THE UNIVERSITY OF CHICAGO

TOPICS ON BAYESIAN INFERENCE SAMPLING ALGORITHMS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY
BUMENG ZHUO

CHICAGO, ILLINOIS
JUNE 2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I would like to thank my advisor Chao Gao, for his constant support and encouragement. He is such a great mentor, advisor and collaborator. He is always thoughtful, enthusiastic, and supportive of my graduate study, and he encourages me to keep moving forward. I would also like to thank all faculty and staff in the Department of Statistics where I have enjoyed my wonderful five years of graduate school. I am thankful to the HELIOS group members for their constructive comments on research and my fellow graduate students for great academic environment here. Finally, I would like to thank my parents, Guilong Zhuo and Yuyun Lin, for their love, encouragement, and patience.

ABSTRACT

Bayesian approach for inference has become one of the central interests in statistical inference, due to its advantages for expressing uncertainty in probability space, and the main hurdle is to sample from posterior distribution in various cases. Markov chain Monte Carlo (MCMC) is the most popular technique while theoretical properties have not yet been well understood. In this work, we study and propose several new sampling algorithms in cases of discrete variable sampling, when prior distribution is not smooth or even unbounded, and when likelihood itself is analytically unavailable and computationally intractable. The approaches we develop can be well applied to many real world applications such as community detection, image uncertainty quantification, and so on.

Chapter 2 studies computational complexity of a Metropolis-Hasting algorithm for Bayesian community detection. We first establish a posterior strong consistency result for a natural prior distribution on stochastic block models under the optimal signal-to-noise ratio condition in the literature, and then give a set of conditions that guarantee rapid mixing of a simple Metropolis-Hasting Markov chain. Chapter 3 proposes a novel sampling algorithm for non-smooth posterior sampling problem, motivated by the idea of unadjusted Langevin Monte Carlo algorithm and Tweedie's posterior mean formula. We provide rigorous non-asymptotic convergence analysis, and show that it outperforms other algorithms in some aspects. Chapter 4 provides an approximate Bayesian distribution via approximate Bayesian computation (ABC) that naturally incorporates prior information with loss function, when likelihood is not accessible. Asymptotic contraction results and Bernstein-von Mises type of property are proved for proposed Bayesian distribution under certain conditions.

CHAPTER 1

INTRODUCTION

Bayesian inference is fundamental primitive in statistical learning, and the central hurdle is to draw samples from posterior distribution. MCMC is the most well-known sampling method, in which the equilibrium distribution of the Markov chain matches the target distribution. Despite its popularity in Bayesian statistics and many other areas, there are limited understanding to the convergence of MCMC algorithms as well as some restrictions when likelihood itself is not analytically available. Examples of those are the application on Bayesian community detection, Bayesian sparse linear regression, and Bayesian generative model analysis.

In Chapter 2, we consider both statistical and computational performance of a simple Metropolis-Hastings algorithm for community detection problem. The primary goal is to study computational complexity for recovering the community memberships in a social network using a simple Markov chain. To be concrete, we construct a Bayesian model with some specific prior, analyze the performance of the corresponding posterior distribution, and provide a rapidly mixing time bound for the induced Metropolis-Hastings algorithm. The mixing time of a Markov chain is the number of iterations required to get close enough to the target distribution, in the sense that the total variation distance is bounded above by some small constant ε.

In Chapter 3, we tackle the problem of sampling from non-smooth or even unbounded density function. Inherited from Langevin Monte Carlo (LMC) algorithm, we propose a novel sampling algorithm that first smooths the density function through Tweedie's transformation, and then applies LMC algorithm. Tweedie's transformation reveals some advantages compared with Moreau-Yosida envelope studied in [28] and Gaussian smoothing studied in [15], as the first method can only be applied to unbounded and convex function, and the second method yields slower convergence rate. The non-asymptotic convergence analysis is performed in this work in terms of total variation distance. The proposed LMC algorithm via Tweedie's formula also works in non-convex cases but needs to be studied case by case according to its spectral gap.

In Chapter 4, we turn to generative model analysis, where the likelihood does not have closed-

form expression and cannot be numerically evaluated. We consider a special case where a frequentist estimate $\widehat{\theta}_Y$ is given, obtained by optimizing some loss function $L_n(\theta, Y)$ with observation $Y = (Y_i)_{i=1}^n$, and we aim to naturally incorporate prior information to get a Bayesian distribution. The proposed Bayesian distribution is constructed by prior distribution and the sampling distribution of $\widehat{\theta}_Y$, which yields asymptotic contraction result. In order to sample from the Bayesian distribution, we adopt modified ABC algorithm to draw samples, and show that the approximate Bayesian distribution induced by ABC also shares asymptotic contraction result as well as Bernstein-von Mises type of property, in the sense that it asymptotically recovers the frequentist distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta_0)$ with underlying true θ_0.

1.1 Summary

This thesis is intended to investigate various sampling algorithms to perform Bayesian inference in different cases. In Chapter 2, we focus on sampling discrete random variable on the topic of community detection. In Chapter 3, we work on sampling from some non-smooth and even unbounded distribution, a common issue in sparse Bayesian regression problems. In Chapter 4, we propose a new type of Bayesian distribution for generative model, and adopt modified ABC algorithms to do Bayesian inference, when the likelihood is not tractable.

1.2 Notation

Throughout the thesis, we will use the following notation. For positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n = o(b_n)$, $a_n = O(b_n)$ and $a_n \lesssim b_n$, if $\lim_{n\to\infty}(a_n/b_n) = 0$, $\limsup_{n\to\infty}(a_n/b_n) < \infty$, $\max\{a_n - b_n, 0\} = o(1)$, respectively. For an integer d, we use $[d]$ to denote $\{1, 2, \ldots, d\}$. For a set S, we write $\mathbb{I}\{S\}$ as its indicator function and $|S|$ as its cardinality. For a vector $v \in \mathbb{R}^d$, its norms are defined by $\|v\|_1 = \sum_{i=1}^d |v_i|$, $\|v\|^2 = \sum_{i=1}^d v_i^2$, and $\|v\|_\infty = \max_{1\le i\le d} |v_i|$. The Hamming error of two binary vectors $v_1, v_2 \in \{0,1\}^d$ is defined by $H(v_1, v_2) = \sum_{i=1}^d \mathbb{I}\{v_1(i) \neq v_2(i)\}$. For a matrix $A \in \mathbb{R}^{K\times K}$, its norms are defined by $\|A\|_\infty = \max_{i,j\in[K]} |A_{ij}|$, $\|A\|_1 = \sum_{i,j\in[K]} |A_{ij}|$, and

$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$. The L_2 norm of matrix coincides with the the spectral norm, and we will generally omit the subscript to simply write it as $\|A\|$. The notation $\mathbb{P}$ and $\mathbb{E}$ are generic probability and expectation operators whose distribution is determined from the context. We use C, c and their variants to denote absolute constants, and the values may vary from line to line. For any two distributions P and Q, the total variation distance is defined by $\|P-Q\|_{\mathsf{TV}} = \frac{1}{2}\int |dP - dQ|$, and the KL divergence is defined by $D(P\|Q) = \int dP \log \frac{dP}{dQ}$. For simplicity, we write $D(p\|q)$ to denote $D(\text{Bernoulli}(p)\|\text{Bernoulli}(q))$ for $p,q \in [0,1]$. For any two numbers a and b, we use $a \wedge b$ and $a \vee b$ to denote $\min\{a,b\}$ and $\max\{a,b\}$ respectively.

CHAPTER 2

MIXING TIME OF METROPOLIS-HASTINGS FOR BAYESIAN COMMUNITY DETECTION

Markov Chain Monte Carlo (MCMC) is a popular sampling technique, in which the equilibrium distribution of the Markov chain matches the target distribution. Most attention to date has been focused on Bayesian applications in order to sample from posterior distributions. Despite its popularity in Bayesian statistics and many other areas, its theoretical properties are not well understood, not to mention the limited theory for computational efficiency of MCMC algorithms, where the pivotal interest lies in the analysis of mixing time. The mixing time of a Markov chain is the number of iterations required to get close enough to the target distribution, in the sense that the total variation distance is bounded above by some small constant ε. We call the Markov chain rapidly mixing (resp. slowly mixing) if the mixing time grows at most polynomially (resp. exponentially) with respect to the sample size of the problem. One central research interest is to determine whether a designed Markov chain is rapidly mixing or slowly mixing.

Though slow mixing is the central hurdle of Markov chains, to the best of our knowledge, there has been little work on the theoretical analysis of mixing time. A series of studies have made efforts to design efficient Markov chains [75, 83, 55, 45, 77, 83] without providing theoretical guarantees. Over the past fifteen years, a surge of research has led to breakthroughs in the understandings of geometric ergodicity of Markov chains [67, 54, 26, 73], and several elegant techniques were developed to characterize the mixing property [12, 40, 26, 73, 53, 66, 50]. Canonical path is one main tool to show rapid mixing of Markov chains, and the idea is to design a set of paths between all pairs of points such that no edge is "overloaded" (congested). The method of canonical paths heavily relies on the graph structure, and the design of low congestion canonical paths remains a highly non-trivial artwork especially for exponentially large state space of Markov chains, which limits the range of applications. However, in some statistical problems, the construction of canonical paths may take advantage of the underlying model, and quantitative bounds for the convergence

rate and mixing time of Markov chains can be obtained under some general conditions. Yang et al. [85] was one of the first to apply the canonical paths idea to a Bayesian variable selection problem, and obtained an explicit upper bound for the mixing time under some mild conditions. Inheriting their ideas, this paper applies the same technique to a Bayesian community detection problem.

Motivated by the computational advantages of Gibbs sampling, a Bayesian point of view of community detection was first suggested in [74] with only two communities. The approach was further extended in [60, 43] to incorporate adjusted priors on community proportions as well as edge probabilities and allow for the case of more than two communities. There has been little theoretical analysis of Bayesian community detection until very recently, when the consistency results of posterior distribution were obtained by [79]. However, they required the expected degree of a node to be at least of order $\log^2 n$ to ensure the strong consistency of Bayesian posterior mode, which is a suboptimal condition for strong consistency [1, 2, 89]. Compared with the work on statistical performance, little work has been done on the computational efficiency to sample from the posterior distribution, and it was once suggested that the mixing time for high dimensional Bayesian community detection should scale exponentially, because the Markov chain must eventually go over all possible states.

2.1 Bayesian Community Detection

Networks have arisen in various areas of applications and have attracted a surge of research interests in fields such as physics, computer science, social sciences, biology, and statistics [36, 58, 81, 32, 16, 86, 87, 88]. In the realm of network analysis, community detection has emerged as a fundamental task that provides insights of the underlying structure. Great advances have been made on community detection recently with a remarkable diversity of models and algorithms developed in different areas [35, 59, 42]. Among various statistical models, the stochastic block model (SBM), first proposed in [44], is one of the most prominent generative model that depicts the network topologies and incorporates the community structure. It is arguably the simplest model of a graph with communities and has been widely applied in social, biological and communication networks.

Much effort has been devoted to SBM-based methods and their asymptotic properties have also been studied recently [11, 14, 10].

In this section, we give a precise formulation of the community detection problem and introduce a Bayesian approach. Then, we present the posterior strong consistency result.

2.1.1 Problem formulation

Consider an unweighted and undirected network with n nodes and K communities. The adjacency matrix is denoted by $A \in \{0,1\}^{n\times n}$, $A = A^T$, and $A_{ii} = 0$, for all $i \in [n]$. The edges are independently generated as Bernoulli variable with $\mathbb{E}A_{ij} = P_{ij}$, for all $i < j$. Here, P_{ij} denotes the connectivity probability for nodes i and j, and depends on the communities that the two nodes are assigned to. In this paper, we focus on a homogeneous SBM and assume $P_{ij} = p$ if two nodes are from the same community and $P_{ij} = q$ otherwise. We call p (resp. q) as the within-community (resp. between-community) connectivity probability and assume $p > q$ to satisfy the "assortative" property. Extensions to heterogenous SBMs are straightforward, but will not be considered in the paper for the sake of the presentation.

Let $Z \in [K]^n$ denote a label assignment vector, where Z_i is the community label for the ith node. Let $B \in [0,1]^{K\times K}$ be a symmetric connectivity probability matrix and thus $P_{ij} = B_{Z_iZ_j}$ with $B_{aa} = p$ for all $a \in [K]$, and $B_{ab} = q$ for all $a \neq b$. According to the description of the model, the likelihood formula can be written as

$$p(A|Z,B) = \prod_{i<j} B_{Z_iZ_j}^{A_{ij}} \left(1 - B_{Z_iZ_j}\right)^{1-A_{ij}}. \tag{1}$$

We use Z^* to denote the underlying true label assignment vector, and further assume that

$$\frac{n}{\beta K} \leq \sum_{i=1}^{n} \mathbb{I}\{Z_i^* = k\} \leq \frac{\beta n}{K}, \text{ for all } k \in [K], \tag{2}$$

where $\beta \geq 1$ is an absolute constant. It indicates that the all community sizes are of the same order.

When $\beta = 1 + o(1)$, all communities have almost the same sizes. Furthermore, we assume K is a known constant, $p, q \to 0$ and $p \asymp q$ throughout the paper. To conclude, this paper focuses on a sparse homogeneous SBM with a finite number of communities.

Note that community detection is a clustering problem, and thus any label assignment gives an equivalent result after a label permutation. To be specific, let

$$\Gamma(Z) = \{\sigma \circ Z : \sigma \in \mathscr{P}_K\}, \tag{3}$$

where $\mathscr{P}_K$ stands for the set of all permutations on $[K]$, and then any $Z' \in \Gamma(Z)$ leads to an equivalent clustering structure. Hence, with the identifiability issue, our ultimate goal is to reconstruct the community structure, or equivalently, to recover the community label assignment Z^* up to a label permutation.

2.1.2 A Bayesian model for community detection

In addition to the likelihood formula of the adjacency matrix A given in (1), we put a uniform prior on Z over a set S_α, where S_α is the set of all feasible label assignments depending on a hyperparameter α. The connectivity probabilities B_{ab} for $1 \le a \le b \le K$ receive independent Beta priors. More precisely, the Bayesian model is given by

$$\begin{aligned}
&\text{stochastic block model:} && p(A|Z,B) = \prod_{i<j} B_{Z_i Z_j}^{A_{ij}} \left(1 - B_{Z_i Z_j}\right)^{1-A_{ij}}, \\
&\text{label assignment prior:} && \pi(Z) \propto \mathbb{I}\{Z \in S_\alpha\}, \\
&\text{connectivity probability prior:} && B_{ab} \overset{\text{iid}}{\sim} \text{Beta}(\kappa_1, \kappa_2), \quad 1 \le a \le b \le K,
\end{aligned}$$

where $\kappa_1, \kappa_2 > 0$ measure the prior information of the connectivity probabilities and have negligible effects on the results when the sample size is large enough. This is essentially the same set-up

in [79], except that we introduce a uniform prior over set S_α. The key set S_α is defined by

$$S_\alpha = \left\{ Z : \sum_{i=1}^{n} \mathbb{I}\{Z_i = k\} \in \left[\frac{n}{\alpha K}, \frac{\alpha n}{K} \right], \text{ for all } k \in [K] \right\}, \tag{4}$$

where the hyperparameter α controls the size of the feasible set S_α, which rules out those models whose group sizes differ too much. We require $\alpha > \beta$ so that $Z^* \in S_\alpha$. As will be clarified in Section 2.3, this additional constraint seems to be necessary for the rapidly mixing according to our practical experiments.

The induced posterior distribution can be expressed as

$$\begin{aligned} \Pi(Z|A) &\propto \int_{[0,1]^{K(K+1)/2}} \prod_{a \le b} B_{ab}^{O_{ab}(Z)} (1 - B_{ab})^{n_{ab}(Z) - O_{ab}(Z)} d\Pi(B) \\ &\propto \prod_{a \le b} \text{Beta}(O_{ab}(Z) + \kappa_1, n_{ab}(Z) - O_{ab}(Z) + \kappa_2), \qquad \text{for } Z \in S_\alpha, \end{aligned}$$

and it follows that for $Z \in S_\alpha$,

$$\log \Pi(Z|A) = \sum_{a \le b} \log \text{Beta}(O_{ab}(Z) + \kappa_1, n_{ab}(Z) - O_{ab}(Z) + \kappa_2) + Const, \tag{5}$$

where $n_{ab}(Z) = n_a(Z) n_b(Z)$, $n_{aa}(Z) = n_a(Z)(n_a(Z) - 1)/2$ for all $a \neq b \in [K]$. We use $n_a(Z)$ to denote the size of community a, i.e., $n_a(Z) = |\{i : Z_i = a\}|$. We use $O_{ab}(Z)$ to denote the number of connected edges between communities a and b, which takes the formula $O_{ab}(Z) = \sum_{i,j} A_{ij} \mathbb{I}\{Z_i = a, Z_j = b\}$ and $O_{aa}(Z) = \sum_{i<j} A_{ij} \mathbb{I}\{Z_i = Z_j = a\}$ for all $a \neq b \in [K]$. Note that the posterior distribution is permutation symmetric, i.e.,

$$\Pi(Z|A) = \Pi(Z'|A), \quad \text{for all } Z' \in \Gamma(Z). \tag{6}$$

2.1.3 Posterior strong consistency

Before stating the theoretical properties of the proposed Bayesian model, we introduce some useful quantities. The first quantity I plays a crucial part in the minimax theory [89],

$$I = -2\log(\sqrt{pq} + \sqrt{(1-p)(1-q)}),$$

which is the Rényi divergence of order 1/2 between Bernoulli(p) and Bernoulli(q). It can be shown that when $p, q \to 0$,

$$I = (1+o(1))(\sqrt{p} - \sqrt{q})^2.$$

Then, we introduce an effective sample size to simplify the presentation of the results. As mentioned in [89], the minimax misclassification error rate is determined by that of classifying two communities of the smallest sizes. When $K = 2$, the hardest case is when one has two communities of the same size $n/2$. When $K > 2$, the hardest case is when one has two communities of sizes $n/K\beta$. Thus, we define

$$\bar{n} = \begin{cases} \dfrac{n}{2}, & \text{for } K = 2, \\ \dfrac{n}{K\beta}, & \text{for } K > 2, \end{cases} \tag{7}$$

as the effective sample size of the problem. The following result characterizes the statistical performance of the posterior distribution $\Pi(Z|A)$ under mild conditions.

Theorem 2.1.1 (Posterior strong consistency)**.** *Recall that* $\Gamma(Z) = \{\sigma \circ Z : \sigma \in \mathscr{P}_K\}$*, where* $\mathscr{P}_K$ *stands for the set of all permutations on* $[K]$*. Suppose that*

$$\liminf_{n\to\infty} \frac{\bar{n}I}{\log n} > 1, \tag{8}$$

and the feasible set S_α *satisfies that* $\alpha - \beta$ *is a positive constant. Then, we have that*

$$\mathbb{E}[\Pi(Z \in \Gamma(Z^*)|A)] \geq 1 - n\exp(-(1-\eta_n)\bar{n}I) = 1 - o(1)$$

for a large n and some positive sequence η_n tending to 0 as $n \to \infty$, and the expectation is with respect to the data-generating process.

We defer the proof of the theorem to Section 2.4. It is worth noting that the condition required in Theorem 4.2.1 is identical to the fundamental limits required for exact label recovery [1, 2, 89]. In the special case of two communities of equal sizes, we require $nI > 2\log n$ to guarantee the strong consistency result. Hence, Theorem 4.2.1 implies that under our Bayesian framework, posterior strong consistency holds under the optimal condition.

We can also compare the statistical performance of our model with other Bayesian community detection approaches. The first Bayesian SBM was suggested by [74], who considered two communities and proposed a uniform prior for both community proportions and the connectivity probabilities. It was further extended for more communities with Dirichlet priors on community proportions and Beta priors on the connectivity probabilities. However, the field of Bayesian SBM grows in a slow pace due to lack of theoretical analysis in terms of statistical consistency. Recently, van der Pas and van der Vaart [79] proved that the strong consistency result holds under a condition where the expected degree satisfies $\lambda_n \gg \log^2 n$. In contrast, our model introduces a feasible set S_α and proposes a uniform prior for label assignment Z on set S_α. It results in the strong consistency of posterior distribution under the condition that $n(p-q)^2/p \gtrsim \log n$, much weaker than the condition required in [79].

2.2 Rapidly mixing of a Metropolis-Hastings algorithm

In this section, we propose a modified Metropolis-Hastings random walk, and analyze its statistical performance as well as the computational complexity. Due to the identifiability of the problem, the rapidly mixing property is analyzed in clustering space that will be defined in the sequel.

2.2.1 A Metropolis-Hastings algorithm

A general Metropolis-Hastings algorithm is an iterative procedure consisting of two steps:

Step 1 For the current state X_t, generate an $X' \sim q(x|X_t)$, where $q(x|X_t)$ is the proposal distribution defined on the same state space.

Step 2 Move to the new state X' with acceptance probability $\rho(X_t, X')$, and stay in the original state X_t with probability $1 - \rho(X_t, X')$, where the acceptance probability is given by

$$\rho(X_t, X') = \min\left\{1, \frac{p(X')q(X_t|X')}{p(X_t)q(X'|X_t)}\right\},$$

where $p(\cdot)$ is the target distribution.

In this paper, we are sampling the community label assignment $Z \in [K]^n$. In particular, we take the *single flip update* as the proposal distribution, which is to choose an index $j \in [n]$ uniformly at random, and then randomly choose $c \in [K] \setminus \{Z_t(j)\}$ to assign a new label. The whole procedure is presented in Algorithm 1.

Algorithm 1: A Metropolis-Hastings algorithm for Bayesian community detection

Input: Adjacency matrix $A \in \{0,1\}^{n\times n}$,

number of communities K,

initial community assignment Z_0,

inverse temperature parameter ξ,

maximum number of iterations T.

Output: Community label assignment Z_T.

for each $t \in \{0,1,2,\ldots,T\}$ **do**

Choose an index $j \in [n]$ uniformly at random;

Randomly assign a new label for index j from the set $[K] \setminus \{Z_t(j)\}$ to get a new assignment Z';

$Z_{t+1} = Z'$ with probability

$$\rho(Z_t, Z') = \min\left\{1, \frac{\Pi^{\xi}(Z'|A)}{\Pi^{\xi}(Z_t|A)}\right\},$$

otherwise set $Z_{t+1} = Z_t$.

