geq -4n\gamma_2 \log \gamma_2$. Using np^* is sufficiently large and $C_2 np^* \leq (D^*)^2$, we have

$$\frac{8}{p^*} = \frac{8n}{np^*} \geq 8ne^{-\frac{\sqrt{C_2 np^*}}{32}} \geq 8ne^{-D^*/32} \geq 4ne^{-D^*/32}.$$

By $C_2 np^* \leq (D^*)^2$ again, we have $C_2 n \leq \frac{(D^*)^2}{p^*}$, so either D^* or $1/p^*$ is greater than $\sqrt{C_2 n}$, which is greater than $1032 \log n$ when n is sufficiently large. Hence

$$\frac{n^2 p^* h_1(\tau_2)}{4\beta^2 K^2} \geq 8D^* \vee \frac{8}{p^*} \geq 4ne^{-D^*/32} \vee (1032 \log n) \geq -4n\gamma_2 \log \gamma_2.$$

Therefore, by Lemma 6, with failing probability at most,

$$\exp\Big[-\frac{n^2 p^* h_1(\tau_2)}{4\beta^2 K^2} - 2n\gamma_2 \log \gamma_2\Big] \leq \exp\Big(-4D^* \vee \frac{4}{p^*}\Big) \leq \exp(-2D^*). \quad (7.18)$$

we have $\|\hat{P} - P\|_\infty \leq C_5(8\beta K\gamma_2 + \tau_2)p^*$, where we recall that $\gamma_2 := 3e^{-D^*/16} + 129/n$ and $\tau_2 = \frac{16\beta K(1 \vee \sqrt{p^* D^*})}{np^*}$.

Second likelihood ratio test The arguments will be similar as the first likelihood ratio test. We define Z_j by new ρ, $\hat{P}$ and $\tilde{z}'$, i.e.,

$$Z_j = 1\{\exists \tilde{P} \in B(\rho), \max_{\ell \neq z_i} Y_{iz_i\ell}(\tilde{P}, \tilde{z}') \geq 0\}.$$

The likelihood ratio test $\hat{z}_j \leftarrow \mathcal{L}(A_{j*}, \hat{P}, \tilde{z}')$ in step 12 succeed to recover z_j if $Z_j = 0$. We apply Lemma 11 with $\rho := \rho_2 := C_5(8\beta K\gamma_2 + \tau_2)p^*$, and let $\eta^* = \max_{k \neq \ell} \eta(p_{k*}, p_{\ell*})$, then we have

$$\begin{aligned} \mathbb{P}(Z_j = 1) \leq &K \exp(C_6 n(\rho_2 + p^*(\gamma_2 + 1/n)))\eta^* \\ &+ 3\exp(-2D^*) + n^{-4} + 2\Big(\frac{n}{3}\Big)^{-4}, \end{aligned} \quad (7.19)$$

where $3\exp(-2D^*)$ comes from the failing probability of first parameter estimation (7.15), first likelihood ratio test (7.17), and second parameter estimation (7.18), n^{-r} is the failing probability of spectral clustering in step 3, and

$2\left(\frac{n}{3}\right)^{-r}$ is the failing probabilities in step 8. For the first term of (7.19), we have

$$\begin{aligned}C_6 n\rho_2 &= C_6 n C_5(8\beta K\gamma_2 + \tau_2)p^* \\ &= 8C_5C_6\beta K(3np^*e^{-D^*/16} + 129p^*) + 16C_5C_6\beta K(1 \vee \sqrt{p^*D^*}).\end{aligned}$$

Since $D^* \geq \sqrt{C_2 np^*}$, we have $np^*e^{-D^*/16} = O(1)$ for sufficiently large D^*. Clearly, $129p^* \leq 129 = O(1)$ and $16C_5C_6\beta K = O(1)$ by assumption 3.3. Thus,

$$Cn\rho_2 = O(1 + \sqrt{p^*D^*}).$$

For the other part of the first term, we similarly have

$$C_6 np^*(\gamma_2 + 1/n) = 3C_6 np^*e^{-D^*/16} + C_6 130p^* = O(1 + \sqrt{p^*D^*}).$$

To handle the second term of (7.19), we apply Lemma 1, then we have

$$\eta(p_{k*}, p_{\ell*}) \geq C_7\Big(\sqrt{np^*}\max_{j\in[n]}\Big|\log\frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})}\Big|\Big)^{-1}\exp(-D_{\alpha^*}(p_{k*}\|p_{\ell*}))$$

where C_7 is the constant of the lower bound in the lemma. Then by (3.3),

$$\Big|\log\frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})}\Big| \leq \log\frac{\omega}{\varepsilon},$$

Since $D^* \geq \sqrt{C_2 np^*}$, then for sufficiently large D^*, we have

$$C_7\Big(\sqrt{np^*}\max_{j\in[n]}\Big|\log\frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})}\Big|\Big)^{-1} \geq 3\exp(-D^*)$$