The Markov chain induced by Algorithm 1 is characterized by the transition matrix, which takes the form as

$$P(Z,Z') = \begin{cases} \dfrac{1}{n(K-1)} \min\left\{1, \dfrac{\Pi^{\xi}(Z'|A)}{\Pi^{\xi}(Z|A)}\right\}, & \text{if } H(Z,Z') = 1, \\ 1 - \displaystyle\sum_{Z' \neq Z} P(Z,Z'), & \text{if } Z' = Z, \\ 0, & \text{if } H(Z,Z') > 1, \end{cases} \tag{9}$$

where $H(Z,Z')$ is the Hamming error between the two label assignments Z, Z'. The inverse temperature parameter ξ satisfies that $\xi \geq 1$. The algorithm is sampling from the scaled distribution $\widetilde{\Pi}(Z|A)$, where $\widetilde{\Pi}(Z|A) \propto \Pi^{\xi}(Z|A)$ for any $Z \in [K]^n$. As $\xi \to \infty$, the probability mass of $\widetilde{\Pi}(\cdot|A)$ concentrates on the global maximum of $\Pi(\cdot|A)$, in which case the algorithm is deterministic and

reduces to a label switching algorithm, as discussed in [11]. When $\xi = 1$, asymptotically the algorithm is sampling from the true posterior distribution. The possible choices of ξ will be discussed in the sequel.

The parameter T in Algorithm 1 is the total number of iterations required. As long as the Markov chain mixes after T, according to Theorem 4.2.1, Z_T recovers the true community label assignment up to a label permutation with high probability, i.e., $Z_T \in \Gamma(Z^*)$ where $\Gamma(Z^*)$ is as defined in (3). Even though Theorem 4.2.1 is only stated for $\xi = 1$, it is easy to see that its conclusion also holds for $\widetilde{\Pi}(\cdot|A)$ for a general $\xi \geq 1$.

Due to the identifiability issue, the theoretical analysis of mixing time will be performed in the clustering space $\{\Gamma(Z) : Z \in S_\alpha\}$, where $\Gamma(Z)$ is defined in (3). We denote the state in the clustering space at time t as $\Gamma_t = \Gamma(Z_t)$, where Z_t is generated from Algorithm 1. The graphical model of the sequence $\{\Gamma_t\}_{t\geq 0}$ is illustrated by Figure 2.1.

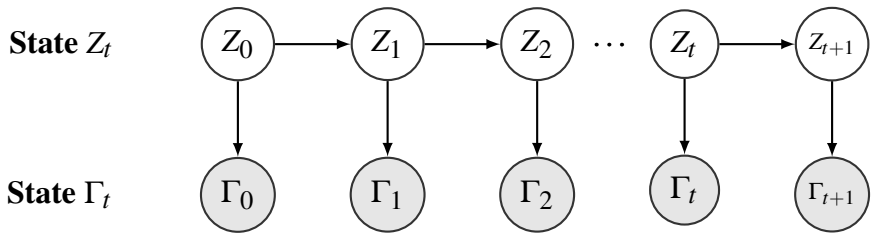


Figure 2.1: Updating process of $\Gamma(Z_t)$.

Proposition 2.2.1. *The sequence $\{\Gamma_t\}_{t\geq 0}$ induced by Algorithm 1 is a Markov chain.*

Proof. The proof relies on the permutation symmetry of the posterior distribution given by (6). We first introduce a distance between two clustering structures Γ and Γ', defined by

$$\check{H}(\Gamma,\Gamma') = \min_{Z\in\Gamma, Z'\in\Gamma'} H(Z,Z'). \tag{10}$$

When $\check{H}(\Gamma_{t+1},\Gamma_t) \leq 1$, we have

$$\mathbb{P}\{\Gamma_{t+1} \mid \Gamma_s,\ s\leq t\} = \sum_{Z\in\Gamma_t} \mathbb{P}\{\Gamma_{t+1} \mid Z_t = Z\} \cdot \mathbb{P}\{Z_t = Z \mid \Gamma_s,\ s\leq t\}. \tag{11}$$

The equality holds since given Z_t, Γ_{t+1} and $\{\Gamma_s : s \le t\}$ are independent. We proceed to calculate $\mathbb{P}\{\Gamma_{t+1} \mid Z_t = Z\}$. In the case of $\check{H}(\Gamma_{t+1}, \Gamma_t) \le 1$, it is obvious that for any $Z \in \Gamma_t$, there exists a unique $Z' \in \Gamma_{t+1}$ such that $H(Z, Z') \le 1$. Thus, we have that

$$\mathbb{P}\{\Gamma_{t+1} \mid Z_t = Z\} = \sum_{\widetilde{Z} \in \Gamma_{t+1}} P(Z, \widetilde{Z}) = P(Z, Z'). \tag{12}$$

By (9), the transition probability $P(Z, Z')$ only depends on the ratio of $\Pi(Z'|A)$ and $\Pi(Z|A)$, and

$$\frac{\Pi(Z'|A)}{\Pi(Z|A)} = \frac{\Pi(\Gamma(Z')|A)}{\Pi(\Gamma(Z)|A)} = \frac{\Pi(\Gamma_{t+1}|A)}{\Pi(\Gamma_t|A)}, \tag{13}$$

which only depends on Γ_t and Γ_{t+1}. It follows that

$$\mathbb{P}\{\Gamma_{t+1} \mid Z_t\} = \mathbb{P}\{\Gamma_{t+1} \mid \Gamma_t\}.$$

Hence, plug the above identity into (11), and we have

$$\begin{aligned}
\mathbb{P}\{\Gamma_{t+1} \mid \Gamma_s,\ s \le t\} &= \sum_{Z \in \Gamma_t} \mathbb{P}\{\Gamma_{t+1} \mid \Gamma_t\} \cdot \mathbb{P}\{Z_t = Z \mid \Gamma_s,\ s \le t\} \\
&= \mathbb{P}\{\Gamma_{t+1} \mid \Gamma_t\} \cdot \sum_{Z \in \Gamma_t} \mathbb{P}\{Z_t = Z \mid \Gamma_s,\ s \le t\} \\
&= \mathbb{P}\{\Gamma_{t+1} \mid \Gamma_t\}.
\end{aligned}$$

When $\check{H}(\Gamma_{t+1}, \Gamma_t) > 1$, it is obvious that

$$\mathbb{P}\{\Gamma_{t+1} \mid \Gamma_s,\ s \le t\} = 0 = \mathbb{P}\{\Gamma_{t+1} \mid \Gamma_t\}.$$

Therefore, $\{\Gamma_t\}_{t \ge 0}$ is a Markov chain by combining the conclusions of the two cases. □

According to the above proposition and its proof, we can define the transition matrix $\check{P}$ from

state Γ to Γ' as

$$\check{P}(\Gamma,\Gamma') = \begin{cases} \dfrac{1}{n(K-1)} \min\left\{1, \left[\dfrac{\Pi(\Gamma'|A)}{\Pi(\Gamma|A)}\right]^{\xi}\right\}, & \text{if } \check{H}(\Gamma,\Gamma') = 1, \\ 1 - \displaystyle\sum_{\Gamma' \neq \Gamma} \check{P}(\Gamma,\Gamma'), & \text{if } \Gamma' = \Gamma, \\ 0, & \text{if } \check{H}(\Gamma,\Gamma') > 1. \end{cases} \tag{14}$$

We perform the analysis of mixing time for the Markov chain $\{\Gamma_t\}_{t\geq 0}$. Write $\check{S}_\alpha = \{\Gamma(Z) : Z \in S_\alpha\}$ for simplicity, and we define the target distribution in the clustering space as $\check{\Pi}(\Gamma|A) = \sum_{Z\in\Gamma} \widetilde{\Pi}(Z|A)$ for any $\Gamma \in \check{S}_\alpha$. We show in the next section that $\{\Gamma_t\}_{t\geq 0}$ is rapidly mixing to the target distribution $\check{\Pi}(\cdot|A)$.

2.2.2 Main results

Before stating the main theorem, we first review the definition of ε-mixing time, as well as the loss function that we need for the community detection problem.

ε-mixing time. Let $\Gamma_0 = \Gamma(Z_0)$ be the initial state of the chain. The total variation distance to the stationary distribution after t iterations is

$$\Delta_{Z_0}(t) = \left\|\check{P}^t(\Gamma_0,\cdot) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}} = \frac{1}{2} \sum_{\Gamma\in\{\Gamma(Z):Z\in S_\alpha\}} \left|\check{P}^t(\Gamma_0,\Gamma) - \check{\Pi}(\Gamma|A)\right|,$$

where $\check{P}^t(\Gamma_0,\cdot)$ and $\check{\Pi}(\cdot|A)$ are both distributions defined in the clustering space. The ε-mixing time for Algorithm 1 starting at Z_0 is defined by

$$\tau_\varepsilon(Z_0) = \min\left\{t \in \mathbb{N} : \Delta_{Z_0}(t') \leq \varepsilon \text{ for all } t' \geq t\right\}. \tag{15}$$

It is the minimum number of iterations required to ensure the total variation distance to the stationary distribution is less than some tolerance threshold ε.

Loss function. We introduce the misclassification proportion as a loss function, which is defined by

$$\ell(Z,Z^*) = \frac{1}{n}\check{H}(\Gamma(Z),\Gamma(Z^*)), \tag{16}$$

where $\check{H}(\cdot,\cdot)$ in defined in (10).

To this end, let us show that the proposed modified Metropolis-Hastings algorithm in Section 2.2.1 gives a rapidly mixing Markov chain $\{\Gamma_t\}_{t\geq 0}$ under the following conditions.

Condition 2.1. There exist some positive sequences $\eta = \eta(n)$ and $\gamma_0 = \gamma_0(n)$ such that

$$\inf_{B,Z^*} \mathbb{P}\{\ell(Z_0,Z^*) \leq \gamma_0\} \geq 1-\eta.$$

We proceed to state the conditions for γ_0.

Condition 2.2. Suppose the sequence γ_0 in Condition 2.1 satisfies one of the following cases:

- Case 1: there are only two communities, i.e., $K = 2$, and

$$(1-K\gamma_0)^4 nI \to \infty, \;\; (1-K\gamma_0)(1-K\beta\gamma_0)n \to \infty, \tag{17}$$

  where $\beta \geq 1$ is defined in (2).

- Case 2: there are more than 2 communities, i.e., $K \geq 3$, and

$$\gamma_0 = o(1). \tag{18}$$

Condition 2.1 and Condition 2.2 require that the misclassification number of the initial label assignment is less than the minimum community size $n/K\beta$ with high probability. Consider the special situation where $K = 2$ and $\beta = 1 + o(1)$, i.e., the underlying two community share the same size asymptotically, Condition 2.2 is satisfied when the initial misclassification proportion is $1/2 - \varepsilon$ for some sequence $\varepsilon \to 0$. When $K \geq 3$, we require a stronger condition that the initial

label assignment need to be weakly consistent, i.e., the initial misclassification error goes to 0 as $n \to \infty$. The initial condition can be easily satisfied by algorithms such as spectral clustering [52, 70, 21, 31].

Condition 2.3. Suppose $\limsup_{n\to\infty} \log n/\bar{n}I = 1-\varepsilon_0$. With the hyperparameter ξ defined in Algorithm 1, one of the following cases holds:

- Case 1: there are only two communities, i.e., $K = 2$, and

$$\xi > (1-\varepsilon_0)\left\{\frac{1}{2\varepsilon_0} \vee \frac{\alpha^2}{(1-K\gamma_0)^4}\right\}, \tag{19}$$

  where α is defined in (4), and γ_0 is defined in Condition 2.1.

- Case 2: there are more than 2 communities, i.e., $K \geq 3$, and

$$\xi > \frac{1-\varepsilon_0}{2\varepsilon_0}. \tag{20}$$

Note that the condition for the inverse temperature hyperparameter ξ also depends on the signal condition (ε_0) and initialization condition (γ_0). The condition of ξ is provided to ensure the strong rapidly mixing property in the worst scenario. Note that with stronger initialization condition for the case of $K \geq 3$, the condition of ξ is slightly weaker than the case of $K = 2$.

Here are some intuitive understandings of Condition 2.3. Theorem 4.2.1 shows that posterior strong consistency holds under the condition $\liminf_{n\to\infty} \bar{n}I/\log n > 1$. Suppose for two label assignments Z_1, Z_2 that $\Pi(Z_1|A) > \Pi(Z_2|A)$, with hyperparameter $\xi \geq 1$, the posterior ratio $\Pi^{\xi}(Z_1|A)/\Pi^{\xi}(Z_2|A)$ gets enlarged, and the Markov chain is more certain to move towards the maximum point. However, the value of ξ is also constrained by the initialization Z_0. With a larger ξ, $\Pi^{\xi}(Z_0|A)$ is smaller and it takes longer for the Markov chain $\{\Gamma_t\}_{t\geq 0}$ to get mixed. The special case is that when the initialization Z_0 is weakly consistent, or equivalently, $\ell(Z_0, Z^*) \to 0$ as $n \to \infty$, then the value of ξ only depends on ε_0. It gives the following alternative condition that can replace Condition 2.2 and Condition 2.3.

Condition 2.4. Denote $\limsup_{n\to\infty} \log n/\bar{n}I = 1-\varepsilon_0$. The positive sequence γ_0 defined in Condition 2.1 and the hyperparameter ξ satisfy that

$$\gamma_0 = o(1), \;\; \xi \geq \frac{1-\varepsilon_0}{2\varepsilon_0}. \tag{21}$$

Theorem 2.2.1 (Rapidly mixing)**.** *The initial label assignment is denoted by Z_0. Suppose Conditions (2.1, 2.2, 2.3) or Conditions (2.1, 2.4) are satisfied. Then, the ε-mixing time of the modified Metropolis-Hastings algorithm is upper bounded by*

$$\tau_\varepsilon(Z_0) \leq 4Kn^2 \max\left\{\gamma_0, n^{-\tau}\right\} \cdot \left(\xi \log\left(\Pi(Z_0|A)^{-1}\right) + \log(\varepsilon^{-1})\right) \tag{22}$$

with probability at least $1 - C_1 n^{-C_2} - \eta$ for some constant $C_1, C_2 > 0$, where τ is a sufficiently small constant, and η is defined in Condition 2.1.

Remark 2.2.1. It is classical to perform theoretical analysis on a lazy version of Markov chain, which has probability 1/2 of staying unchanged, and the other probability 1/2 of updating the state. Theorem 2.2.1 is proved for the lazy Markov chain induced by Algorithm 1, i.e., the corresponding transition matrix is $(\check{P}+I)/2$. The same tricks are widely used in [85, 50, 8, 56]. It is worth noting that this is only for the proof, and in practice, we still use the original transition matrix in Algorithm 1.

Theorem 2.2.1 implies the mixing time depends on the initialization Z_0 and the choice of ξ. In order to show that the mixing time is at most a polynomial of n, we still need the following lemma to lower bound the initial posterior value $\Pi(Z_0|A)$.

Lemma 2.2.1. *Under the conditions of Theorem 4.2.1, we have*

$$\log \Pi(Z_0|A) \geq -C_3 n^2 I \cdot \ell(Z_0, Z^*), \tag{23}$$

with probability at least $1 - C_4 n^{-C_5}$ for some positive constants C_3, C_4, C_5.

Theorem 2.2.1 and Lemma 2.2.1 jointly imply that $\tau_\varepsilon(Z_0) \lesssim n^2(n^2 I + \log(\varepsilon^{-1}))$ with high probability, which demonstrates that the Markov chain of Metropolis-Hastings algorithm is rapidly mixing. To the best of our knowledge, (22) is the first explicit upper bound on the mixing time of the Markov chain for Bayesian community detection.

Note that the target distribution of Algorithm 1 is $\widetilde{\Pi}(\cdot|A) \propto \Pi^{\xi}(\cdot|A)$. Since $\xi \geq 1$ and the posterior strong consistency property still holds for $\widetilde{\Pi}(\cdot|A)$, Theorem 2.2.1 shows that Algorithm 1 will find the maximum a posteriori in polynomial time with high probability.

Corollary 2.2.1. *Under the condition of Theorem 2.2.1, for any iteration number T such that $T \geq C_6 n^2(n^2 I + \log(\varepsilon^{-1}))$ for some constant C_6, the output Z_T of the Algorithm 1 satisfies that $Z_T \in \Gamma(Z^*)$ with high probability, or equivalently, $\ell(Z_T, Z^*) = 0$.*

The following corollary focuses on the case of $\xi = 1$, and gives explicit conditions for the Markov chain to converge to the posterior distribution $\Pi(\cdot|A)$.

Corollary 2.2.2. *When $nI/\log n \to \infty$, suppose Condition 2.2 holds, and we can take $\xi = 1$ in Algorithm 1, which reduces to the standard Metropolis-Hastings algorithm sampling from $\Pi(\cdot|A)$. We have that the ε-mixing time of the Markov chain is upper bounded by $O(n^2(n^2 I + \log(\varepsilon^{-1})))$ with high probability.*

The conditions of the above results can be weakened in the case where the connectivity probability matrix B is known. When B is known, there is no need to put a prior on B. Thus, the posterior distribution can be simplified as

$$\begin{aligned} \log \Pi(Z|A) = &\log \frac{p(1-q)}{q(1-p)} \sum_{i<j} A_{ij} \mathbb{I}\{Z_i = Z_j\} - \\ &\log \frac{1-q}{1-p} \sum_{i<j} \mathbb{I}\{Z_i = Z_j\} + Const, \quad \text{for } Z \in S_\alpha. \end{aligned} \tag{24}$$

The posterior formula is essentially the same as likelihood, while we restrict Z inside the feasible set S_α. It can be shown that the posterior strong consistency property still holds in this case.

Theorem 2.2.2 (posterior strong consistency)**.** *Suppose that* $\limsup_{n\to\infty} \bar{n}I/\log n > 1$*, and the feasible set* S_α *satisfies that* $\alpha - \beta$ *is a positive constant, then it follows that*

$$\mathbb{E}\left[\Pi(Z \in \Gamma(Z^*)|A)\right] \geq 1 - o(1),$$

with high probability, and the expectation is with respect to the data-generating process.

Condition 2.5. Suppose $\limsup_{n\to\infty} \log n/\bar{n}I = 1 - \varepsilon_0$. Assume the positive sequence γ_0 defined in Condition 2.1 and the hyperparameter ξ satisfy one of the following conditions:

- Case 1:

$$\gamma_0 = o(1), \;\; \xi \geq \frac{1-\varepsilon_0}{2\varepsilon_0}. \tag{25}$$

- Case 2:

$$(1 - K\alpha\gamma_0)^2 nI \to \infty, \;\; \xi > \begin{cases} (1-\varepsilon_0)\left(\dfrac{1}{2\varepsilon_0} \vee \dfrac{\alpha}{4(1-K\alpha\gamma_0)}\right), & \text{for } K = 2, \\ (1-\varepsilon_0)\left(\dfrac{1}{2\varepsilon_0} \vee \dfrac{\alpha}{4\beta(1-K\alpha\gamma_0)}\right), & \text{for } K \geq 3. \end{cases} \tag{26}$$

Condition 2.5 yields the rapidly mixing property when the connectivity matrix B is known.

Theorem 2.2.3 (Rapidly mixing)**.** *Suppose we start the algorithm at* Z_0*, and Conditions (2.1, 2.5) hold. Then, the* ε*-mixing time of the Metropolis-Hastings algorithm induced by Equation* (24) *is upper bounded by*

$$\tau_\varepsilon(Z_0) \leq 4Kn^2 \max\left\{\gamma_0, n^{-\tau}\right\} \cdot \left(\xi \log \Pi(Z_0|A)^{-1} + \log(\varepsilon^{-1})\right),$$

with probability at least $1 - C_7 n^{-C_8} - \eta$ *for some constants* C_7, C_8*, where* τ *is a sufficiently small constant, and* η *is as defined in Condition 2.1.*

Compare the result with Theorem 2.2.1, and we can see that Theorem 2.2.3 obtains the same upper bound with slightly weaker conditions for the initialization Z_0 and hyperparameter ξ.

2.3 Numerical Results

In this section, we study the numerical performance of the Metropolis-Hastings algorithm.

Balanced networks. In this setting, we generate networks with 2500 nodes, and 5 communities, each of which consists of 500 nodes. Figure 2.2 shows the trajectories of the Markov chains (each denoted by a black line). By posterior strong consistency, the true label assignment receives the highest posterior probability (denoted by the red line), and the Markov chains converge rapidly to the stationarity (within $40n$ iterations), demonstrating the rapidly mixing property.

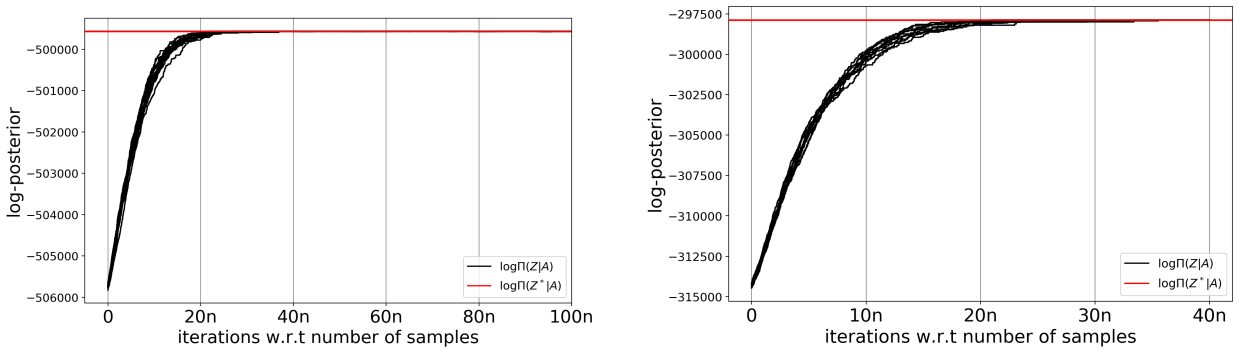


Figure 2.2: Log-posterior probability versus the number of iterations. Each black curve corresponds to a trajectory of the chain (20 chains in total), and the red horizontal line represents the log-posterior probability at the true label assignment. (a) A network with $p = 0.48$ and $q = 0.32$. (b) A network with $p = 0.3$ and $q = 0.1$.

Heterogeneous networks. In this setting, we generate networks with 2000 nodes and 4 communities of sizes 200, 400, 600, and 800, respectively. The connectivity matrix is set as

$$B = \begin{pmatrix} 0.50 & 0.29 & 0.35 & 0.25 \\ 0.29 & 0.45 & 0.25 & 0.30 \\ 0.35 & 0.25 & 0.50 & 0.35 \\ 0.25 & 0.30 & 0.35 & 0.45 \end{pmatrix}.$$

The algorithm still performs well. As shown in Figure 2.3, the posterior strong consistency still hold, and the Markov chains rapidly converge to the stationarity.

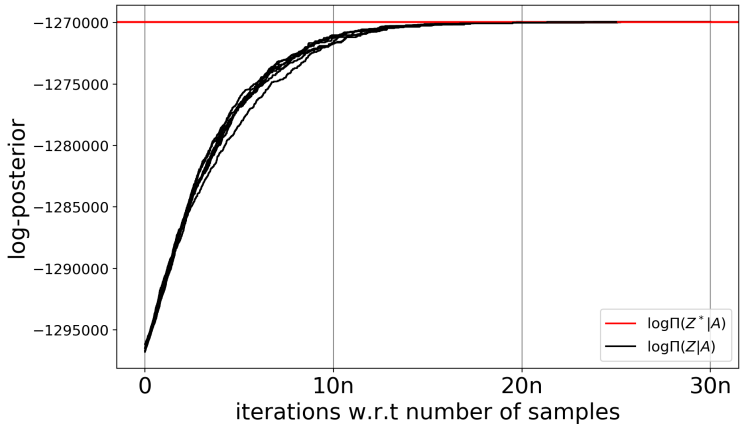
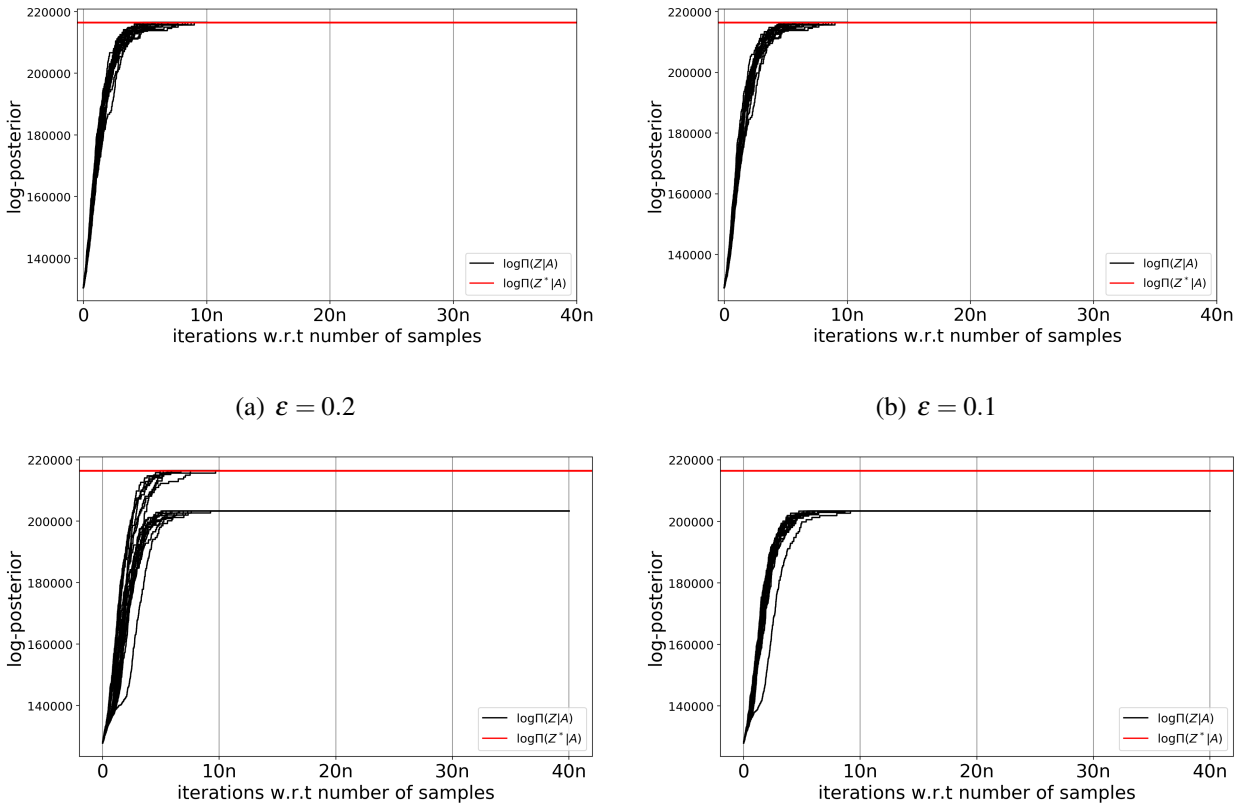


Figure 2.3: Log-posterior probability versus the number of iterations. Each black curve corresponds to a trajectory of the chain (20 chains in total), and the red horizontal line represents the log-posterior probability at the true label assignment.

Necessity of the initialization condition. We show that the initialization condition required by our main theorems is necessary by numerical experiments. Consider the network with two communities of size 270 and 460, and the connectivity probabilities are set to be $p = 10^{-1}$, $q = 10^{-8}$. The initial label assignment Z_0 satisfies $\ell(Z_0, Z^*) = (1-\varepsilon)/2\alpha$, and then Condition 2.5 is equivalent to $\varepsilon > 0$ and $\varepsilon^2 nI \to \infty$. In simulations, we run experiments for $\varepsilon = 0.2, 0.1, -0.1, -0.2$, and the results are shown as below.



(a) $\varepsilon = 0.2$

(b) $\varepsilon = 0.1$

(c) $\varepsilon = -0.1$

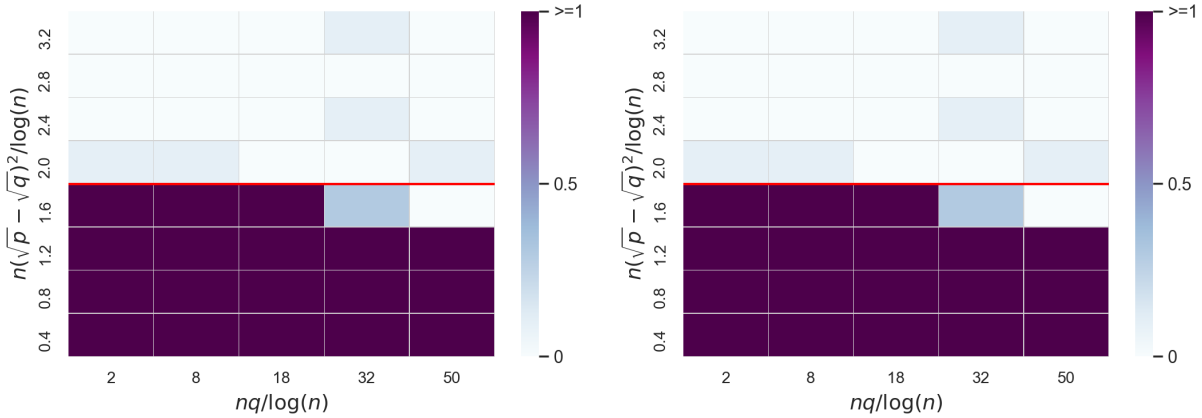
(d) $\varepsilon = -0.2$

Figure 2.4: Log-posterior probability versus the number of iterations. The initial label assignment Z_0 is constructed so that the labels of the community of size 270 are all correct, and there are $n(1-\varepsilon)/2\alpha$ labels in the community of size 460 are incorrect. Each black curve corresponds to a trajectory of the chain (20 chains in total), and the red horizontal line represents the log-posterior probability at the true label assignment.

Figure 2.4 shows that when $\varepsilon < 0$, it is very likely for the algorithm to get stuck at some local maximum, and does not converge to the stationary distribution.

Fundamental limit of the signal condition. We check that the fundamental limit of the signal condition can be achieved by the Metropolis-Hastings algorithm. We generate homogeneous networks with 1000 nodes and 2000 nodes, and each has two communities of equal sizes. Figure 2.5 is the heatmap of the number of misclassified samples, where every rectangular block represents

one setting with different values of p and q. In each setting, we run 20 experiments with independent initializations and adjacency matrices, and the value of each block is the average number of misclassified samples in the 20 experiments. Figure 2.5 shows that when $nI > 2\log n$, we are able to exactly recover the underlying true label assignment, and the result of simulation coincides with the posterior strong consistency property in Section 2.1.3.



(a) Network with 1000 nodes

(b) Network with 2000 nodes

Figure 2.5: The heatmap of the number of misclassified samples. The red line in each plot represents the fundamental limit with $K = 2$.

2.4 Proofs

The posterior strong consistency property, Theorem 4.2.1 and Theorem 2.2.2, is proved in Section 2.4.1. The main result of the paper, Theorem 2.2.1, is proved in Section 2.4.2. Other technical proofs are in Appendix.

2.4.1 Proof of posterior strong consistency

We first state the proof in the case where the connectivity probability matrix B is known (Theorem 2.2.3). Then, by similar techniques, we have the result of Theorem 4.2.1. To distinguish the two cases, we denote the posterior distribution as $\Pi_0(\cdot|A)$ with a known connectivity probability matrix. In this section, we use $d(Z,Z^*) = n\ell(Z,Z^*) = m(Z)$ to denote the number of mistakes for

the label assignment Z. For simplicity, we also write m for $m(Z)$ with a slight abuse of notation.

Proof of Theorem 2.2.2

We first state a lemma in order to prove the theorem.

Lemma 2.4.1 (Lemma 5.4 in [89]). *For any constants $\alpha > \beta \geq 1$, let $Z \in S_\alpha$ be an arbitrary assignment satisfying that $d(Z,Z^*) = m$ with $0 < m < n$. Then, for the $\Pi_0(Z|A) = \Pi(Z|A)$ defined in* (24)*, we have*

$$\mathbb{P}\left\{\Pi_0(Z|A) > \Pi_0(Z^*|A)\right\} \leq \begin{cases} \exp\left(-(\bar{n}m - m^2)I\right), & m \leq \dfrac{n}{2K}, \\ \exp\left(-d_{\alpha,\beta}\dfrac{nmI}{K}\right), & m > \dfrac{n}{2K}, \end{cases}$$

where $d_{\alpha,\beta}$ is some positive constant that only depends on α, β.

Proof of Theorem 2.2.2. Recall that for any $Z' \in \Gamma(Z)$, we have $\Pi_0(Z'|A) = \Pi_0(Z|A)$. For each $Z \in S_\alpha$, let $G_Z = \{\Pi_0(Z|A) > \Pi_0(Z^*|A)\}$, and define $G = \cup_{Z \in S_\alpha} G_Z$. Let P_Z denote the likelihood function for the assignment Z. With the uniform prior on S_α, we have

$$\begin{aligned}
\mathbb{E}\Pi_0(Z \notin \Gamma(Z^*)|A) =& P_{Z^*}\Pi_0(Z \notin \Gamma(Z^*)|A)\mathbb{I}\{G^c\} + P_{Z^*}\Pi_0(Z \notin \Gamma(Z^*)|A)\mathbb{I}\{G\} \\
\leq& P_{Z^*}\sum_{Z \notin \Gamma(Z^*)} \frac{P_Z}{P_{Z^*} + \sum_Z P_Z}\mathbb{I}\{G_Z^c\} + \sum_{Z \notin \Gamma(Z^*)} P_{Z^*}(G_Z) \\
\leq& P_{Z^*}\sum_{Z \notin \Gamma(Z^*)} \frac{P_Z}{P_{Z^*}}\mathbb{I}\{G_Z^c\} + \sum_{Z \notin \Gamma(Z^*)} P_{Z^*}(G_Z) \\
=& \sum_{Z \notin \Gamma(Z^*)} P_Z(G_Z^c) + P_{Z^*}(G_Z) \\
=& 2\sum_{Z \notin \Gamma(Z^*)} \mathbb{P}\left\{\Pi_0(Z|A) > \Pi_0(Z^*|A)\right\},
\end{aligned}$$

where the last inequality is due to symmetry. We also have

$$|\{\Gamma : \exists Z \in \Gamma, s.t.\ d(Z,Z^*) = m\}| \leq \binom{n}{m}(K-1)^m \leq \min\left\{\left(\frac{enK}{m}\right)^m, K^n\right\}.$$

Note that $\{Z : Z \notin \Gamma(Z^*)\}$ is equivalent to set $\{Z : m(Z) \geq 1\}$. With the condition that $\bar{n}I > \log n$, it follows by Lemma 2.4.1 that

$$\begin{aligned}\mathbb{E}\Pi_0(Z \notin \Gamma(Z^*)|A) &\leq 2\sum_{1\leq m\leq n/2K}\binom{n}{m}K^m \exp\left(-(\bar{n}m - m^2)I\right) + 2\sum_{m>n/2K} K^n \exp\left(-d_{\alpha,\beta}mnI/K\right) \\ &\leq 2\sum_{1\leq m\leq n/2K}\binom{n}{m}K^m \exp\left(-(\bar{n}m - m^2)I\right) + 2nK^n \exp\left(-Cn^2 I\right) \qquad (27)\end{aligned}$$

for some constant C. We proceed to upper bound the first term in (27). It follows that

$$\sum_{1\leq m\leq n/2K}\binom{n}{m}K^m \exp\left(-(\bar{n}m - m^2)I\right) \leq \sum_{1\leq m\leq n/2K}(enK)^m \exp\left(-(\bar{n}m - m^2)I\right) = \sum_m P_m$$

where $P_m = (enK)^m \exp\left(-(\bar{n}m - m^2)I\right)$. The ratio of P_m and P_1 is calculated as

$$\frac{P_m}{P_1} = (enK)^{m-1}\exp(-\bar{n}I(m-1) + (m^2-1)I) = \left(enK\exp\left(-\bar{n}I + (m+1)I\right)\right)^{m-1}.$$

Define $m' = \varepsilon' n$ for some positive sequence $\varepsilon' = \varepsilon'_n$ with $\varepsilon' \to 0$ and $\varepsilon' nI \to \infty$. Then, $\sum_{1\leq m\leq n/2K} P_m$ can be split into summation of $\sum_{1\leq m<m'} P_m$ and $\sum_{m'\leq m\leq n/2K} P_m$, where

$$\begin{aligned}\sum_{m=1}^{m'-1} P_m = P_1 \sum_{m=1}^{m'-1}\frac{P_m}{P_1} &\leq P_1 \sum_{m=1}^{m'-1}\left(enK\exp\left(-\bar{n}I + m'I\right)\right)^{m-1} \\ &\leq enK\exp\left(-\bar{n}I + I\right)\cdot\left(1 + 2enK\exp(-\bar{n}I + \varepsilon' nI)\right),\end{aligned}$$

and there exists some constant C such that

$$\sum_{m'<m\leq n/2K} P_m \leq nK^n \exp(-\varepsilon'(C-\varepsilon')n^2 I) \leq \exp(-n).$$

Hence, by combining all parts and based on the condition that $\bar{n}I > \log n$, we have $\Pi_0(Z \notin \Gamma(Z^*)|A) \leq Cn\exp(-\bar{n}I)$ for some constant C and for a large n.

□

Proof of Theorem 4.2.1

Lemma 2.4.2. *Let $Z \in S_\alpha$ be an arbitrary assignment with $d(Z,Z^*) = m > 0$. If $p,q \to 0$ and $p \asymp q$, there exists some positive sequence $\gamma = \gamma_n$ with $\gamma \to 0$ and $\gamma^2 nI \to \infty$, such that for the $\Pi(Z|A)$ defined in* (5)*, we have*

$$\mathbb{P}\left\{\max_{Z\in S_\alpha : m>\gamma n} \log\frac{\Pi(Z|A)}{\Pi(Z^*|A)} \geq -C_1\gamma n^2 I\right\} \leq 4\exp(-n),$$

and

$$\mathbb{P}\left\{\max_{Z\in S_\alpha : m\leq\gamma n} \log\frac{\Pi(Z|A)}{\Pi(Z^*|A)} - \log\frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} - C_2\gamma mnI > 0\right\} \leq n\exp\left(-(1-o(1))\bar{n}I\right),$$

for some constants C_1, C_2. Here, $\Pi_0(Z|A)$ is the posterior probability with known connectivity probabilities.

The proof of Lemma 2.4.2 is deferred later. We now state another lemma that is based on Proposition 5.1 in [89].

Lemma 2.4.3 (Proposition 5.1 in [89])**.** *For any $Z \in S_\alpha$ where $d(Z,Z^*) = m < n/2K$,*

$$\mathbb{E}\sqrt{\frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)}} \leq \exp\left(-\bar{n}mI + m^2 I\right).$$

Proof of Theorem 4.2.1. With Lemma 2.4.2 and Lemma 2.4.3, we divide S_α into a large mistake region and a small mistake region according to whether $m > \gamma n$, where γ is a positive sequence defined in Lemma 2.4.2.

Large mistake region. For $m > \gamma n$, by Lemma 2.4.2, with probability at least $1 - 4\exp(-n)$,

$$\sum_{Z \in S_\alpha : m > \gamma n} \frac{\Pi(Z|A)}{\Pi(Z^*|A)} \leq nK^n \exp(-C_1 \gamma n^2 I) \leq \exp(-n).$$

Denote $\mathscr{E} = \{\sum_{Z \in S_\alpha : m > \gamma n} \frac{\Pi(Z|A)}{\Pi(Z^*|A)} \leq \exp(-n)\}$, and it follows directly that

$$\mathbb{E}\left[\sum_{Z \in S_\alpha : m > \gamma n} \Pi(Z|A)\right] \leq \mathbb{E}\left[\Pi(Z : m > \gamma n|A)\mathbb{I}\{\mathscr{E}\}\right] + \mathbb{P}\{\mathscr{E}^c\} \leq 5\exp(-n).$$

Small mistake region. For $m \leq \gamma n$, let $G_Z = \{\Pi_0(Z|A) > \Pi_0(Z^*|A)\}$. Let θ denote all unknown parameters and θ_0 denote the underlying true parameters respectively. Define

$$\mathscr{F} = \{\max_{Z \in S_\alpha : m \leq \gamma n} \log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} - \log \frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} - C_2 \gamma mnI > 0\}$$

as in Lemma 2.4.2. Then, we have

$$\begin{aligned}
&\mathbb{E}\Pi(Z : 1 \leq m \leq \gamma n|A) \\
\leq & P_{Z^*,\theta_0}\Pi(Z : 1 \leq m \leq \gamma n|A)\mathbb{I}\{\mathscr{F}^c\} + \mathbb{P}\{\mathscr{F}\} \\
\leq & \sum_{Z:1 \leq m \leq \gamma n} P_{Z^*,\theta_0} \frac{\Pi(Z|A)}{\Pi(Z^*|A)} \mathbb{I}\{G_Z^c, \mathscr{F}^c\} + \sum_{Z:1 \leq m \leq \gamma n} P_{Z^*,\theta_0}\mathbb{I}\{G_Z\} + \mathbb{P}\{\mathscr{F}\} \\
\leq & \sum_{Z:1 \leq m \leq \gamma n} P_{Z^*,\theta_0} \frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} \mathbb{I}\{G_Z^c\} \exp\left(C_2 \gamma mnI\right) + \sum_{Z:1 \leq m \leq \gamma n} P_{Z^*,\theta_0}\mathbb{I}\{G_Z\} + \mathbb{P}\{\mathscr{F}\} \\
\leq & \sum_{Z:1 \leq m \leq \gamma n} \exp\left(C_2 \gamma mnI\right) P_{Z,\theta_0}\mathbb{I}\{G_Z^c\} + \sum_{Z:1 \leq m \leq \gamma n} P_{Z^*,\theta_0}\mathbb{I}\{G_Z\} + \mathbb{P}\{\mathscr{F}\} \\
\leq & n\exp\left(-(1 - o(1))\bar{n}I\right).
\end{aligned}$$

Recall that $\Pi_0(\cdot|A)$ denotes the posterior distribution with knowledge of the connectivity probabilities. The second inequality is due to $\Pi(Z^*|A) \leq 1$. The third inequality is due to the definition of the event $\mathscr{F}$. The last two inequalities hold by Lemma 2.4.1 and symmetry.

Combine the two regions, and then

$$\mathbb{E}\Pi(Z \notin \Gamma(Z^*)|A) \leq n\exp(-(1-o(1))\bar{n}I).$$

The proof is complete. □

2.4.2 Proofs of Theorem 2.2.1 and Theorem 2.2.3

Backgrounds on mixing time

Consider a reversible, irreducible, and aperiodic Markov chain on a discrete space Ω that is completely specified by a transition matrix $P \in [0,1]^{|\Omega|\times|\Omega|}$ with stationary distribution Π. Let $\omega \in \Omega$ be the initial state of the chain, and then the total variation distance to the stationary distribution after t iterations is

$$\Delta_\omega(t) = \left\| P^t(\omega,\cdot) - \Pi \right\|_{\mathsf{TV}},$$

where $P^t(\omega,\cdot)$ is the distribution of the chain after t iterations. The ε-mixing time starting at ω is given by

$$\tau_\varepsilon(\omega) = \min\left\{ t \in \mathbb{N} : \Delta_\omega(t') \leq \varepsilon \text{ for all } t' \geq t \right\}.$$

With this notation, we say a Markov chain is rapidly mixing if $\tau_\varepsilon(\omega)$ is $O\left(\text{poly}(\log(|\Omega|/\varepsilon))\right)$ in the case where $|\Omega|$ scales exponentially to the problem size n. This means we only need to update the Markov chain for $\text{poly}(n)$ steps in order to obtain *good* samples from the stationary distribution.

The explicit bound for the mixing time through the spectral gap is

$$\tau_\varepsilon(\omega) \leq \frac{-\log \Pi(\omega) + \log(1/\varepsilon)}{Gap(P)}, \tag{28}$$

where $Gap(P)$ represents the spectral gap of the transition matrix P, defined by $Gap(P) = 1 - \max\{|\lambda_2(P)|, |\lambda_{\min}(P)|\}$, where $\lambda_2(P)$, $\lambda_{\min}(P)$ are the second largest and the smallest eigenvalues of the transition matrix P. See the paper [82] for this bound.