As a result,

$$\eta^* \geq C_7\Big(\sqrt{np^*}\max_{j\in[n]}\Big|\log\frac{p_{kj}(1-p_{\ell j})}{p_{\ell j}(1-p_{kj})}\Big|\Big)^{-1}\exp(-D^*) \geq 3\exp(-2D^*). \quad (7.20)$$

Combining these results, we have

$$\mathbb{P}(Z_j = 1) \leq \exp(C_8(1 + \sqrt{p^*D^*}))\eta^* + 3\Big(\frac{n}{3}\Big)^{-4}. \quad (7.21)$$

for some constants C_8. We will consider two cases:

Case 1: suppose $D^* \leq 2\log n$, then using $(D^*)^2 \geq C_2 np^*$,

$$p^*D^* = \frac{np^*D^*}{n} \leq \frac{(D^*)^3}{C_2 n} \leq \frac{C_9^3(\log n)^2}{n} = O(1)$$

and by (7.20),

$$\eta^* \geq e^{-2D^*} \geq e^{-4\log n} = n^{-4} \gtrsim 3\Big(\frac{n}{3}\Big)^{-4}.$$

Thus, (7.21) can be bounded by $\mathbb{P}(Z_j = 1) = O(\eta^*)$, and

$$\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] \leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(\hat{z}_j \neq z_j) \leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(Z_j = 1) = O(\eta^*).$$

Case 2: suppose $n\eta^* = o(1)$, we further divide this case into two situation. Suppose the condition $D^* \leq 2\log n$ is still satisfied, then $\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] = O(\eta^*)$. Therefore, by the Markov inequality,

$$\mathbb{P}(\mathrm{Mis}(\hat{z}, z) > 0) = \mathbb{P}(\mathrm{Mis}(\hat{z}, z) \geq 1/n) \leq n\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] = O(n\eta^*) = o(1).$$

Now suppose $D^* > 2\log n$, then using $\eta^* \lesssim e^{-D^*}$, for sufficiently large D^*, by (7.21), we have

$$\begin{aligned}\mathbb{P}(Z_j = 1) &\lesssim \exp(C_8(1 + \sqrt{p^* D^*}))e^{-D^*} + 3\Big(\frac{n}{3}\Big)^{-4} \\ &\leq e^{-\frac{2D^*}{3}} + 3\Big(\frac{n}{3}\Big)^{-4} \leq n^{-4/3} + 3\Big(\frac{n}{3}\Big)^{-4} \lesssim n^{-4/3}.\end{aligned}$$

Therefore,

$$\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] \leq \frac{1}{n}\sum_{j=1}^{n} \mathbb{P}(Z_j = 1) = O(n^{-4/3}).$$

By the Markov inequality again, we have

$$\mathbb{P}(\mathrm{Mis}(\hat{z}, z) > 0) \leq n\mathbb{E}[\mathrm{Mis}(\hat{z}, z)] = O(n \cdot n^{-4/3}) = o(1).$$

8. Data driven adjacency regularization

We provide the procedure of data driven regularization in step 14 of Algorithm 1. We note that this is only one of possible way to regularized the adjacency matrix. We refer readers to original paper [22] for more details.

Algorithm 2 Data-driven adjacency regularization

1: **Input:** $n \times n$ adjacency matrix A and regularization parameter τ. (Default: $\tau = 3$)
2: **Output:** Regularized adjacency matrix A_{re}.
3: Form the degree sequence $D_i = \sum_{j=1}^{n_2} A_{ij}$ for $i = 1, \ldots, n$ and the corresponding order statistics: $D_{(1)} \geq D_{(2)} \geq \cdots \geq D_{(n_1)}$.
4: Let $\bar{D} = \frac{1}{n_1}\sum_{i=1}^{n_1} D_i$ and $\alpha = \lfloor n_1/\bar{D} \rfloor$.
5: Set $\hat{d}_1 = \tau D_{(\alpha)}$ and $\mathcal{I} = \{i : D_i \geq \hat{d}_1\}$.
6: For every $i \in \mathcal{I}$, replace A_{i*} and A_{*i} by 0 to obtain A_{re}.

References

[1] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952. MR0057518

[2] Claude E Shannon, Robert G Gallager, and Elwyn R Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels. i. *Information and Control*, 10(1):65–103, 1967. MR0210513

[3] Richard Blahut. Hypothesis testing and information theory. *IEEE Transactions on Information Theory*, 20(4):405–417, 1974. MR0396072

[4] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015. MR3473334

[5] Chao Gao, Zongming Ma, Anderson Y Zhang, Harrison H Zhou, et al. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, 2018. MR3845014

[6] Zhixin Zhou and Arash A Amini. Optimal bipartite network clustering. *Journal of Machine Learning Research*, 21(40): 1–68, 2020. MR3828566