Preparation

Suppose $P(\cdot,\cdot)$ in (9) is the transition matrix introduced in Algorithm 1 defined in the label assignment space S_α, and $\check{P}(\cdot,\cdot)$ in (14) is the transition matrix of $\{\Gamma_t\}_{t\geq 0}$ defined in the clustering space $\check{S}_\alpha = \{\Gamma(Z) : Z \in S_\alpha\}$. The stationary distribution for P and $\check{P}$ are denoted as $\widetilde{\Pi}$ and $\check{\Pi}$ respectively. We require a good initializer, and use the following lemma to guarantee that all possible states visited by the algorithm remain in a good region with high probability.

Lemma 2.4.4. *Suppose we start at a fixed initializer Z_0 with $\ell(Z_0, Z^*) \leq \gamma_0$ where γ_0 satisfies Condition 2.2, 2.4, or 2.5. Then, the number of misclassified nodes in any polynomial running time can be upper bounded by*

$$m = n \cdot \ell(Z, Z^*) \leq n \max\left\{\gamma_0, n^{-\tau}\right\} + \log^2 n, \tag{29}$$

with probability at least $1 - \exp(-\log^2 n)$, where $\tau > 0$ is a sufficiently small constant.

The proof of Lemma 2.4.4 is deferred to Section A.1.1. Note that Lemma 2.4.4 is stated conditioning on a fixed initial label assignment Z_0 with $\ell(Z_0, Z^*) \leq \gamma_0$. This is slightly different from the original initialization conditions where we use Z_0 dependent on data. A simple union bound will lead to the final conclusion. Lemma 2.4.4 quantifies the maximum possible number of classification mistakes when starting at a good initializer. Here, $(\log n)^2$ is chosen for simplicity and can be replaced by any sequence $\nu_n \gg \log n$.

Let $\mathscr{G}(\gamma_0)$ denote a good region with respect to the initial misclassification proportion, defined by

$$\mathscr{G}(\gamma_0) = \{Z \in S_\alpha : m \le n\max\{\gamma_0, n^{-\tau}\} + \log^2 n\}, \tag{30}$$

where τ is a sufficiently small constant. Accordingly, we can define a good region in the clustering space as

$$\check{\mathscr{G}}(\gamma_0) = \{\Gamma(Z) : Z \in \mathscr{G}(\gamma_0)\},$$

and Lemma 2.4.4 ensures that for any T that is a polynomial of n, $\{\Gamma_t\}_{0\le t\le T}$ stays inside $\check{\mathscr{G}}(\gamma_0)$ with high probability. Sometimes we write $\mathscr{G}(\gamma_0)$ and $\check{\mathscr{G}}(\gamma_0)$ as $\mathscr{G}$ and $\check{\mathscr{G}}$ for simplicity. Then, we modify the distributions and transition matrices according to the regions $\mathscr{G}$ and $\check{\mathscr{G}}$. Denote the modified distributions as $\widetilde{\Pi}_g(Z|A) \propto \Pi^{\xi}(Z|A)\mathbb{I}\{Z \in \mathscr{G}\}$ for all $Z \in S_\alpha$, and $\check{\Pi}_g(\Gamma|A) \propto \check{\Pi}(\Gamma|A)\mathbb{I}\left\{\Gamma \in \check{\mathscr{G}}\right\}$ for all $\Gamma \in \check{S}_\alpha$. Define in the label assignment space the new transition matrix $P_g(\cdot,\cdot)$ corresponding to $\widetilde{\Pi}_g(\cdot|A)$, by replacing $\Pi^{\xi}(\cdot|A)$ with $\widetilde{\Pi}_g(\cdot|A)$ in (9). Define in the clustering space the new transition matrix $\check{P}_g(\cdot,\cdot)$ corresponding to $\check{\Pi}_g(\cdot|A)$, by replacing $\Pi(\cdot|A)$ with $\Pi(\cdot|A)\mathbb{I}\left\{\cdot \in \check{\mathscr{G}}\right\}$ in (14).

With these notations, we proceed to bound the total variation error between the distribution of Γ_T and $\check{\Pi}(\cdot|A)$ after T steps for some T that is a polynomial of n.

Lemma 2.4.5 (TV difference)**.**

$$\mathbb{E}\left\|\check{\Pi}_g(\cdot|A) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}} \le n\exp(-(1-o(1))\bar{n}I).$$

Proof. We have

$$\mathbb{E}\left\|\check{\Pi}_g(\cdot|A) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}} \le 2\mathbb{E}\check{\Pi}(\Gamma \notin \check{\mathscr{G}}|A) \le 2\mathbb{E}\check{\Pi}(\Gamma \ne \Gamma(Z^*)|A) \le n\exp(-(1-o(1))\bar{n}I),$$

The second inequality is due to the definition of $\check{\mathscr{G}}$. The last inequality directly follows by Theorem 4.2.1 or Theorem 2.2.3, and the condition that $\xi \ge 1$. □

Thus, by triangle inequality, we can decompose the total variation bound at time T as

$$\begin{aligned}&\left\|\check{P}^T(\Gamma_0,\cdot)-\check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}}\\ \leq&\left\|\check{P}^T(\Gamma_0,\cdot)-\check{P}_g^T(\Gamma_0,\cdot)\right\|_{\mathsf{TV}}+\left\|\check{P}_g^T(\Gamma_0,\cdot)-\check{\Pi}_g(\cdot|A)\right\|_{\mathsf{TV}}+\left\|\check{\Pi}_g-\check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}},\end{aligned} \tag{31}$$

where T is the number of iterations. Lemma 2.4.4 implies that the first term is 0 with high probability for $T \leq \text{poly}(n)$, since the algorithm stays in the region $\mathscr{G}(\gamma_0)$. The third term can be upper bounded by Lemma 2.4.5. Therefore, the remaining proof is to adopt the canonical path approach to bound the second term in (31).

For the purpose of the proof, we replace the transition matrix $\check{P}_g$ by its lazy version, which has a probability of 1/2 at staying at its current state, and another probability of 1/2 at updating the state. The same technique can be also found in [85, 50, 8, 56]. It is worth noting that this technique is only for the proof.

Canonical path

Given an ergodic Markov chain $\mathscr{C}$ induced by the lazy transition matrix $\check{P}_g$ in the discrete state space $\check{\mathscr{G}}$, we define a weighted directed graph $G(\mathscr{C}) = (V,E)$, where the vertex set $V = \check{\mathscr{G}}$ and an edge between an ordered pair (Γ,Γ') is included in E with weight $Q(\Gamma,\Gamma') = \check{\Pi}_g(\Gamma)\check{P}_g(\Gamma,\Gamma')$ whenever $\check{P}_g(\Gamma,\Gamma') > 0$. A canonical path ensemble $\mathscr{T}$ is a collection of simple paths $\{\mathscr{T}_{x,y}\}$ in the graph $G(\mathscr{C})$, one between each ordered pair (x,y) of distinct vertices. As shown in [73], for any choice of canonical path $\mathscr{T}$, the spectral gap of the transition matrix $\check{P}_g$ can be lower bounded by

$$Gap(\check{P}_g) \geq \frac{1}{\rho(\mathscr{T})\ell(\mathscr{T})},$$

where $\ell(\mathscr{T})$ is the length of the longest path in $\mathscr{T}$, and $\rho(\mathscr{T})$ is the *path congestion* parameter defined by $\rho(\mathscr{T}) = \max_{(\Gamma,\Gamma')\in E(\mathscr{C})} \frac{1}{Q(\Gamma,\Gamma')} \sum_{\mathscr{T}_{x,y}\ni(\Gamma,\Gamma')} \check{\Pi}_g(x)\check{\Pi}_g(y)$.

In order to apply the approach, we need to construct an appropriate canonical path ensemble

$\mathscr{T}$ in the discrete state space $\check{\mathscr{G}}$. First, we construct a unique canonical path from any clustering Γ to the underlying true clustering Γ^*, where $\Gamma^* = \Gamma(Z^*)$. Suppose for any label assignment Z, we define a function $g : S_\alpha \to S_\alpha$ such that

$$g(Z) = \begin{cases} \underset{Z' \in \mathscr{B}(Z) \cap S_\alpha}{\arg\max} \ \Pi(Z'|A), & \text{if } Z \notin \Gamma(Z^*), \\ Z, & \text{if } Z \in \Gamma(Z^*), \end{cases} \tag{32}$$

where

$$\mathscr{B}(Z) = \{Z' : d(Z', Z^*) = d(Z, Z^*) - 1, H(Z, Z') = 1\}.$$

We use $\mathscr{B}(Z) \cap S_\alpha$ to denote the set of available states that have fewer mistakes than the current state Z. Here, $\mathscr{B}(Z) \cap S_\alpha$ is always non-empty for $Z \notin \Gamma(Z^*)$. Here, $g(Z)$ is the *optimal* state in $\mathscr{B}(Z)$ in the sense that $g(Z)$ maximizes the posterior distribution. Then, for any current state $\Gamma \in \check{S}_\alpha$, we define the next state $\check{g}(\Gamma)$ to be

$$\check{g}(\Gamma) = \{g(Z) : Z \in \Gamma\}$$

Since for any $Z \in \Gamma$, $g(Z)$ gives the equivalent result, and thus $\check{g}(\Gamma) \in \check{S}_\alpha$ is well defined. Hence, the canonical path from any current state $\Gamma \neq \Gamma(Z^*)$ is a greedy path, and the number of mistakes keeps decreasing along the canonical path.

Second, we construct a unique canonical path between any two states Γ and $\widetilde{\Gamma}$, defined by $\mathscr{T}_{\Gamma,\widetilde{\Gamma}} = \mathscr{T}_{\Gamma,\Gamma^*} - \mathscr{T}_{\widetilde{\Gamma},\Gamma^*}$. The operations on simple paths are the same as defined in [1]. It is worth noting that the construction of the canonical path is data dependent, i.e., for different adjacency matrix A, the construction of the canonical path might be different.

Let $\Lambda(\Gamma) = \left\{\widetilde{\Gamma} \in \check{\mathscr{G}} : \Gamma \in \mathscr{T}_{\widetilde{\Gamma},\Gamma^*}\right\}$ denote the set of all precedents states before Γ along the canonical path. Let $\mathscr{E} = \left\{(\Gamma, \Gamma') \in E(\mathscr{C}) : \Gamma \in \Lambda(\Gamma')\right\}$ denote the ordered adjacent pairs along the

canonical path. It follows that

$$\begin{aligned}\rho(\mathscr{T}) &= \max_{(\Gamma,\Gamma')\in E(\mathscr{C})} \frac{1}{Q(\Gamma,\Gamma')} \sum_{\mathscr{T}_{x,y}\ni(\Gamma,\Gamma')} \check{\Pi}_g(x)\check{\Pi}_g(y) \\ &\le \max_{(\Gamma,\Gamma')\in\mathscr{E}} \frac{1}{Q(\Gamma,\Gamma')} \sum_{x\in\Lambda(\Gamma),y\in\check{\mathscr{G}}} \check{\Pi}_g(x)\check{\Pi}_g(y) \\ &= \max_{(\Gamma,\Gamma')\in\mathscr{E}} \frac{\check{\Pi}_g(\Lambda(\Gamma))}{Q(\Gamma,\Gamma')} = \max_{\Gamma\in\check{\mathscr{G}}} \frac{\check{\Pi}_g(\Lambda(\Gamma))}{Q(\Gamma,\check{g}(\Gamma))},\end{aligned}$$

where we simply take maximum only over all states in the discrete space $\check{\mathscr{G}}$. By the definition of Algorithm 1 and the lazy transition matrix, $Q(\Gamma,\Gamma')$ can be expressed as

$$Q(\Gamma,\check{g}(\Gamma)) = \check{\Pi}_g(\Gamma)\check{P}_g(\Gamma,\check{g}(\Gamma)) = \frac{1}{2(K-1)n}\min\left\{\check{\Pi}_g(\Gamma),\check{\Pi}_g(\check{g}(\Gamma))\right\}.$$

It leads to the bound for the *congestion parameter* as

$$\begin{aligned}\rho(\mathscr{T}) &\le 2(K-1)n\max_{\Gamma\in\check{\mathscr{G}}} \frac{\check{\Pi}_g(\Lambda(\Gamma))}{\min\left\{\check{\Pi}_g(\Gamma),\check{\Pi}_g(\check{g}(\Gamma))\right\}} \\ &= 2(K-1)n\max_{\Gamma\in\check{\mathscr{G}}} \frac{\check{\Pi}(\Lambda(\Gamma)|A)}{\min\left\{\check{\Pi}(\Gamma|A),\check{\Pi}(\check{g}(\Gamma)|A)\right\}} \\ &= 2(K-1)n\max_{\Gamma\in\check{\mathscr{G}}} \left\{\frac{\check{\Pi}(\Lambda(\Gamma)|A)}{\check{\Pi}(\Gamma|A)}\cdot\max\left\{1,\frac{\check{\Pi}(\Gamma|A)}{\check{\Pi}(\check{g}(\Gamma)|A)}\right\}\right\},\end{aligned}$$

where $\check{\Pi}(\Gamma|A) = \sum_{Z\in\Gamma}\widetilde{\Pi}(Z|A)$ for all $\Gamma\in\check{S}_\alpha$.

Lemma 2.4.6. *Recall that $g(Z)$ is the next optimal state of Z defined in* (32)*, and $\mathscr{G}(\gamma_0)$ is defined in* (30)*. Suppose Z_0 is given with $\ell(Z_0,Z^*)\le\gamma_0$ where γ_0 satisfies Condition 2.2, 2.4, or 2.5. Suppose ξ satisfies Condition 2.3, 2.4 or 2.5. Then, we have*

$$\max_{Z\in\mathscr{G}(\gamma_0)} \frac{\widetilde{\Pi}(Z|A)}{\widetilde{\Pi}(g(Z)|A)} \le \exp(-C\bar{n}I)$$

for some constant $C > 1-\varepsilon_0$ with probability at least $1-C_1n^{-C_2}$.

The proof of Lemma 2.4.6 is deferred to Section A.1.4 and Section A.1.5. By Lemma 2.4.6

and by permutation symmetry, we have

$$\max_{\Gamma\in\check{\mathscr{G}}}\frac{\check{\Pi}(\Gamma|A)}{\check{\Pi}(\check{g}(\Gamma)|A)}=\max_{Z\in\mathscr{G}}\frac{\widetilde{\Pi}(Z|A)}{\widetilde{\Pi}(g(Z)|A)}\leq\exp(-C\bar{n}I)\leq\exp(-C\bar{n}I),$$

for some constant $C>1-\varepsilon_0$ with high probability. Denote $\widetilde{m}=n\max\{\gamma_0,n^{-\tau}\}+\log^2 n$ for simplicity, and it follows that

$$\begin{aligned}\max_{\Gamma\in\check{\mathscr{G}}}\frac{\check{\Pi}(\Lambda(\Gamma)|A)}{\check{\Pi}(\Gamma|A)}\leq&1+\max_{m\leq\widetilde{m}}\sum_{l=1}^{\widetilde{m}-m}\binom{n-m}{l}(K-1)^l\exp(-Cl\bar{n}I)\\ \leq&1+C'n\exp(-C\bar{n}I),\end{aligned}$$

for some constants C' and $C>1-\varepsilon_0$. Then, we have

$$\begin{aligned}\rho(\mathscr{T})&\leq 2(K-1)n\max_{\Gamma\in\check{\mathscr{G}}}\left\{\frac{\check{\Pi}(\Lambda(\Gamma)|A)}{\check{\Pi}(\Gamma|A)}\cdot\max\left\{1,\frac{\check{\Pi}(\Gamma|A)}{\check{\Pi}(g(\Gamma)|A)}\right\}\right\}\\ &\leq 2(K-1)n(1+C'n\exp(-C\bar{n}I)).\end{aligned}$$

Furthermore, since the canonical path is defined within $\check{\mathscr{G}}$, we can upper bound the length of the longest path by

$$\ell(\mathscr{T})\leq 2n\max\{\gamma_0,n^{-\tau}\}+2\log^2 n.$$

Recall that $\Gamma_0=\Gamma(Z_0)$. By Lemma 2.4.4 and Lemma 2.4.5, together with (28) and the strong consistency property of $\check{\Pi}_g(\cdot|A)$, we have that for any constant $\varepsilon\in(0,1)$,

$$\left\|\check{P}^T(\Gamma_0,\cdot)-\check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}}\leq\varepsilon, \tag{33}$$

holds for any

$$T\geq 4(K-1)n^2\max\{\gamma_0,n^{-\tau}\}\left(-\xi\log\Pi(Z_0|A)+\log\varepsilon^{-1}\right)(1+o(1)), \tag{34}$$

for large n with probability at least $1 - C_3 n^{-C_4}$ for some constants C_3, C_4, where $\Pi^{\xi}(Z_0|A) \leq \check{\Pi}_g(\Gamma_0|A)$ always holds. Finally, if $\mathbb{P}\{\ell(Z_0, Z^*) \leq \gamma_0\} \geq 1 - \eta$, then the conclusions of Theorem 2.2.1 and Theorem 2.2.3 can be obtained by a simple union bound argument.

Coupling

We require T to be at most a polynomial of n so that Lemma 2.4.4 holds. Thus, the previous total variation bound (33) holds only for $T \leq \text{poly}(n)$. In order to bound the mixing time defined in (15), we further use coupling approach to show the total variation bound holds for any $t \geq T$.

We call a probability measure w over $\Omega \times \Omega$ is a coupling of (u, v) if its two marginals are u and v respectively. Before the proof, we first state the following lemma to relate the total variation to the coupling.

Lemma 2.4.7 (Proposition 4.7 in [50]). *For any coupling w of (u, v), if the random variables (X, Y) is distributed according to w, then we have*

$$\|u - v\|_{\mathsf{TV}} \leq \mathbb{P}\{X \neq Y\}.$$

Back to our problem, in order to upper bound $\left\|\check{P}^t(\Gamma_0, \cdot) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}}$ for any $t \geq T$, we first create a coupling of these two distributions as follows. Consider two copies of the Markov chain X_t and Y_t both with the transition matrix $\check{P}$:

- Let $X_0 = \Gamma_0$, and $Y_0 \sim \check{\Pi}(\cdot|A)$.
- If $X_t \neq Y_t$, then sample X_{t+1} and Y_{t+1} independently according to $\check{P}(X_t, \cdot)$ and $\check{P}(Y_t, \cdot)$ respectively.
- If $X_t = Y_t$, then sample $X_{t+1} \sim \check{P}(X_t, \cdot)$ and set $Y_{t+1} = X_{t+1}$.

Thus, it is obviously that for any $t \geq 1$, $Y_t \sim \check{\Pi}(\cdot|A)$, and $X_t \sim \check{P}^t(\Gamma_0, \cdot)$. Set

$$T = 4Kn^2 \max\{\gamma_0, n^{-\tau}\} \left(-\xi \log \check{\Pi}(Z_0|A) + \log \varepsilon^{-1}\right)(1 + o(1))$$

defined in (34). By Lemma 2.4.7 and (34), we have for any $t \geq T$,

$$\begin{aligned}\left\|\check{P}^t(\Gamma_0,\cdot) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}} &\leq \mathbb{P}\{X_t \neq Y_t\} \leq \mathbb{P}\{X_T \neq Y_T\} \\ &= 1 - \mathbb{P}\{X_T = Y_T\} \\ &\leq 1 - \mathbb{P}\{X_T = Y_T = \Gamma^*\} \\ &\leq 2 - \mathbb{P}\{X_T = \Gamma^*\} - \mathbb{P}\{Y_T = \Gamma^*\}.\end{aligned}$$

By (33), we have

$$\left\|\check{P}^T(\Gamma_0,\cdot) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}} = \max_S \left|\check{P}^T(\Gamma_0,S) - \check{\Pi}(S|A)\right| \leq \varepsilon$$

with high probability. Together with the strong consistency result, it yields

$$\begin{aligned}\left\|\check{P}^t(\Gamma_0,\cdot) - \check{\Pi}(\cdot|A)\right\|_{\mathsf{TV}} &\leq 2 - \mathbb{P}\{X_T = \Gamma^*\} - \mathbb{P}\{Y_T = \Gamma^*\} \\ &\leq 1 - \left(\check{\Pi}(\Gamma^*|A) - \varepsilon\right) - \check{\Pi}(\Gamma^*|A) \\ &\leq \varepsilon(1 + o(1)),\end{aligned}$$

with probability at least $1 - Cn^{-C'}$ for some constants C, C'. Here, the high probability statement is with respect to the data generation process, i.e., adjacency matrix A.

To combine, we reach the result that for any constant $\varepsilon \in (0,1)$,

$$\tau_\varepsilon(Z_0) \leq 4Kn^2 \max\left\{\gamma_0, n^{-\tau}\right\} \cdot \left(\xi \log \Pi(Z_0|A)^{-1} + \log(\varepsilon^{-1})\right),$$

with high probability where $\tau_\varepsilon(Z_0)$ is defined in (15).

CHAPTER 3

UNADJUSTED LANGEVIN MONTE CARLO VIA TWEEDIE'S FORMULA

In the area of Bayesian computation, Markov Chain Monte Carlo (MCMC) is the most popular technique, in which the equilibrium distribution of the Markov chain matches the target distribution. Most recent works have been done on proposing novel MCMC schemes and studying the computational efficiency [24, 39, 68, 90], where the central interest is to analyze the mixing time of Markov chain, also referred to as the convergence rate. However, the non-asymptotic analyses of mixing time with respect to dimension depends heavily on the smoothness of distribution, which is usually not satisfied in sparse Bayesian inference with non-smooth prior distribution.

Among a great variety of MCMC techniques, Langevin Monte Carlo (LMC) is one of the most famous method that is derived from stochastic Langevin diffusion process. The continuous dynamic diffusion process was first proposed in [71, 61], and the corresponding stochastic differential equation (SDE) is constructed based on the negative log-density function (referred to as *potential function*) of the target distribution. Under some general conditions, the SDE admits one stochastic solution that converges to the target distribution [69]. In practice, Euler-Maruyama discretization is adopted to approximate the stochastic solution [61], referred to as LMC algorithm, which can be viewed as iterative updating procedure to approximately sample from the target distribution. Over the past five years, a surge of research has led to breakthrough in the understanding of the convergence rate of LMC. Dalalyan [23] first established the non-asymptotic convergence analysis in total variation for the LMC algorithm, targeting the log-concave and smooth density function with a good initialization. Durmus and Moulines [27] followed the work in [23] and studied the convergence rate in wasserstein distance. Cheng and Bartlett [18] reformulated the sampling problem as a minimization problem in a probability measure space, and provided the result in terms of Kullback-Leibler divergence. However, nearly all well-established results for LMC algorithm require convex and smooth condition of potential function, which is a big limitation to application.

In this work, we aim to tackle the problem of sampling from non-smooth or even unbounded density function, and provide non-asymptotic convergence analysis under certain conditions. Motivated by Tweedie's formula, we propose a novel LMC sampling scheme where, at each step, the point was updated through a posterior mean calculation. We show that it is equivalent to first smooth the density function, and then apply LMC algorithm. The smoothing scheme studied in this work is referred to as Tweedie's transformation. Tweedie's transformation reveals advantages compared with Moreau-Yosisa envolope studied in [29] and Gaussian smoothing studied in [15]. Moreau-Yosisa envolope (also referred to as proximal method) cannot be applied to an unbounded density function and the smoothness bound requires convexity condition, which restricts its application in area of Bayesian regression with sparse priors. Gaussian smoothing technique yields slower convergence rate than Tweedie's transformation both under Lipschitz condition, since its smoothness constant after smoothing depends on the dimension.

The non-asymptotic convergence analysis is performed in this work in terms of total variation distance. In particular, we focus on the minimum number of iteration steps required to guarantee the total variation distance is less than some predefined threshold ε (also referred to as mixing time in Chapter 2). When the potential function is convex and smooth, the minimum number of iterations in theory scales as $\mathscr{O}\left(d/\varepsilon^2\right)$ with dimension d. Without smoothness condition, we are able to show the number of iterations scales as $\mathscr{O}\left(d^3/\varepsilon^6\right)$ under certain conditions. It is worth noting that the proposed algorithm also works in non-convex cases, but needs to be studied case by case according to its spectral gap.

3.1 Bayesian Formulation and Computation

In this section, we formulate the problem and introduce the Langevin Monte Carlo (LMC) algorithm.

3.1.1 Problem Formulation

In this work, we focus on sampling from a probability distribution on $\mathbb{R}^d$ with density $\mu(x)$ given by

$$\mu(x) = \frac{e^{-U(x)}}{\int_{\mathbb{R}^d} e^{-U(y)} dy},$$

for some measureable function $U : \mathbb{R}^d \to \mathbb{R}$, referred to as the *potential function*. Throughout the chapter, we require U to satisfy the following conditions:

Condition 3.1. $U = f + g$, where $f, g : \mathbb{R}^d \to \mathbb{R}$ satisfy that:

(i) f is strongly convex with positive constant m and smooth with positive constant M, i.e., for any $x, x' \in \mathbb{R}^d$,

$$f(x) - f(x') - \nabla f(x')^T (x - x') \geq \frac{m}{2} \|x - x'\|_2^2,$$
$$\|\nabla f(x) - \nabla f(x')\|_2 \leq M \|x - x'\|_2^2$$

(ii) g is separable, i.e., $g(x) = \sum_{i=1}^d g_i(x_i)$.

We assume f to be strongly convex for simplicity. If f is convex with $m = 0$, we can always add a small convexity into the function and then tune the magnitude to bound the convergence result. Notice that the function g is allowed to be non-smooth, non-convex, and even unbounded. One typical example is sparse Bayesian inference. In this settings, we focus on sampling from the posterior distribution taking the form of $\mu(x) \propto \exp(-f(x) - \log \pi(x))$, where $f(x)$ is the negative log-likelihood function that is smooth and strongly convex, and $\pi(x)$ is some sparse prior that is not necessarily upper bounded.

Due to intractable integration in the denominator, direct inference of $\mu(x)$ is generally infeasible, which leads to Markov Chain Monte Carlo (MCMC). In this work, in order to tackle the non-smoothness and even unboundedness, we propose a new MCMC methodology and establish theoretical convergence rate.

3.1.2 MCMC: Unadjusted Langevin Monte Carlo

One of the most popular MCMC methods is unadjusted Langevin Monte Carlo (LMC), derived from the discretization of over-damped Langevin diffusion. It was first introduced in the physics literature [61], and attracts more attentions in computational statistics community [37, 38]. Suppose we want to sample from $\bar{\mu}(x) \propto \exp(-\bar{U}(x))$ for a general function $\bar{U} : \mathbb{R}^d \to \mathbb{R}$ that is continuously differentiable. Then, the stochastic differential equations (SDE) of corresponding Langevin diffusion process is given by

$$dL_t = -\nabla \bar{U}(L_t)dt + \sqrt{2}dB_t, \tag{1}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. Under additional mild assumptions, this SDE admits a unique strong solution $(L_t)_{t\geq 0}$. In addition if $\int \exp(-\bar{U}(x))dx < \infty$, then the unique invariant distribution of the semi-group associated with the Langevin SDE is given by $\bar{\mu}(x) \propto \exp(-\bar{U}(x))$ [48]. In practice, the discretization of the Langevin difussion is obtained by the Euler-Maruyama discretization scheme, and leads to the discrete time Markov chain $(X_k)_{k\geq 0}$, that for all $k \geq 0$,

$$\text{LMC:} \quad X_{k+1} = X_k - h\nabla \bar{U}(X_k) + \sqrt{2h}\xi_{k+1}, \tag{2}$$

where $h > 0$ is the step size, and $(\xi_k)_{k\geq 1}$ is a sequence of iid d-dimensional standard Gaussian variables. The process $(X_k)_{k\geq 0}$ can be viewed as samples approximately drawn from $\bar{\mu}$. The proposed iterated algorithm is also referred to as Unadjusted Langevin Algorithm (ULA) in some literature. Under the smoothness condition of $\bar{U}(x)$, the divergence between continuous time process and discretized time process can be well bounded through the following lemma.

Lemma 3.1.1 (Discretization approximation, Lemma 2 in [23])**.** *If $\bar{U} : \mathbb{R}^d \to R$ satisifies the second inequality in Condition 3.1 with smoothness constant M, and $x^* \in \mathbb{R}^d$ is one stationary point satisfying $\nabla \bar{U}(x^*) = 0$. For any $k \geq 0$, let $\mathbb{P}_L^{\nu_0,kh}$ and $\mathbb{P}_X^{\nu_0,kh}$ be respectively the distributions of Langevin diffusion after time kh and its discretization process after k steps, with the same initial*

value $X_0 \sim \nu_0$. Then, if the step size $h \leq \frac{1}{2M}$, it holds that

$$D\left(\mathbb{P}_L^{\nu_0,kh} \middle\| \mathbb{P}_X^{\nu_0,kh}\right) \leq \frac{M^3h^2}{6}\left(\mathbb{E}_{\nu_0}\left[\|X_0 - x^*\|_2^2\right] + 2khd\right) + \frac{dkM^2h^2}{4}.$$

By Lemma 3.1.1, the divergence between two processes only depends on the smoothness of potential function and the initialization.

3.2 Unadjusted Langevin Monte Carlo with Tweedie's Transformation

We present our novel sampling algorithm in this section, which combines LMC with Tweedie's transformation to tackle the non-smooth or even unbounded issue. We first introduce Tweedie's transformation through a toy example, study its theoretical properties and present some graphical results in various cases. Then, we propose main algorithm along with some theoretical convergence guarantee.

3.2.1 Tweedie's Transformation

Let $g : \mathbb{R}^d \to \mathbb{R}$ be any measurable function with $\int_{\mathbb{R}^d} e^{-g(x)}dx < \infty$, and define the Tweedie's transformation of g as

$$g^\lambda(x) = -\log \int e^{-g(y) - \frac{\|x-y\|^2}{2\lambda}} (2\pi\lambda)^{-d/2} dy = -\log \mathbb{E}\left[e^{-g(x-\sqrt{\lambda}z)}\right], \text{ where } z \sim \mathcal{N}(0, I_d). \quad (3)$$

The transformation stems from Gaussian kernel convolution, and we have the following lemma.

Lemma 3.2.1. *Suppose $y \in \mathbb{R}^d$ follows distribution $q(y) \propto e^{-g(y)}$, $z \sim \mathcal{N}(0, I_d)$, $y \perp\!\!\!\perp z$ and define $x = y + \sqrt{\lambda}z$. Let $g^\lambda(x)$ be defined as in* (3)*. Then, it follows that*

(i) Invariant normalization constant: $\int \exp\left(-g^\lambda(x)\right) dx = \int \exp\left(-g(x)\right) dx$;

(ii) Convexity preservation: if $g(y)$ is convex, then $g^\lambda(x)$ is convex;

(iii) $g^\lambda(x) \in C^\infty$, i.e., $g^\lambda(x)$ is infinite smooth;

(iv) Tweedie's formula: $\nabla g^{\lambda}(x) = \frac{1}{\lambda}(x - \mathbb{E}[y|x])$. *When g is differentiable, we further have* $\nabla g^{\lambda}(x) = \mathbb{E}[\nabla g(y)|x]$;

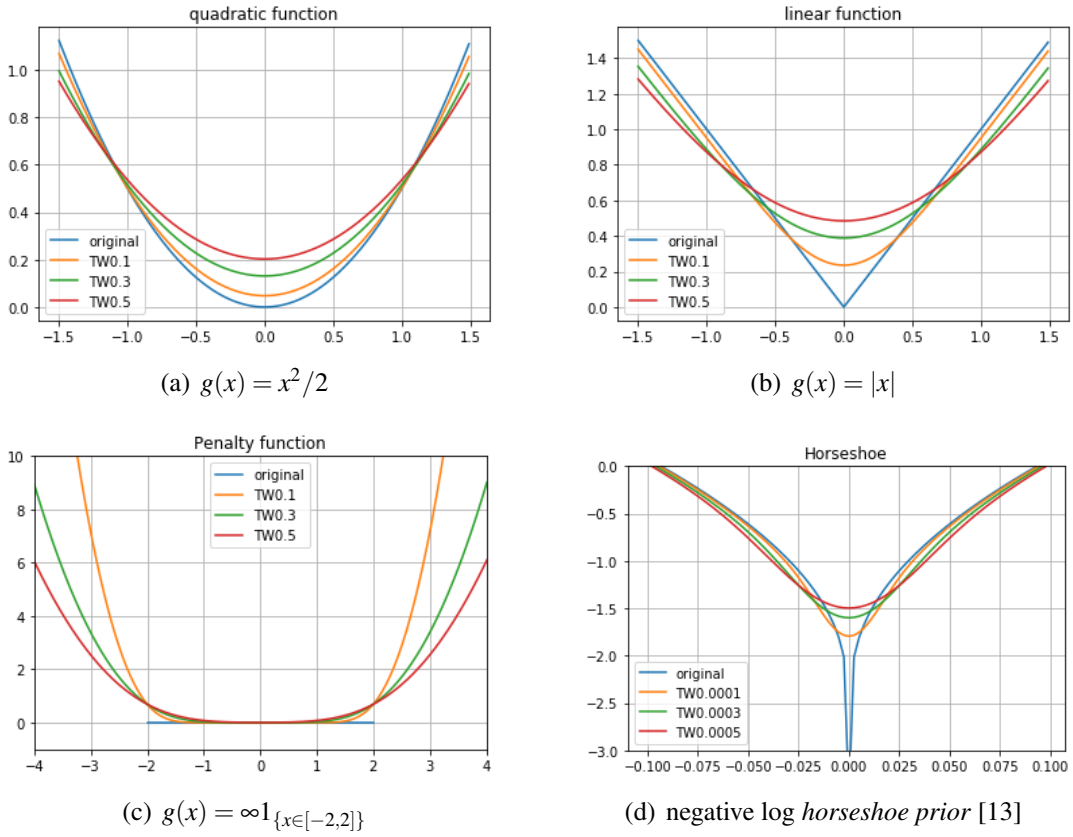
(v) Smoothness: $\nabla^2 g^{\lambda}(x) = \frac{1}{\lambda} I_d - \frac{1}{\lambda}\mathrm{Cov}(z|x)$. *When g is twice differentiable, we further have* $\nabla^2 g^{\lambda}(x) = \mathbb{E}\left[\nabla^2 g(y)|x\right] - \mathrm{Cov}\left(\nabla g(y)|x\right)$;

The proof is deferred to Section 3.4.2. In Lemma 3.2.1, $g^{\lambda}(x)$ stands for the negative log-likelihood of x after Gaussian convolution, and its derivative recovers *Tweedie's formula* [30]. When $\lambda \to 0$, $g^{\lambda}(x)$ point-wisely converges to $g(x)$. In addition, from (i,v) in Lemma 3.2.1, it follows that $\left\|\nabla^2 g^{\lambda}(x)\right\|_2 \leq \frac{1}{\lambda}$ always holds if $g(x)$ is convex. Worth noting that from (v) in Lemma 3.2.1, when g is convex and smooth with smoothness constant M_g, then it directly follows that $g^{\lambda}(x)$ is convex and smooth with smoothness constant $\min\left\{1/\lambda, M_g\right\}$. Hence, Tweedie's transformation in (3) provides a smooth approximation for g and we can make it arbitrarily close to g by tunning the value of λ.

We pause here to provide some toy examples to better understand the Tweedie's transformation. Results are collected in Figure 3.1. For the sake of clear presentation, all transformations are applied onto one-dimensional functions.

From Figure 3.1, we observe that Tweedie's transformation works well to various kinds of functions, and all functions after transformation are smooth, even if the function itself is not upper bounded (as in plot (c)) or lower bounded (as in plot (d)). As λ becomes smaller, g^{λ} is getting closer to g. As shown in plots (a,b,c), when original function g is convex, the resulting g^{λ} is also convex, which coincides with Lemma 3.2.1. We will discuss in detail about horseshoe function in Section 3.3.

From computational point of view, if g is separable, i.e., $g(x) = \sum_{i=1}^{n} g_i(x_i)$, then Tweedie's transformation only involves one-dimensional integration, which is tractable in practice.



(a) $g(x) = x^2/2$

(b) $g(x) = |x|$

(c) $g(x) = \infty 1_{\{x \in [-2,2]\}}$

(d) negative log *horseshoe prior* [13]

Figure 3.1: Tweedie transformation applied onto quadratic function, absolute function, constraint function, and log horseshoe function. The blue line in each plot represents the original function g, and other lines with different colors represent g^λ using different values of λ. The formula of horseshoe prior is given by (4), and here we take $\tau = 0.1$.

3.2.2 Proposed LMC with Tweedie's Transformation

Recall that we want to sample from $\mu(x) \propto \exp(-U(x))$, where potential function $U = f + g$ satisfying Condition 3.1. When g is not smooth or even unbounded, we focus on smoothed version $U^\lambda = f + g^\lambda$ via Tweedie's transformation on g. It leads to an alternative target density $\mu^\lambda \propto \exp\left(-U^\lambda\right)$. Combined with LMC in (2) and Tweedie's formula in Lemma 3.2.1, we then propose the following novel algorithm.

Algorithm 2: Unadjusted Langevin Monte Carlo via Tweedie's Formula (TDLMC)

Input: Initialization $X_0 \sim \nu_0$,

Smoothing level λ,

Step size h,

Number of iteration steps K.

for each $k \in \{0, 1, 2, \ldots, K\}$ **do**

Draw $\xi_{k+1} \sim \mathcal{N}(0, I_d)$;

Update X_{k+1} according to

$$X_{k+1} = \left(1 - \frac{h}{\lambda}\right) X_k - h\nabla f(X_k) + \frac{h}{\lambda}\mathbb{E}[y|X_k] + \sqrt{2h}\xi_{k+1},$$

where

$$p(y|X_k) \propto \exp\left(-g(y) - \frac{\|X_k - y\|_2^2}{2\lambda}\right).$$

We refer to Algorithm 2 as TDLMC for simplicity. The crucial part of TDLMC algorithm is appropriately choosing value of λ. When λ is small enough, the divergence between distributions μ and μ^λ is negligible, while λ also controls the smoothness of potential function U^λ, which further controls the accuracy of approximation with discrete time Langevin process by Lemma 3.1.1. Notice that when $\lambda \to 0$, TDLMC reduces to the original LMC algorithm.

We further compare the TDLMC algorithm with MYULA algorithm proposed in [29]. Worth noting that TDLMC is using posterior mean in each update, and MYULA is using posterior mode to update. Hence, MYULA does not apply to some cases where g is not lower bounded.

3.2.3 Main Results

In this section, we present a detailed theoretical analysis of TDLMC algorithm with fixed smoothing level λ and step size h. Recall that the original potential function is $U = f + g$ where f satisfies Condition 3.1. Without smoothness in g, we apply Tweedie's transformation on g to obtain g^λ as

defined in (3), and construct smoothed potential function $U^\lambda = f + g^\lambda$. Recall that $\mu \propto e^{-U}$ is the target distribution, and thus $\mu^\lambda \propto e^{-U^\lambda}$ serves as a surrogate distribution. In order to construct the convergence rate of sampling from μ, the key points are to bound the divergence between μ and μ^λ, as well as to show the convergence rate of sampling from μ^λ.

To begin with, we study the convergence rate in convex case.

Condition 3.2. $g : \mathbb{R}^d \to \mathbb{R}$ is convex.

Condition 3.3. $g : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz, i.e., $|g(x) - g(y)| \le L\|x - y\|$ for any $x, y \in \mathbb{R}^d$.

Condition 3.1 and Condition 3.2 guarantee that U^λ is strongly convex by Lemma 3.2.1. Condition 3.3 ensures that the total variation between μ and μ^λ are well bounded by the following lemma.

Lemma 3.2.2. *Suppose Condition 3.3 holds with $L\sqrt{\lambda d} \le 1/2$, then we have that*

$$\sup_x \left| g(x) - g^\lambda(x) \right| \le 28L\sqrt{\lambda d},$$

and

$$\left\| \mu - \mu^\lambda \right\|_{\mathsf{TV}} \le CL\sqrt{\lambda d}$$

for some universal constants $C > 0$.

Proof. By the definition of $g^\lambda(x)$, we have

$$\left| g(x) - g^\lambda(x) \right| = \left| \log \mathbb{E}\left[e^{g(x) - g(x - \sqrt{\lambda} z)} \right] \right|, \text{ where } z \sim \mathcal{N}(0, I_d).$$

It follows that

$$\sup_x \left| g(x) - g^\lambda(x) \right| \le \log \mathbb{E}\left[e^{\left| g(x) - g(x - \sqrt{\lambda} z) \right|} \right] \le \log \mathbb{E}\left[e^{L\sqrt{\lambda}\|z\|} \right] \le 28L\sqrt{\lambda d},$$

where the last inequality holds by Lemma 3.4.1. Then, we have for any $x \in \mathbb{R}^d$,

$$\begin{aligned}\log \frac{\mu^\lambda(x)}{\mu(x)} &= \log \frac{e^{-f(x)-g^\lambda(x)}/e^{-f(x)-g(x)}}{\int e^{-f(x)-g^\lambda(x)}dx/\int e^{-f(x)-g(x)}dx} \\ &= g(x) - g^\lambda(x) - \log \mathbb{E}_\mu\left[e^{g(x)-g^\lambda(x)}\right] \le 56L\sqrt{\lambda d}.\end{aligned}$$

It follows that

$$\left\|\mu - \mu^\lambda\right\|_{\mathsf{TV}} = \int \left|\mu(x) - \mu^\lambda(x)\right| dx \le \sup_x \left|\frac{\mu^\lambda(x)}{\mu(x)} - 1\right| \le CL\sqrt{\lambda d},$$

for some universal constant C. □

It leads to the following main theorem.

Theorem 3.2.1. *Let X_0 be initialization draw from ν_0, and $\nu_{K,h}$ be the distribution of X_K after K iterations of TDLMC Algorithm with step size h. If Conditions (3.1, 3.2, 3.3) hold with $L\sqrt{\lambda d} \le 1/2$ and $h(M+1/\lambda) \le 1/2$, then we have*

$$\left\|\nu_{K,h} - \mu\right\|_{\mathsf{TV}} \le C\left(e^{-mKh}\sqrt{D\left(\nu_0\middle\|\pi^\lambda\right)} + hM_\lambda^{3/2}\sqrt{\mathbb{E}_{\nu_0}\|X_0 - x^*\|^2} + hM_\lambda\sqrt{Kd} + L\sqrt{\lambda d}\right),$$

where $M_\lambda = M + 1/\lambda$ with M defined in Condition 3.1, x^* is the minimizer of U^λ, and C is some universal constant.

Notice that we are using fixed step size in Theorem 3.2.1 for simplicity, but it can be easily generalized to decreasing step size h_K, since the smoothing procedure doesn't depend on the step size. Theorem 3.2.1 involves the initial distribution. With a warm start, we can simplify the result in the following corollary.

Corollary 3.2.1. *Under all assumptions of Theorem 3.2.1, assume initialization X_0 is a warm start and follows $\mathcal{N}\left(x^*, \sigma^2 I_d\right)$ for some fixed variance, if we set*

$$\lambda \le C_1\frac{\varepsilon^2}{d}, \qquad h \le C_2\frac{\varepsilon^2\lambda^2}{d}, \qquad K = C_3\frac{1}{h}$$

then $\left\|\nu_{K,h}-\mu\right\|_{\mathsf{TV}} \leq \varepsilon$*, where* $\nu_{K,h}$ *is the distribution of* X_K *after K iterations of TDLMC Algorithm with step size h. Here,* C_1, C_2, C_3 *are constants depending on* m, M, L*.*

We pause here to do some literature review and compare the results. The best rate given so far in convex and smooth case using LMC is $h = \mathscr{O}\left(\varepsilon^2/d\right)$ and $K = \mathscr{O}\left(d/\varepsilon^2\right)$ in terms of total variation bound, ignoring the smoothness and convexity constant [18, 23]. However, without smoothness condition, we have the bound on mixing time (total number of iterations need) $K = \mathscr{O}\left(d^3/\varepsilon^6\right)$ by Corollary 3.2.1. Under the same condition, our result is better than the result in [15] whose mixing time bound is $\mathscr{O}\left(d^5/\varepsilon^7\right)$ with Gaussian smoothing, since their smoothness bound depends on dimension d after transformation. Moreau-Yosida proximal transformation proposed in [28] can also address non-smoothness, but the final rate is not precise as the constant C still involves the dimension d. Also notice that Moreau-Yosida only works for g that is lower bounded and requires g to be convex so that g^λ after transformation is smooth. In contrast, Tweedie's transformation works for unbounded g as well, and is smooth even when g is not convex.