[7] Michael Nussbaum, Arleta Szkoła, et al. The chernoff lower bound for symmetric quantum hypothesis testing. *The Annals of Statistics*, 37(2):1040–1057, 2009. MR2502660

[8] Marco Dalai. Lower bounds on the probability of error for classical and classical-quantum channels. *IEEE Transactions on Information Theory*, 59(12):8027–8056, 2013. MR3142280

[9] Ke Li et al. Second-order asymptotics for quantum hypothesis testing. *The Annals of Statistics*, 42(1):171–189, 2014. MR3178460

[10] Lin Zhou, Vincent YF Tan, and Mehul Motani. Second-order asymptotically optimal statistical classification. *arXiv preprint* arXiv:1806.00739, 2018. MR3625993

[11] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011. MR2893856

[12] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013. MR3174850

[13] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in sparse stochastic block models. *arXiv preprint* arXiv:1312.2050, 2013. MR3285605

[14] Donniell E Fishkind, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe. Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications*, 34(1):23–39, 2013. MR3032990

[15] Van Vu. A simple svd algorithm for finding hidden partitions. *arXiv preprint* arXiv:1404.3918, 2014. MR3734334

[16] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing (STOC)*, pages 694–703. ACM, 2014. MR3238997

[17] Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint* arXiv:1412.7335, 2014.
[18] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015. MR3473374
[19] Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721, 2017. MR3694314
[20] Marianna Pensky and Teng Zhang. Spectral clustering in the dynamic stochastic block model. *arXiv preprint* arXiv:1705.01204, 2017. MR3914178
[21] Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint* arXiv:1709.09565, 2017.
[22] Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47):1–47, 2019. MR3948087
[23] Emmanuel Abbe, Enric Boix, Peter Ralli, and Colin Sandon. Graph powering and spectral robustness. *arXiv preprint* arXiv:1809.04818, 2018. MR4065193
[24] Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *arXiv preprint* arXiv:1406.5647, 2014. MR3766949
[25] Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016. MR3493879
[26] Afonso S Bandeira. Random laplacian matrices and convex relaxations. *Foundations of Computational Mathematics*, pages 1–35, 2015. MR3777782
[27] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3-4):1025–1049, 2016. MR3520025
[28] Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing (STOC)*, pages 814–827. ACM, 2016. MR3536616
[29] Federico Ricci-Tersenghi, Adel Javanmard, and Andrea Montanari. Performance of a community detection algorithm based on semidefinite programming. In *Journal of Physics: Conference Series*, volume 699, page 012015. IOP Publishing, 2016.
[30] Naman Agarwal, Afonso S Bandeira, Konstantinos Koiliaris, and Alexandra Kolla. Multisection in the stochastic block model using semidefinite programming. In *Compressed Sensing and its Applications*, pages 125–162. Springer, 2017. MR3751736
[31] Amelia Perry and Alexander S Wein. A semidefinite program for unbal-

anced multisection in the stochastic block model. In *Sampling Theory and Applications (SampTA), 2017 International Conference on*, pages 64–67. IEEE, 2017.

[32] Yingjie Fei and Yudong Chen. Achieving the bayes error rate in synchronization and block models by sdp, robustly. *arXiv preprint* arXiv:1904.09635, 2019a.

[33] Yingjie Fei and Yudong Chen. Achieving the bayes error rate in stochastic block model by sdp, robustly. In *Conference on Learning Theory (COLT)*, pages 1235–1269, 2019b.

[34] Arash A Amini, Aiyou Chen, Peter J Bickel, Elizaveta Levina, et al. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013. MR3127859

[35] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017. MR3687603

[36] Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016. MR3546450

[37] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory (COLT)*, pages 391–423, 2015.

[38] Gavin E Crooks. On measures of entropy and information. *Tech. Note*, 9:v4, 2017. MR3608468

[39] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2006. MR2239987

[40] Sergio Verdú. Asymptotic error probability of binary hypothesis testing for poisson point-process observations (corresp.). *IEEE Transactions on Information Theory*, 32(1):113–115, 1986. MR0831567

[41] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987. MR0902027

[42] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing (STOC)*, pages 69–75. ACM, 2015. MR3388184

[43] Anderson Y Zhang and Harrison H Zhou. Theoretical and computational guarantees of mean field variational inference for community detection. *arXiv preprint* arXiv:1710.11268, 2017.

[44] Koenraad MR Audenaert, John Calsamiglia, Ramón Munoz-Tapia, Emilio Bagan, Ll Masanes, Antonio Acin, and Frank Verstraete. Discriminating states: The quantum chernoff bound. *Physical review letters*, 98(16):160501, 2007.

[45] IG Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. In *Doklady Mathematics*, volume 82, pages 862–864. Springer, 2010. MR2790498

[46] Louis HY Chen and Qi-Man Shao. A non-uniform berry–esseen bound via stein's method. *Probability theory and related fields*, 120(2):236–254, 2001. MR1841329

[47] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015. MR3588285

[48] Vasek Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979. MR0534946