Before proceeding with the non-convex case, we pause here to present a crucial role in the study of continuous process convergence rate, known as *Poincaré constant* (also referred to as *spectral gap*). The Poincaré constant of an arbitrary distribution π is given by

$$\gamma^* = \inf\left\{ \frac{\int \|\nabla f\|^2 d\pi}{\int f^2 d\pi} : \quad f \in C^1(\mathbb{R}^d) \cap L^2(\pi), \quad f \neq 0, \quad \int f d\pi = 0 \right\}.$$

By Theorem 4.2.5 in [4], we can well establish the convergence rate of continuous process via spectral gap. Hence, it leads to the following main theorem.

Theorem 3.2.2. *Let* X_0 *be initialization draw from* ν_0*, and* $\nu_{K,h}$ *be the distribution of* X_K *of TDLMC algorithm with step size h. Suppose Condition 3.1 and Condition 3.3 hold with* $L\sqrt{\lambda d} \leq 1/2$ *and* $h(M + 2/\lambda) \leq 1/2$*, then we have*

$$\left\|\nu_{K,h}-\mu\right\|_{\mathsf{TV}} \leq C\left(e^{-\gamma^* Kh/2}\sqrt{\chi^2\left(\nu_0\|\pi^\lambda\right)} + hM_\lambda^{3/2}\sqrt{\mathbb{E}_{\nu_0}\|X_0 - x^*\|^2} + hM_\lambda\sqrt{Kd} + L\sqrt{\lambda d}\right),$$

where $M_\lambda = M + 2/\lambda$ with M defined in Condition 3.1, x^* is the minimizer of U^λ, γ^* is the spectral gap of μ^λ, and C is some universal constants.

Worth noting that the smoothness of g^λ after Tweedie's transformation does not require the convexity of g by Lemma 3.4.2, and hence this is so far the only theorem establishing the convergence rate in non-convex and non-smooth case. Here, γ^* is the spectral gap of $\mu^\lambda \propto e^{-f-g^\lambda}$, which may depend on dimension d in some cases. The crucial part is to lower bound γ^*. Some related work has been elaborated in [65, 17, 84] and it will not be the focus of this Chapter.

3.3 Numerical Results

In order to show the validity of Tweedie's transformation as well as the fast convergence of TDLMC algorithm, we first present simulation result in one dimensional space, aiming to match the sampled distribution with the theoretical distribution [1]. Then, we move experiments to large dimensional cases.

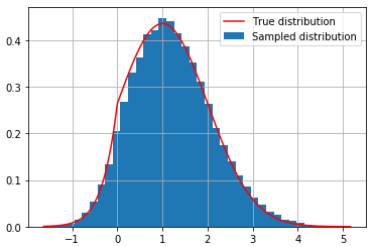
Gaussian with Laplace prior. Suppose we want to sample from $\mu \propto e^{-\frac{(x-a)^2}{2}-L|x|}$, which can be viewed as posterior distribution for Gaussian observation with Laplace prior. Here, we fix $a = 5$ and set initialization $X_0 = a$, and study how TDLMC algorithm performs with different value of L. We choose $\lambda \propto 1/L^2$ according to Lemma 3.2.2. It is better to set smaller step size h such that the approximation of discrete time Langevin process is more accurate, but it requires more iterations to converge. For simplicity, we set $h = \lambda$ and $K = 60,000$, and the results are shown below.

From Figure 3.2, we see that Tweedie's transformation and Algorithm 2 are valid and work well to sample from non-convex distribution, as all histograms well fit red curves.

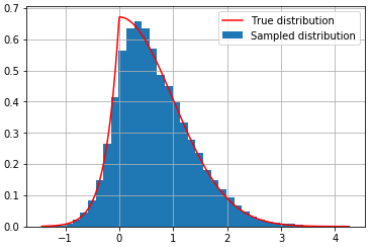
Gaussian with horseshoe prior. We then sample from $\mu \propto e^{-\frac{(x-a)^2}{2}} p(x)$, where $p(x)$ takes the form as horseshoe prior [13]. Horseshoe prior can be viewed as normal prior with random variance of the form

$$x|\mu \sim \mathcal{N}(0,\mu), \quad \mu \sim \pi(u) \propto u^{-1/2}(1+u/\tau^2)^{-1}. \tag{4}$$

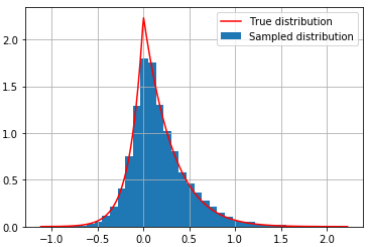
1. Code is available in `https://github.com/zhuobumeng/Tweedie_sampling/`



(a) $L = 1$



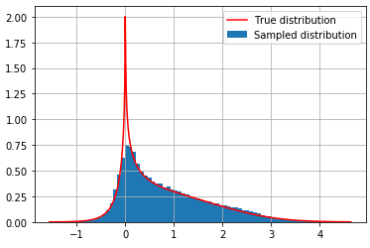
(b) $L = 2$



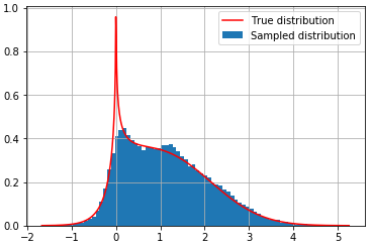
(c) $L = 5$

Figure 3.2: Tweedie transformation applied onto Laplace prior. Blue areas denote the histogram of sampled data using Algorithm 2, and red curves represent the true distribution. Here, we choose $\lambda = 0.05/L^2$.

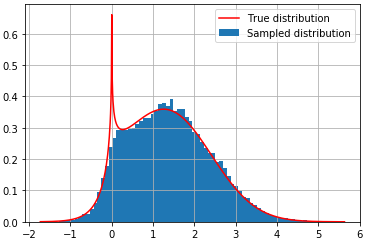
We plot the potential function of horseshoe prior in Figure 3.1 by taking $\tau = 1$. As shown in Figure 3.1 plot (d), the original function is unbounded, not smooth, and not convex, such that common sampling methods, including Moreau-Yosida proximal sampling, fail. To show the advantages of our proposed algorithm, we do experiment in one dimension, and match the sampled data to its true distribution. We again set $a = 2$, $X_0 = 2$, $K = 60,000$, and only change the value of τ. Here, τ controls the variance of normal prior, and as τ gets larger, the prior is less informative and the posterior is more toward the observation data ($a = 2$ in our case). Since we require step size to be smaller when τ is smaller, we choose $\lambda \propto \sqrt{\tau}$ heuristically, and set the step size $h = \lambda$. The results are shown below:



(a) $\tau = 0.5$



(b) $\tau = 1$



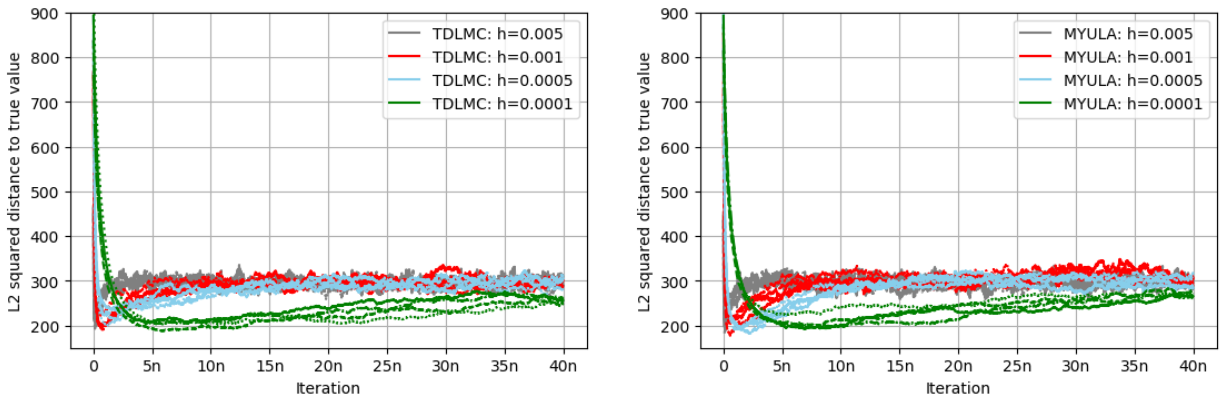
(c) $\tau = 2$

Figure 3.3: Tweedie transformation applied onto horseshoe prior. Blue areas denote the histogram of sampled data using Algorithm 2, and red curves represent the true distribution. Here, we choose $\lambda = h = 0.01\sqrt{\tau}$.

Large dimension Bayesian sparse linear regression. In this part, we study how fast the algo-

rithm converges to stationary distribution. We focus on sampling from $\mu \sim e^{-\frac{\|y-Ax\|^2}{2}} p(x)$ where $x, y \in \mathbb{R}^n$. Here, $A \in \mathbb{R}^{n \times n}$ is design matrix, and $p(x)$ is sparse prior. It can be viewed as Bayesian sparse linear regression problem, and we want to sample from the posterior. For the likelihood model, We set sample size $n = 100$, and use independent design matrix where $A_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. We set sparsity $s = 10$, i.e., the first 10 elements of underlying true x^* is 0, and all remaining elements are set to be $10\sqrt{2\log n/n}$, larger than the universal threshold of $\sqrt{2\log n/n}$. Observation y is generated through likelihood model.

We begin with experiment using Laplace prior, given by $p(x_i) \propto e^{-L|x_i|}$ for each $i \in \{1, \ldots, n\}$. We choose $L = \sqrt{2n\log n}$, and try to sample from the corresponding posterior. This is non-smooth distribution sampling problem, and we present below the results using our algorithm TDLMC via Tweedie's transformation and the MYULA algorithm in [29] via Moreau-Yosida envolope.



(a) TDLMC (b) MYULA

Figure 3.4: Comparison of TDLMC and MYULA algorithms. In each plot, each line represents one Markov chain. Here, x-axis represents the number of iterations in terms of the sample size, and y-axis represents the L_2 squared distance from the current point to the true value x^*, given by $\|X_k - x^*\|^2$. In each plot, different colors represent different step size h and smoothing level λ, and we set $h = \lambda$ for simplicity. In each color, there are 4 lines, representing 4 different replicates of experiments. The initializations are chosen to be standard Gaussian.

From Figure 3.4, we can see that both two algorithms converge to stationary distribution very fast. We can also see that for both two algorithms, when step size is smaller, it takes longer time for Markov chains to converge. In this case, the performances of two algorithms are very similar, but

the application of MYULA is limited.

We then shift gears to sparse Bayesian regression with horseshoe prior, given by (4), where the density is unbounded at 0. MYULA is using posterior mode and it fails in this case. It is worth noting that so far, to the best of our knowledge, TDLMC is the only sampling algorithms that can sample from the posterior distribution with horseshoe prior when design matrix is not orthogonal. The prior $p(x_i)$ for each $i \in \{1,\dots,n\}$ is defined by (4). The experiment design is the same as regression with Laplace prior. The regularization parameter τ in horseshoe prior is set to be $\frac{p}{n^{3/2}}$ according to [78]. The result is shown below.

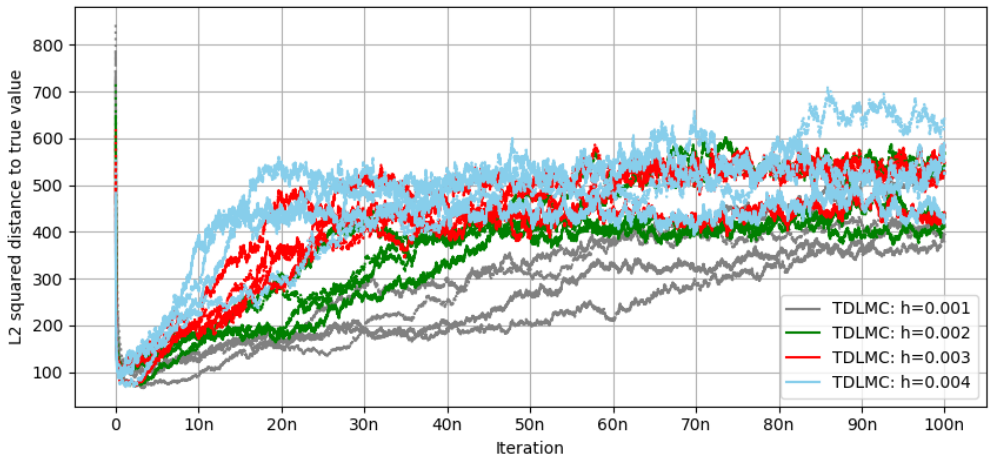


Figure 3.5: TDLMC applied on sparse Bayesian regression with horseshoe prior. Each line represents one Markov chain. Here, x-axis represents the number of iterations in terms of the sample size, and y-axis represents the L_2 squared distance from the current point to the true value x^*, given by $\|x_k - x^*\|^2$. In each plot, different colors represent different step size h and smoothing level λ, and we set $h = \lambda$ for simplicity. In each color, there are 4 lines, representing 4 different replicates of experiments. The initializations are chosen to be standard Gaussian.

As seen in Figure 3.5, TDLMC works well for sampling from non-convex, non-smooth and even unbounded distribution. When the step size is slightly larger, Markov chains reaches plateau faster.

3.4 Proofs

In this section, we present the technical proofs for lemmas and theorems. Some other technical proofs can be found in Appendix.

3.4.1 Proof of Lemma 3.2.1

Proof of Lemma 3.2.1. The first claim holds by Proposition 3.5 in [72]. For other two claims, we slightly change notation here for general proof. Suppose $y \sim p$ defined in $\mathbb{R}^d$, $z \sim \mathcal{N}(0, \lambda I_d)$, $y \perp\!\!\!\perp z$ and $x = y + z$. Denote $m(x)$ (resp. $p(y)$, $q(z)$) as $\exp(-\phi_m(x))$ (resp. $\exp(-\phi_p(y))$, $\exp(-\phi_q(z))$). It follows that

$$m(x) = p * q(x) = \int p(x-z)q(z)dz = \int p(y)q(x-y)dy.$$

If the density function of y is log-concave, then the density function of x is log-concave. The strongly log-concave property is also preserved [72]. And we have

$$\begin{aligned}
\nabla\phi_m(x) &= -\nabla \log m(x) = -\frac{\nabla m(x)}{m(x)} = -\frac{\int p(y)\nabla q(x-y)dy}{m(x)} \\
&= \frac{\int \nabla\phi_q(x-y)p(y)q(x-y)dy}{m(x)} = \mathbb{E}\left[\nabla\phi_q(z) \mid y+z=x\right] \\
&= \mathbb{E}\left[\nabla\phi_p(y) \mid y+z=x\right].
\end{aligned}$$

The last equation holds when ϕ_p is differentiable. It directly gives the Tweedie's formula that

$$\nabla\phi_m(x) = \mathbb{E}\left[\nabla\phi_q(z)|x\right] = \mathbb{E}\left[\frac{x-y}{\lambda} \,\middle|\, x\right] = \frac{x - \mathbb{E}[y|x]}{\lambda}.$$

As shown above,

$$m(x) \propto \int \exp\left(-\phi_p(y) - \frac{\|y-x\|^2}{2\lambda}\right) dy,$$

is an average version of the density p after Gaussian kernel convolution. When $\lambda \to 0$, $m(x)$ pointwise converges to $p(x)$. Furthermore, $\phi_m(x)$ is a smooth and convex approximation of $\phi_p(x)$. It

follows that

$$
\begin{aligned}
&\nabla^2\phi_m(x)\\
=&\nabla\mathbb{E}\left[\nabla\phi_q(z)|x\right]=\nabla\left(\int q(x-y)\cdot\nabla\phi_q(x-y)p(y)dy\cdot\frac{1}{m(x)}\right)\\
=&\int\nabla q(z)\cdot\nabla\phi_q(x-y)^T p(y)dy/m(x)+\int q(z)\cdot\nabla^2\phi_q(z)p(y)dy/m(x)-\frac{\int q(z)\nabla\phi_q(z)p(y)dy}{m(x)}\cdot\frac{\nabla m(x)}{m(x)}\\
=&\mathbb{E}\left[\nabla^2\phi_q(z)\;\middle|\;x\right]-\mathbb{E}\left[\nabla\phi_q(z)\cdot\nabla\phi_q(z)^T\;\middle|\;x\right]+\mathbb{E}\left[\nabla\phi_q(z)\mid x\right]\cdot\mathbb{E}\left[\nabla\phi_q(z)\mid x\right]^T\\
=&\mathbb{E}\left[\nabla^2\phi_q(z)\;\middle|\;x\right]-\mathrm{Var}\left(\nabla\phi_q(z)\mid x\right)\\
=&\mathbb{E}\left[\nabla^2\phi_p(y)\;\middle|\;x\right]-\mathrm{Var}\left(\nabla\phi_p(y)\mid x\right)
\end{aligned}
$$

The last equality holds if $\nabla^2\phi_p(y)$ exists. □

Lemma 3.4.1. *Suppose $Z\sim\mathcal{N}(0,I_d)$. If $L\sqrt{\lambda}\le 1/2$, we have* $\log\mathbb{E}\left[e^{L\sqrt{\lambda}\|z\|}\right]\le 28L\sqrt{\lambda d}$.

Proof of Lemma 3.4.1. We have

$$
\mathbb{E}\left[e^{L\sqrt{\lambda}\|z\|}\right]=\int_0^\infty\mathbb{P}\left\{e^{L\sqrt{\lambda}\|z\|}\ge t\right\}dt\le a+\int_a^\infty\mathbb{P}\left\{e^{L\sqrt{\lambda}\|z\|}\ge t\right\}dt,
$$

where $a=e^{2L\sqrt{\lambda d}}$. It follows that

$$
\begin{aligned}
&\int_a^\infty\mathbb{P}\left\{e^{L\sqrt{\lambda}\|z\|}\ge t\right\}dt=\int_a^\infty\mathbb{P}\left\{\|z\|^2\ge\left(\frac{\log(t)}{L\sqrt{\lambda}}\right)^2\right\}dt\\
\le&\int_a^\infty\exp\left(-\frac{\left(\frac{\log(t)}{L\sqrt{\lambda}}\right)^2-d}{8}\right)dt\le\int_a^\infty\exp\left(-\frac{\left(\frac{\log(t)}{L\sqrt{\lambda}}\right)^2}{16}\right)dt\\
=&\int_{\log(a)}^\infty\exp\left(-\frac{y^2}{16L^2\lambda}+y\right)dy\\
=&\int_{\log(a)}^\infty\exp\left(-\frac{(y-8L^2\lambda)^2}{16L^2\lambda}\right)\frac{1}{4L\sqrt{\lambda}}dy\cdot\exp\left(4L^2\lambda\right)4L\sqrt{\lambda}\\
\le&8\pi L\sqrt{\lambda}\exp(4L^2\lambda)
\end{aligned}
$$

□

3.4.2 Proof of Theorem 3.2.1 and Theorem 3.2.2

Proof of Theorem 3.2.1 and Theorem 3.2.2. We begin with the proof of Theorem 3.2.1. Suppose we have continuous time diffusion process defined in (1) and discretized LMC defined in (2), with potential function U^λ, given by $U^\lambda = f + g^\lambda$. Suppose two processes start at the same initial distribution ν_0. For the sake of the presentation, we use μ_t to denote the distribution of continuous time process at time t, and use $\nu_{k,h}$ to denote the distribution of LMC after k steps with step size h.

Continuous time convergence. It is easy to see that U^λ is m-strongly convex by Condition 3.1, Condition 3.2, and Lemma 3.2.1. By Bakry-Émery theorem ([80], Th. 21.2 and Remark 21.4), we have that the distribution $\mu^\lambda(x) \propto \exp(-U^\lambda(x))$ satisfies a logarithmic Sobolev inequality with constant m, i.e., for any π that is absolutely continuous w.r.t. μ^λ,

$$D\left(\pi \middle\| \mu^\lambda\right) \leq \frac{1}{2m} \int \frac{|\nabla \rho|^2}{\rho} \mu^\lambda dx, \quad \text{where} \quad \rho = \frac{\pi}{\mu^\lambda}.$$

It follows directly Theorem 5.2.1 in [4], that

$$D\left(\mu_t \middle\| \mu^\lambda\right) \leq \exp(-2mt)\, D\left(\nu_0 \middle\| \mu^\lambda\right).$$

Hence, by Pinsker inequality, we have $\left\|\mu_t - \mu^\lambda\right\|_{\mathsf{TV}} \leq \sqrt{\frac{1}{2} D\left(\mu_t \middle\| \mu^\lambda\right)} \leq \exp(-mt) \sqrt{D\left(\nu_0 \middle\| \mu^\lambda\right)}$.

Discrete time process approximation. By Lemma 3.1.1, at time $t = kh$, we have

$$\left\|\nu_{k,h} - \mu_{kh}\right\|_{\mathsf{TV}} \leq \sqrt{\frac{1}{2} D\left(\nu_{k,h} \middle\| \mu_{kh}\right)} \leq \sqrt{\frac{{M_\lambda}^3 h^2}{6} \left(\mathbb{E}_{\nu_0}\left[\|X_0 - x^*\|_2^2\right] + 2khd\right) + \frac{dk{M_\lambda}^2 h^2}{4}},$$

where M_λ is the smoothness constant of U^λ. By Lemma 3.2.1 and the convexity of g, we know $\|\nabla^2 g^\lambda(x)\| \leq \frac{1}{\lambda}$. Hence, we can simply take $M_\lambda = M + 1/\lambda$, where M is the smoothness constant of f.

Tweedie's transformation. By Lemma 3.2.2, we have that

$$\left\|\mu-\mu^{\lambda}\right\|_{\mathsf{TV}} \leq CL\sqrt{\lambda d},$$

for some constant C.

Finally, due to the triangle inequality of total variation, after combining all three bounds, it directly gives us the result of Theorem 3.2.1.

In order to prove Theorem 3.2.2, without convexity condition, we need to replace the continuous time convergence rate and bound the smoothness of $g^{\lambda}(x)$. By Theorem 4.2.5 in [4], we directly have

$$\left\|\mu_t-\mu^{\lambda}\right\|_{\mathsf{TV}} \leq e^{-\gamma^* t}\chi^2\left(\nu_0\middle\|\mu^{\lambda}\right),$$

where γ^* is the spectral gap of μ^{λ}, and needs to be analyzed case by case. By Lemma 3.4.2, we have $\left\|\nabla^2 g^{\lambda}(x)\right\| \leq \frac{2}{\lambda}$. Hence, the result directly follows. □

Proof of Corollary 3.2.1. It follows that

$$\begin{aligned} D\left(\nu_0\middle\|\mu^{\lambda}\right) &= \int \nu_0(x)\log\frac{\nu_0(x)}{\mu^{\lambda}(x)}dx \\ &= \int \nu_0(x)\log\nu_0(x)dx + \int \nu_0(x)U^{\lambda}(x)dx + \log\int \exp(-U^{\lambda}(x))dx, \end{aligned}$$

where

$$\int \nu_0(x)\log\nu_0(x)dx = \frac{d}{2}\log\frac{1}{2\pi\sigma^2} - \frac{d}{2},$$

and

$$\int \nu_0(x)U^{\lambda}(x)dx \leq \int \nu_0(x)\left(U^{\lambda}(x^*) + \frac{M_{\lambda}}{2}\|x-x^*\|_2^2\right)dx = U^{\lambda}(x^*) + \frac{dM_{\lambda}}{2}\sigma^2,$$

and

$$\int \exp\left(-U^{\lambda}(x)\right) dx \leq \exp\left(-U^{\lambda}(x^*)\right) \int \exp\left(-\frac{m}{2}\|x-x^*\|^2\right) dx = \exp(-U^{\lambda}(x^*)) \left(\frac{2\pi}{m}\right)^{d/2}.$$

Combining all inequalities and ignoring constants expect for step size h, smoothing level λ, and dimension d, it directly gives the result. □

Lemma 3.4.2. *Suppose g satisfies Condition 3.1 and Condition 3.3. After we apply Tweedie's transformation to get $g^{\lambda}(x)$, we have $\left\|\nabla^2 g^{\lambda}(x)\right\| \leq 2/\lambda$.*

Proof. By Condition 3.1 and Condition 3.3, for any $x \in \mathbb{R}^d$, we have $g(x) = \sum_{i=1}^{d} g_i(x_i)$, and each g_i is L-Lipschitz. In addition, we have that g^{λ} is also separable by (3), and hence $\nabla^2 g^{\lambda}$ is a diagonal matrix. Combined with Lemma 3.2.1, we have

$$\partial_{x_i}^2 g^{\lambda}(x) = \frac{1}{\lambda} - \frac{\mathrm{Var}(z_i|x_i)}{\lambda},$$

for $i = 1, \ldots, d$, where $p(z_i|x_i) \propto \exp\left(-g(x_i - z_i) - \frac{z_i^2}{2\lambda}\right)$. By Lemma 3.4.3, we have that when $L\sqrt{\lambda d} \leq 1/2$, we have $\left|\partial_{x_i}^2 g^{\lambda}(x)\right| \leq 2/\lambda$. □

Lemma 3.4.3. *Suppose $g : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz, and $x, y, z \in \mathbb{R}$ satisfy that $y \sim g(x)$, $z \sim \mathcal{N}(0,1)$, $y \perp\!\!\!\perp z$, and $x = y + \sqrt{\lambda} z$. Then, we have*

$$\mathrm{Var}(z|x) \leq \frac{\sqrt{6}}{2} e^{L^2\lambda}.$$

When $L\sqrt{\lambda} \leq 1/2$, $\mathrm{Var}(z|x) \leq 2$.

Proof of Lemma 3.4.3. We have that

$$\mathrm{Var}(z|x) \leq \mathbb{E}\left[z^2 \,\middle|\, x\right] = \frac{\int z^2 e^{-g(x-\sqrt{\lambda}z)} \phi(z) dz}{\int e^{-g(y)} dy} \leq \frac{\int z^2 e^{g(x)-g(x-\sqrt{\lambda}z)} \phi(z) dz}{\int e^{g(x)-g(y)} dy} \leq \frac{\int z^2 e^{L\sqrt{\lambda}|z|} \phi(z) dz}{\int e^{-|x-y|} dy},$$

where $\int e^{-|x-y|}dy = 2$, and

$$\mathbb{E}\left[z^2 e^{L\sqrt{\lambda}|z|}\right] \leq \sqrt{\mathbb{E}\left[z^4\right]\mathbb{E}\left[e^{2L\sqrt{\lambda}|z|}\right]} \leq \sqrt{3\left(\mathbb{E}\left[e^{2L\sqrt{\lambda}z}\right]\right) + \mathbb{E}\left[e^{-2L\sqrt{\lambda}z}\right]} \leq \sqrt{6}e^{L^2\lambda}.$$

□

CHAPTER 4

APPROXIMATE BAYESIAN COMPUTATION WITH BERNSTEIN-VON MISES PROPERTY

Bayesian approach for inference has become one of the central interests in statistical inference, due to its advantage of expressing uncertainty in probabilities rather than using criteria integrating out the whole sample space. The full inference using Bayesian approach is based on the posterior distribution and is widely studied. Though it is convenient, there are still some limitations to the full Bayesian inference, especially in scenarios where the likelihood does not have closed-form expression or cannot be numerically evaluated, while it is accessible to simulate synthetic data from generation model for given parameter of interest. In addition, full inference based on the posterior distribution is not attractive considering model misspecification, when the likelihood is no longer a good criterion for parameter estimation. In the past century, a surge of research works have improved frequentist approaches to data analysis and model fitting, but it still remains unclear, to the best of our knowledge, how to naturally combine the frequentist approach of parameter estimation with prior information, when posterior inference is not preferable.

In recent years, approximate Bayesian computation (ABC) draws great attentions as one of popular Bayesian techniques [7, 22], when other likelihood-based inferences are limited by the difficulty of computing the likelihood. ABC provides approximate inference in generative models with tractable algorithms, and has proven useful in various applications [76, 49]. It is a simulation-based approach that first samples parameter according to prior distribution, and then simulates synthetic data based on generation process given the sampled parameter. If the synthetic data and true data are close enough, then we accept the sampled parameter. The *closeness* is measured according to the distance between summary statistics compared with user-defined threshold level. All ABC samples form an approximate distribution, and is close to posterior distribution, if the selected summary statistic is sufficient and the threshold level is small enough [5]. The choice of summary statistics is crucial and most works have been focused on choosing sufficient summary

statistics in order to increase the information available to the ABC procedure, with great hope of recovering the posterior distribution [9, 63, 57]. However, in most cases, it is not preferable to choose sufficient statistics since we may only focus on one part of information among all data.

In this work, we consider a special case where a frequentist estimation $\widehat{\theta}_Y$ is given, obtained by optimizing some loss function $L_n(\theta, X)$ with observation $Y = (Y_i)_{i=1}^n$ with sample size n, and we aim to answer the following three questions:

- How to naturally incorporate prior information of θ with the loss function?
- How to do Bayesian inference?
- Is there any good property of Bayesian distribution when sample size is large enough?

The proposed Bayesian distribution in this work is constructed by the prior distribution and the sampling distribution of $\widehat{\theta}_Y$. To the best of our knowledge, the Bayesian distribution proposed in this work is the only distribution that not only combines information from loss function and prior distribution, but also possesses Bernstein-von Mises type of property, in the sense that the asymptotic shape of Bayesian distribution is the same as that of the sampling distribution of $\widehat{\theta}_Y$. In addition, when $\widehat{\theta}_Y$ is taken to be maximum likelihood estimate, then the Bayesian distribution asymptotically converges to the full posterior distribution. The sampling distribution is not feasible in most cases, and thus we adopt ABC algorithm to do Bayesian inference. The approximate Bayesian distribution obtained by ABC algorithm is also proved to share Bernstein-von Mises type of property under certain conditions. Due to computational complexity, we further provide an alternative algorithm pseudo ABC that is more efficient.

4.1 Problem Formulation and Background Introduction

In this section, we mathematically formulate our problem of constructing Bayesian distribution for generative model. Then, some background of approximate Bayesian computation is presented, followed by some related work in constructing Bayesian distributions.

4.1.1 Problem Formulation and Generative Models

Throughout this work, we consider distribution class $\{P_\theta : \theta \in \Theta\}$ indexed by $\theta \in \Theta \subset \mathbb{R}^d$. We denote the data by $Y = (Y_1, \ldots, Y_n)$ drawn from $P^n_{\theta_0}$, where θ_0 denotes the true parameter as an interior point in Θ, and n is the sample size. Each observation y_i can be of arbitrary dimension. We consider the asymptotics as $n \to \infty$ and d is fixed. Suppose the loss function is given, denoted by $L_n(\theta, X)$, and the estimate $\widehat{\theta}_Y$ is given by

$$\widehat{\theta}_Y = \arg\min_\theta L_n(\theta, Y). \tag{1}$$

In addition, we have prior information of θ, denoted by distribution $\pi(\theta)$. Our goal is to *naturally* incorporate prior information with the assigned loss function.

Very often, for complex loss functions, such as Tukey's depth function [3] and generative adversarial net, it is difficult to do inference and study the distribution of $\widehat{\theta}_Y$. In this paper, we propose a new type of Bayesian distribution and adopt approximate Bayesian computation to address these issues. We consider inference for purely generative models, i.e., it is feasible to sample from P^n_θ for all $\theta \in \Theta$.

4.1.2 Approximate Bayesian Computation

Approximate Bayesian computation (ABC) is a likelihood-free method for implementing Bayesian inference. It was first proposed in [64], and then improved and popularized in population genetics[7], systems biology [76], finance [62], and statistics [33, 9, 51]. Suppose we have data $Y = (Y_1, \ldots, Y_n) \sim P^n_{\theta_0}$, where $\theta_0 \in \Theta$ is the true parameter. Let π denote the prior distribution on θ. Then the basic ABC procedure is presented below:

(1) Draw a parameter $\theta \sim \pi(\theta)$;

(2) Sample *fake* data $X = (X_1, \ldots, X_n) \sim P^n_\theta$;

(3) Accept θ if $d(w(X), w(Y)) \leq h$;

(4) Repeat procedure and collect all remaining θ;

Here, $d(\cdot,\cdot)$ can be any user-defined distance metric, $w(X), w(Y)$ are summary statistics used to distinguish true data from fake data, and ε is the threshold of ABC. According to this procedure, the final accepted samples are drawn from ABC posterior distribution, given by

$$\pi_h^{ABC}(\theta) \propto \pi(\theta) \int \mathbb{I}\{d(w(X), w(Y)) \leq h\} \, dP_\theta^n(X) \propto \pi(\theta) \mathbb{P}_\theta \{d(w(X), w(Y)) \leq h\},$$

where with a slight abuse of notation, we use $\mathbb{P}_\theta$ to take probability over all variables drawn from P_θ, and hence Y is treated as fixed. It is easy to check that if $w(X), w(Y)$ are sufficient statistics, when $h \to 0$ as $n \to \infty$, and π_h^{ABC} converges to the true posterior distribution. In most cases where the likelihood is not tractable, ABC is adopted to recover the posterior distribution by carefully choosing summary statistics, which is not the focus of this work.

Notice that the above procedure is referred to as *rejection ABC*, and the more general version will be discussed later.

4.1.3 Contribution and Related Works

Prior to our work, there are two types of work of constructing Bayesian distribution in order to incorporate the prior information. One type of work focuses on analytical form, given by $q(\theta) \propto \exp(-\lambda L_n(\theta, Y))\pi(\theta)$ [46, 47, 19, 20], where the loss function plays the role as a scaled negative log-likelihood, and λ is an extra parameter. It can be also viewed as a optimization problem with negative log prior as regularization. The other type of work focuses on ABC procedure replacing distance metric by loss function, i.e., accept sampled parameter θ when the loss on fake data is small enough [41, 57, 9]. However, these two types of work both depend heavily on the scale of loss function, which should be irrelevant considering that the frequentist estimate does not change with the scale of loss function.

In this work, we propose a novel way to add prior distribution on top of frequentist estimation, and propose a feasible algorithm to do inference, along with rigorous proof of nice asymptotic

properties.

4.2 Approximate Bayesian Distribution

In this section, we propose theoretical Bayesian distribution that combines the loss function and the prior distribution. Due to lack of analytical form, we adopt ABC algorithm to sample from the Bayesian distribution, and the output samples are drawn from an approximate Bayesian distribution. Then, we present rigorous theoretical proofs to show both two distributions share Bayesian contraction and Bernstein-von Mises type of property under certain conditions.

4.2.1 Bayesian Formula

We begin with the theoretical Bayesian distribution proposed in this work. Recall that the frequentist estimate $\widehat{\theta}_Y$ is defined in (1) w.r.t. data $Y \sim P_{\theta_0}^n$, and let $\widehat{\theta}_X$ be the same statistics w.r.t. fake data X where $X \sim P_\theta^n$ for some θ, i.e., $\widehat{\theta}_X = \arg\min_\theta L_n(\theta, X)$. We define a new Bayesian distribution, taking the form of

$$q(\theta) = \frac{\pi(\theta) p_{\widehat{\theta}_X|\theta}(\widehat{\theta}_Y)}{\int_\Theta \pi(\theta) p_{\widehat{\theta}_X|\theta}(\widehat{\theta}_Y) d\theta}, \tag{2}$$

where $p_{\widehat{\theta}_X|\theta}(\widehat{\theta}_Y)$ is the density of $\widehat{\theta}_X$ given θ taking value of $\widehat{\theta}_Y$. We pause here to better explain the distribution through a toy example. Suppose $P_\theta = \mathcal{N}(\theta, I_d)$. Let $Y_1, \ldots, Y_n \sim \mathcal{N}(\theta_0, I_d)$ for some true θ_0, and the estimate obtained from some loss function takes the form of $\widehat{\theta}_Y = \bar{Y}$. Then, the proposed Bayesian distribution in this case is given by

$$q(\theta) \propto \pi(\theta) p_{\widehat{\theta}_X|\theta}(\widehat{\theta}_Y) \propto \pi(\theta) p_{\widehat{\theta}_X|\theta}(\widehat{\theta}_Y) \propto \pi(\theta) \exp\left(-\frac{n\|\bar{Y} - \theta\|^2}{2}\right),$$

which boils down to the standard posterior. Worth noting that the construction of the proposed Bayesian distribution works for any loss function that gives consistent estimate. If we take negative log-likelihood as our loss function, then the Bayesian distribution converges to the full posterior distribution asymptotically as $n \to \infty$, since maximum likelihood estimate is asymptotic sufficient.

4.2.2 Bayesian Computation

The Bayesian distribution defined in (2) is intractable as we do not know the density of $\widehat{\theta}_X$ for each given θ. Considering it is easy to generate data from P_θ for each θ, then it is feasible to do inference via ABC. Hence, we provide the first main algorithm via ABC.

Algorithm 3: Basic approximate Bayesian computation

Input: Loss function L_n,

Frequentist estimate $\widehat{\theta}_Y = \arg\min_\theta L_n(\theta, Y)$,

Proposal distribution $S(\theta)$,

Choice of kernel function: $\mathscr{K} : \mathbb{R}^d \to \mathbb{R}$

Large constant $K > \mathscr{K}(0) \max_\theta \frac{\pi(\theta)}{S(\theta)}$,

Bandwidth (or threshold) h,

Number of samples N.

for each $i \in \{1, 2, \ldots, N\}$ **do**

Generate $\theta^{(i)} \sim S(\theta)$;

Sample $X = (X_1, \ldots, X_n) \sim P^n_{\theta^{(i)}}$;

Solve $\widehat{\theta}_X = \arg\min_\theta L_n(\theta, X)$;

Accept $\theta^{(i)}$ with probability

$$\mathscr{K}\left(\frac{\widehat{\theta}_X - \widehat{\theta}_Y}{h}\right)\frac{\pi(\theta)}{K \cdot S(\theta)}$$

Output: A set of sampled parameters $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$

Algorithm 3 is rooted from the ABC rejection algorithm but is more general. In Algorithm 3, we use $\mathscr{K} : \mathbb{R}^d \to \mathbb{R}$ to denote any multidimensional kernel function that must satisfy the following properties:

- $\mathscr{K}(u) \geq 0$ is a symmetric density function;
- $\int_u \mathscr{K}(u)du = 1$, $\int_u u\mathscr{K}(u)du = 0$, and $\int_u uu^T \mathscr{K}(u)du \succ 0$.

We use $\mathscr{K}_h$ to denote the kernel function with bandwidth h, given by $\mathscr{K}_h(u) = h^{-d}\mathscr{K}(u/h)$. In particular, if we take $\mathscr{K}(u) \propto \mathbb{I}\{\|u\| \leq 1\}$ in Algorithm 3, then it gives back the ABC rejection algorithm. Another big difference is that, in ABC procedure, we need to seek for summary statistics $w(X)$ in order to recover the posterior distribution, while here we fix the summary statistic to be the solution of the loss function w.r.t. the fake data and true data.

The samples obtained from Algorithm 3 no longer follow the Bayesian distribution defined in (2), but instead follow an approximate Bayesian distribution, given by

$$q_h(\theta) = \frac{\pi(\theta)\int_X \mathscr{K}_h\left(\widehat{\theta}_X - \widehat{\theta}_Y\right) dP_\theta^n(X)}{\int_\theta \pi(\theta)\int_X \mathscr{K}_h\left(\widehat{\theta}_X - \widehat{\theta}_Y\right) dP_\theta^n(X)d\theta}, \tag{3}$$

where $\mathscr{K}$ is the kernel function defined above with bandwidth h, P_θ^n is the product measure, and $\widehat{\theta}_X$, $\widehat{\theta}_Y$ are solutions of loss function w.r.t. true data $Y \sim P_{\theta_0}^n$ and fake data $X \sim P_\theta^n$. In particular, when we take $\mathscr{K}(u) \propto \mathbb{I}\{\|u\| \leq 1\}$, the approximate Bayesian distribution is simplified to

$$q_h(\theta) \propto \pi(\theta)\mathbb{P}_\theta\left\{\|\widehat{\theta}_X - \widehat{\theta}_Y\| \leq h\right\}.$$

The main idea of Algorithm 3 is to sample parameters according to the prior distribution, and then keep those parameters whose fake data cannot be well distinguished from the true data by the loss function. Hence, the approximate Bayesian distribution defined in (3) is a natural way to combine prior and the loss function, which is also tractable. Notice that both Q and Q_h are constructed given observation Y.

4.2.3 Main Results of Asymptotic Properties

Before diving into main theorems, we first present some assumptions. Recall that we study the asymptotic property with $n \to \infty$ and d fixed. Throughout this section, let θ_0 be an inner point of Θ. Let $\widehat{\theta}_X = \arg\min_\theta L_n(\theta, X)$ for $X \sim P_\theta^n$. We use $B(a,\varepsilon)$ to denote the ε-neighbor of point a, taking the form of $B(a,\varepsilon) = \{\theta : \|\theta - a\| \leq \varepsilon\}$.

Assumption 4.2.1. *The estimate $\widehat{\theta}_X$ is uniformly consistent for all $\theta \in \Theta$, that is, for any $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_\theta \left\{ \left\| \widehat{\theta}_X - \theta \right\| \geq \delta \right\} = 0.$$

Assumption 4.2.2. *There exists an open set neighborhood U of θ_0, such that for any $\varepsilon > 0$, there exists some large constant M, such that*

$$\lim_{n\to\infty} \sup_{\theta\in U} \mathbb{P}_\theta \left\{ \sqrt{n} \left\| \widehat{\theta}_X - \theta \right\| \geq M \right\} \leq \varepsilon.$$

Assumption 4.2.1 implies that the loss function is *good* in the sense that it gives the uniformly consistent estimate for all $\theta \in \Theta$. Assumption 4.2.2 implies that for any θ in a neighbor of θ_0, $\sqrt{n}\|\widehat{\theta}_X - \theta\| = O_P(1)$. Let $F_n(\theta,\cdot)$ denote the distribution of $\sqrt{n}\left(\widehat{\theta}_X - \theta\right)$ with density $f_n(\theta,\cdot)$, where $X \sim P_\theta^n$, and use $F(\theta,\cdot)$ to denote its limit distribution with density $f(\theta,\cdot)$. We have the following stronger assumption to make sure $f_n(\theta,\cdot)$ is close to $f(\theta,\cdot)$.

Assumption 4.2.2'. *There exists an open set neighborhood U of θ_0, such that $f_n(\theta,t) \to f(\theta,t)$ for all t and all $\theta \in U$, and for any fixed large constant M,*

$$\lim_{n\to\infty} \sup_{\theta\in U} \sup_{\|t\|\leq M} \left| \frac{f_n(\theta,t)}{f(\theta,t)} - 1 \right| = 0.$$

Worth noting that Assumption 4.2.2' implies Assumption 4.2.2. In addition, we have the following continuously differentiable assumptions on limit density function.

Assumption 4.2.3. *There exists an open set neighborhood U of θ_0, such that for any fixed constant M, there are corresponding constants L_1, L_2 satisfying that*

$$\sup_{\theta\in U} \sup_{\|t\|\leq M} \|\nabla_\theta \log f(\theta,t)\| \leq L_1, \qquad \sup_{\|t\|\leq M} \|\nabla_t \log f(\theta_0,t)\| \leq L_2, \tag{4}$$

where L_1, L_2 depend on constant M. That is, $f_n(\theta,t)$ is uniformly continuously differentiable w.r.t. θ and t for any $\theta \in U$.

Assumption 4.2.4. *Suppose* $\sup_t f(\theta_0,t) < L$*, i.e., the limit density function is upper bounded.*

Assumption 4.2.3 is a general condition for most density functions, and is easily satisfied if $\nabla_\theta \log f(\theta,t)$ and $\nabla_t \log f(\theta,t)$ are continuous within the constraint set. The final assumption is for prior function as shown below.

Assumption 4.2.5. *There exists some small constant* δ_π *such that*

$$\sup_{\theta \in B(\theta_0,\delta_\pi)} \|\nabla_\theta \log \pi(\theta)\| \leq L_\pi, \tag{5}$$

for some constant L_π*, i.e.,* $\pi(\theta)$ *is continuously differentiable.*

We then proceed to present asymptotic properties of the Bayesian distribution and the approximate Bayesian distribution.

Theorem 4.2.1 (Contraction result)**.** *We use* Q *to denote the Bayesian distribution with density function defined in* (2)*. If Assumptions (4.2.1, 4.2.2', 4.2.3, 4.2.4, 4.2.5) hold, then for any small* $\varepsilon > 0$*, we can find corresponding large constant* M'*, such that*

$$\mathbb{E}_{\theta_0} Q\left(\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \geq M'\right) \leq \varepsilon,$$

for large n, i.e., for any positive sequence $M_n \to \infty$*,* $\mathbb{E}_{\theta_0} Q\left(\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \geq M_n\right) \to 0$*. Here,* $\mathbb{E}_{\theta_0}$ *is taking expectation over the true data generation process* $Y \sim P_{\theta_0}^n$*.*

Theorem 4.2.2 (Contraction result)**.** *We use* Q_h *to denote the approximate Bayesian distribution obtained by Algorithm 3 with bandwidth h. If Assumptions (4.2.1, 4.2.2', 4.2.3, 4.2.4, 4.2.5) hold with* $\sqrt{n}h = O(1)$*, then for any small constant* $\varepsilon > 0$*, there exists a corresponding large constant* M' *such that*

$$\mathbb{E}_{\theta_0} Q_h\left(\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \geq M'\right) \leq \varepsilon,$$

for large n, where $\mathbb{E}_{\theta_0}$ *is taking expectation over the true data generation process* $Y \sim P_{\theta_0}^n$*.*

Theorem 4.2.1 and Theorem 4.2.2 provide the contraction results of two distributions. Some results similar to Theorem 4.2.2 have been provided in [9, 33]. However, they used the probability concentration bound to show the contraction result and thus the bandwidth is required to be $h = \Omega(1/\sqrt{n})$. With above assumptions on density function, we can improve the contraction result with bandwidth $h = O(1/\sqrt{n})$.

The following two theorems character the shape of Bayesian distributions.

Theorem 4.2.3 (Bernstein-von Mises)**.** *Under all conditions in Theorem 4.2.1, suppose variable $\theta \sim Q$ for given $\widehat{\theta}_Y$, then we have that*

$$\sqrt{n}(\widehat{\theta}_Y - \theta) \mid \widehat{\theta}_Y \rightsquigarrow F(\theta_0, \cdot)$$

with high probability, where $F(\theta_0, \cdot)$ is the limit distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta_0)$, asymptotic convergence in distribution is measured via total variation distance, and the probability is taking over the true data generation process $Y \sim P_\theta^n$.

Theorem 4.2.4 (Bernstein-von Mises)**.** *Under all conditions in Theorem 4.2.2, for given θ_Y, if we run the ABC algorithm to sample θ with $h = o(1/\sqrt{n})$, then we have that*

$$\sqrt{n}(\widehat{\theta}_Y - \theta) \mid \widehat{\theta}_Y \rightsquigarrow F(\theta_0, \cdot)$$

with high probability, where $F(\theta_0, \cdot)$ is the limit distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta_0)$, asymptotic convergence in distribution is measured via total variation distance, and the probability is taking over the true data generation process $Y \sim P_\theta^n$.

From Assumption 4.2.1 and Assumption 4.2.2', it directly follows that $\sqrt{n}(\widehat{\theta}_Y - \theta_0) \rightsquigarrow F(\theta_0, \cdot)$. Theorem 4.2.3 and Theorem 4.2.4 prove that the proposed Bayesian distribution and approximate Bayesian distribution possess BvM property and are capable of recovering $F(\theta_0, \cdot)$ with large sample size. The main assumption to verify is Assumption 4.2.2', and thus we present the following lemma as one possible tool.

Lemma 4.2.1. *Let $X_n(\theta) = \sqrt{n}(\widehat{\theta}_X - \theta)$ with $X \sim P_\theta^n$ for simplicity. Suppose there exists some sufficiently small constant $\tilde{\delta}$, such that for all $\theta \in B(\theta_0, \tilde{\delta})$, $X_n(\theta) = X(\theta) + W_n(\theta)$, where $X(\theta) \sim F(\theta, \cdot)$ and $W_n(\theta) = o_P(1)$. If $\sup_{\theta \in B(\theta_0, \tilde{\delta})} \mathbb{E}\|W_n(\theta)\| = o(1)$, and Assumption 4.2.3 and Assumption 4.2.4 hold, then Assumption 4.2.2' holds.*

4.3 Efficient ABC Algorithms

Algorithm 3 provides a feasible way to sample from the Bayesian distribution, but in most cases, for a given loss function, to obtain $\widehat{\theta}_X = \arg\min_\theta L_n(\theta, X)$ for all fake data $X \sim P_\theta^n$ is complicate and inefficient. For instance, if the loss function is given by generative adversarial net, then both generation process and discrimination process involve optimization of neural networks, and thus updating $\widehat{\theta}_X$ for all fake data X is computationally expensive. To reduce the computational cost, in this section we propose an alternative efficient algorithm referred to as pseudo ABC without rigorous proof.

Algorithm 4: Pseudo approximate Bayesian computation

Input: Loss function L_n,

Frequentist estimate $\widehat{\theta}_Y = \arg\min_\theta L_n(\theta, Y)$,

Proposal distribution $S(\theta)$,

Choice of kernel function $\mathscr{K}$,

Large constant $K > \mathscr{K}(0) \max_\theta \frac{\pi(\theta)}{S(\theta)}$,

Tolerance level/bandwidth ε,

A positive sequence $M_n = o(\sqrt{n})$,

Number of iterations N.

for each $i \in \{1, 2, \ldots, N\}$ **do**

Generate $\theta^{(i)} \sim S(\theta)$;

if $\sqrt{n}\|\theta - \widehat{\theta}_Y\| \leq M_n$ **then**

Sample $X = (X_1, \ldots, X_n) \sim P^n_{\theta^{(i)}}$;

Calculate $S_n(\widehat{\theta}_Y, X) = \nabla_\theta L_n(\widehat{\theta}_Y, X)$;

Accept $\theta^{(i)}$ with probability

$$\mathscr{K}\left(\frac{S_n(\widehat{\theta}_Y, X)}{\varepsilon}\right)\frac{\pi(\theta)}{K \cdot S(\theta)}$$

else

Go to next iteration;

Output: A set of parameters $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$

As shown in Algorithm 4, pseudo ABC does not measure the distance between two summary statistics, but measures the norm of gradient. Compared with Algorithm 3, there are two major differences in pseudo ABC. The one is that we restrict all sampled θ into a small region around $\widehat{\theta}_Y$ due to the consistency of $\widehat{\theta}_Y$. The other is that we replace $\widehat{\theta}_X - \widehat{\theta}_Y$ in Algorithm 3 by $S_n(\widehat{\theta}_Y, X)$, the gradient of loss function $L_n(\theta, X)$ at point $\theta = \widehat{\theta}_Y$. Recall that $\widehat{\theta}_X = \arg\min_\theta L_n(\theta, X)$. The intuition is that if $\widehat{\theta}_Y$ is close to $\widehat{\theta}_X$, then the gradient at $\widehat{\theta}_Y$ is close to 0. We use P_ε to denote the

distribution generated by pseudo ABC algorithm, and density function is given by

$$p_\varepsilon(\theta) \propto \pi(\theta) \int \mathscr{K}\left(\frac{S_n(X,\widehat{\theta}_Y)}{\varepsilon}\right) dP_\theta^n(X) \cdot \mathbb{I}\left\{\sqrt{n}\|\theta - \widehat{\theta}_Y\| \le M_n\right\}, \tag{6}$$

where $M_n = o(\sqrt{n})$ is some positive sequence tending to ∞, and ε is the bandwidth of kernel function. In particular, if we take special kernel $\mathscr{K}(u) = \mathbb{I}\{\|u\| \le 1\}$, the corresponding distribution can be simplified as

$$p_\varepsilon(\theta) \propto \pi(\theta)\mathbb{P}_\theta\left\{\left\|S_n(X,\widehat{\theta}_Y)\right\| \le \varepsilon\right\} \mathbb{I}\left\{\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \le M_n\right\}.$$

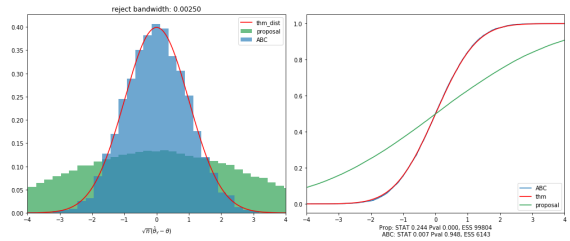
Algorithm 4 depends heavily on the choice of bandwidth ε. Possible improvements of algorithms are to adopt adaptive sequential ABC discussed in [25, 9, 6, 49]. In this work, we only focus on its property of recovering the shape of distribution, and it leaves as future work to better modify pseudo ABC algorithm

4.4 Numerical Results

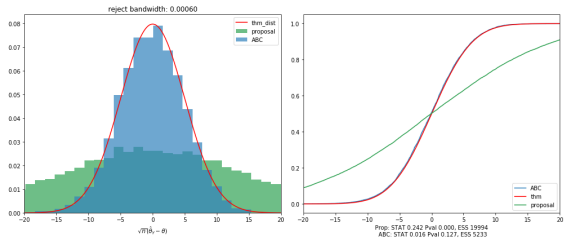
In this section, we present some experimental results by applying pseudo ABC algorithms defined in Algorithm 4, and examine how well the approximate Bayesian distribution recovers the distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta_0)$ with large sample. Notice that this paper does not focus on how to tune the bandwidth of ABC algorithms to get best performance. We first begin with toy experiments in one dimension and compare with the density plots. We then move to multi dimensional cases. We only use exponential kernel $\mathscr{K}_h(\mu) \propto \exp(-\frac{\|u\|^2}{2h})$ and box kernel $\mathscr{K}_h(\mu) \propto \mathbb{I}\{\|\mu\| \le h\}$ in the simulations. The code is available in `https://github.com/zhuobumeng/pseudo_ABC`.

Maximum likelihood estimates. We start with a toy example by taking MLE as our frequentist estimate, where the loss function is given by $L_n(\theta, X) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(Y_i)$, and $S_n(\theta, Y) = -\frac{1}{n}\sum_{i=1}^n \nabla_\theta \log p_\theta(Y_i)$. For the sake of clear presentation, we do one-dimensional experiment in all four cases and compare the resulting distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta)|\theta_Y$ with $\mathscr{N}\left(0, I_0^{-1}\right)$, where I_0

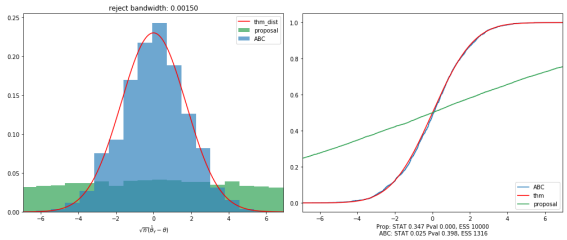
is fisher information matrix at true parameter θ_0. In each experiment, we generate $n = 10,000$ sample data from the true distribution, and then use pseudo ABC algorithm 4 to sample $N = 10,000$ parameters from the Bayesian distribution. For the sake of clear comparison, we use improper uniform prior in order not to induce any bias of distribution. The proposal distributions are set to be the same in all cases. The results are shown below.



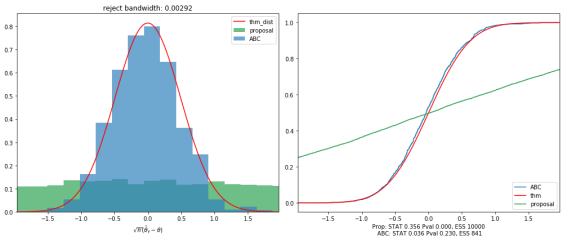
(a) Gaussian distribution $\mathcal{N}(\theta_0, 1)$



(b) Exponential distribution $\exp(\lambda)$



(c) Poisson distribution $Poisson(\lambda)$

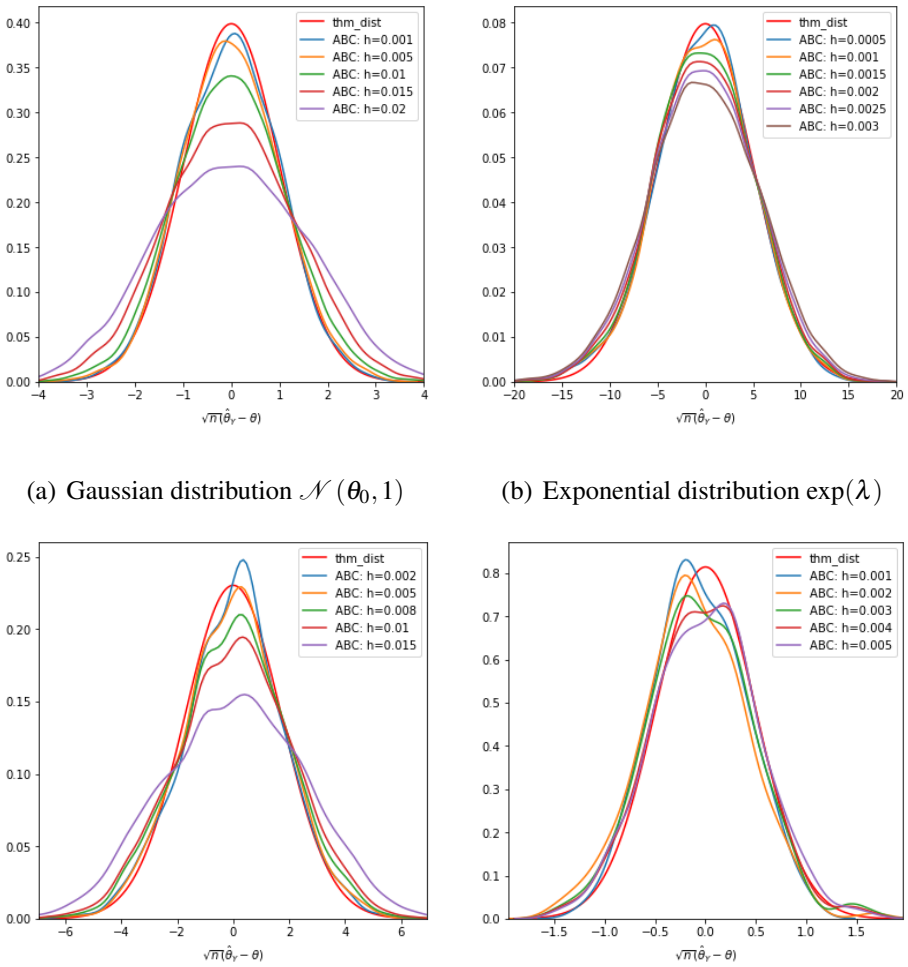


(d) Bernoulli distribution $Bernoulli(p)$

Figure 4.1: The left plot inside each subplot is density plot where green histogram is the proposal distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta)$ for given $\widehat{\theta}_Y$, the blue histogram is the sampled distribution, and the red curve is the theoretical distribution. The right plot inside each subplot is cumulative density plot where the green line is the proposal distribution, the blue line is the sampled distribution, and the red line is theoretical distribution.

In Figure 4.1, top two experiments (a, b) are for samples with continuous values, and the bottom two experiments (c, d) are for samples with discrete values. From Figure 4.1, pseudo ABC algorithm works well to recover the shape of the underlying true distribution, $\mathcal{N}\left(0, I_0^{-1}\right)$, with appropriate value of bandwidth h. As for CDF plots, all blue lines well align with red lines, again implying that approximate Bayesian distribution well recovers the true distribution. In order to see the influence of the choices of bandwidth, we then change different values of bandwidth, and the

result is shown below.



(a) Gaussian distribution $\mathcal{N}(\theta_0, 1)$ (b) Exponential distribution $\exp(\lambda)$

(c) Poisson distribution $Poisson(\lambda)$ (d) Bernoulli distribution $Bernoulli(p)$

Figure 4.2: Inside each subplot, red curve represents the theoretical density curve. Other lines in different color represent fitted density curve with different values of bandwidth h.

From Figure 4.2, when bandwidth is relatively smaller, the density fitted using sampled parameters is closer to the true distribution, which coincides the result in Theorem 4.2.4. We can refer to adaptive ABC procedure [57, 9, 6, 25] to appropriately choose the bandwidth h, which is beyond the scope of this work.

Quantile g-and-k Distribution. We then move to a more complex case. The univariate g-and-k distribution is a unimodal distribution that can characterize data with significant amounts of

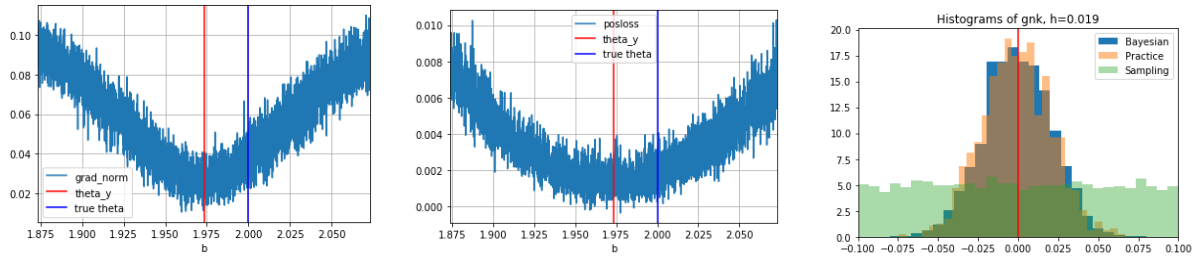
skewness and kurtosis. It is defined in terms of its quantile function:

$$q(r) = a + b\left(1 + 0.8\frac{1-\exp(-gz(r))}{1+\exp(-gz(r))}\right)(1+z(r)^2)^k z(r),$$

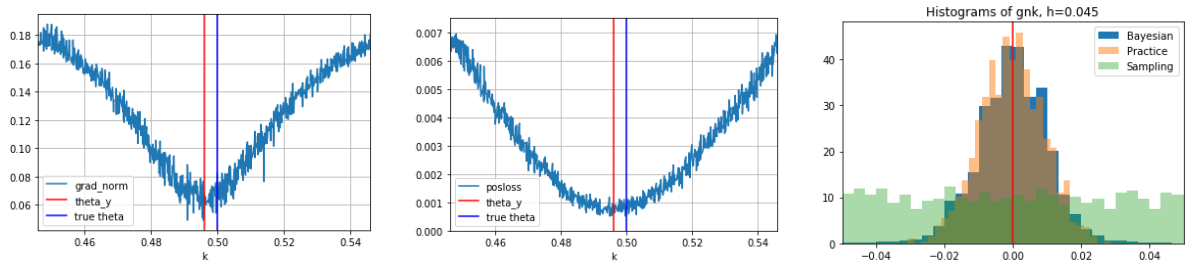
where $z(r)$ refers to r-th quantile of the standard Normal distribution. In this case, the likelihood formula is not analytically available, but it is still accessible to simulate synthetic data. For simplicity, we write the parameter of interest as θ, and represent the data generation as $G_\theta(z)$ where $z \sim \mathcal{N}(0,1)$, and use the function G to represent the transformation. One popular technique to do estimation in this case is adversarial network, which combines the generator and discriminator in one loss function, given by

$$\widehat{\theta}_Y = \arg\min_w \frac{1}{n}\sum_{i=1}^{n} D_w(Y_i) - \mathbb{E}\left[D_w(G_\theta(z))\right] - \frac{\lambda}{2}\|w\|^2,$$

where D is discriminator class usually constructed by neural network, and λ is the regularization constant. Here, we choose $D(x)$ as neural network with two hidden layers, layer-width is 20 for each layer, and choose regularization term $\lambda = 0.01$. Recall that the purpose of experiment is to naturally incorporate prior information, and also recover the frequentist distribution of $\widehat{\theta}_Y$ when sample size is large. We take the true parameters to be $a = 3, b = 2, g = 1, k = 1/2$, generate observation Y as our data with sample size 10,000, and solve optimization to obtain $\widehat{\theta}_Y$. Then, we plug $\widehat{\theta}_Y$ into pseudo ABC algorithm to generate parameters from the Bayesian distribution. Again for the sake of clear presentation, we choose prior distribution to be improper prior, and set proposal distribution as uniform distribution. Here, we present the results of estimating b and k separately, since the estimates for b, k given by the loss function seem consistent.



(a) Absolute value of gradient (b) Loss (c) Histogram



(d) Absolute value of gradient (e) Loss (f) Histogram

Figure 4.3: (a,b,c) are plotted when sampling b, and (d,e,f) are plotted when sampling k. (a,d) are the plots of gradient $S_n(\widehat{\theta}_Y, X)$ and (b,c) are the plots of loss function $L_n(\widehat{\theta}_Y, X)$, both with $X \sim P_\theta^n$. In (a,b,d,e), x-axis denotes the value of parameters θ sampled from the Bayesian distribution, the red vertical lines represent the value of $\widehat{\theta}_Y$ and the blue vertical lines represent θ_0. Here, (c,f) are the histogram plots, where the yellow histograms stand for the true distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta_0)$, obtained by generating data repeatedly and running optimization for 1,000 times, the green histograms stand for the proposal distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta)|\widehat{\theta}_Y$ for given $\widehat{\theta}_Y$, and the blue histograms stand for the approximate Bayesian distribution of $\sqrt{n}(\widehat{\theta}_Y - \theta)|\widehat{\theta}_Y$ for given $\widehat{\theta}_Y$.

The result in Figure 4.3 shows that pseudo ABC algorithm works well that the Bayesian distribution almost perfectly recover the shape of the underlying true distribution. Not to mention that it took more than 15 hours to simulate the true distribution, while it took less than 1.5 hours to run pseudo ABC algorithm. According to first two columns of plots, we can tell that Bayesian distribution centers around $\widehat{\theta}_Y$, and the loss and gradient is smaller when the sampled θ is closer to $\widehat{\theta}_Y$.

Multivariate generative adversarial model. We then move to study the performance on multi-

variate case. We study Gaussian distribution $\mathcal{N}(\theta, I_d)$ with the loss function given by

$$\min_{\theta} \max_{w,b} \frac{1}{n} \sum_{i=1}^{n} \log S(w^T Y_i + b) + \mathbb{E}\left[\log\left(1 - S(w^T G_\theta(Z) + b\right))\right],$$

where $Z \sim \mathcal{N}(0, I_d)$, $G_\theta(Z) = Z + \theta$, and $Y_1, \ldots Y_n \sim \mathcal{N}(\theta_0, I_d)$. It is a complex generative adversarial model but the solution is simply given by sample mean according to [34], which makes it a good example to test the performance of pseudo ABC algorithm on multivariate generative adversarial network. We set $d = 4$ and $n = 10,000$ with true parameter is $\theta_0 = (0.1, 0.1, 0.1, 0.1)^T$. Again we set the prior to be improper uniform prior, and apply pseudo ABC to draw 50,000 samples from the Bayesian distribution, and the result is shown below.

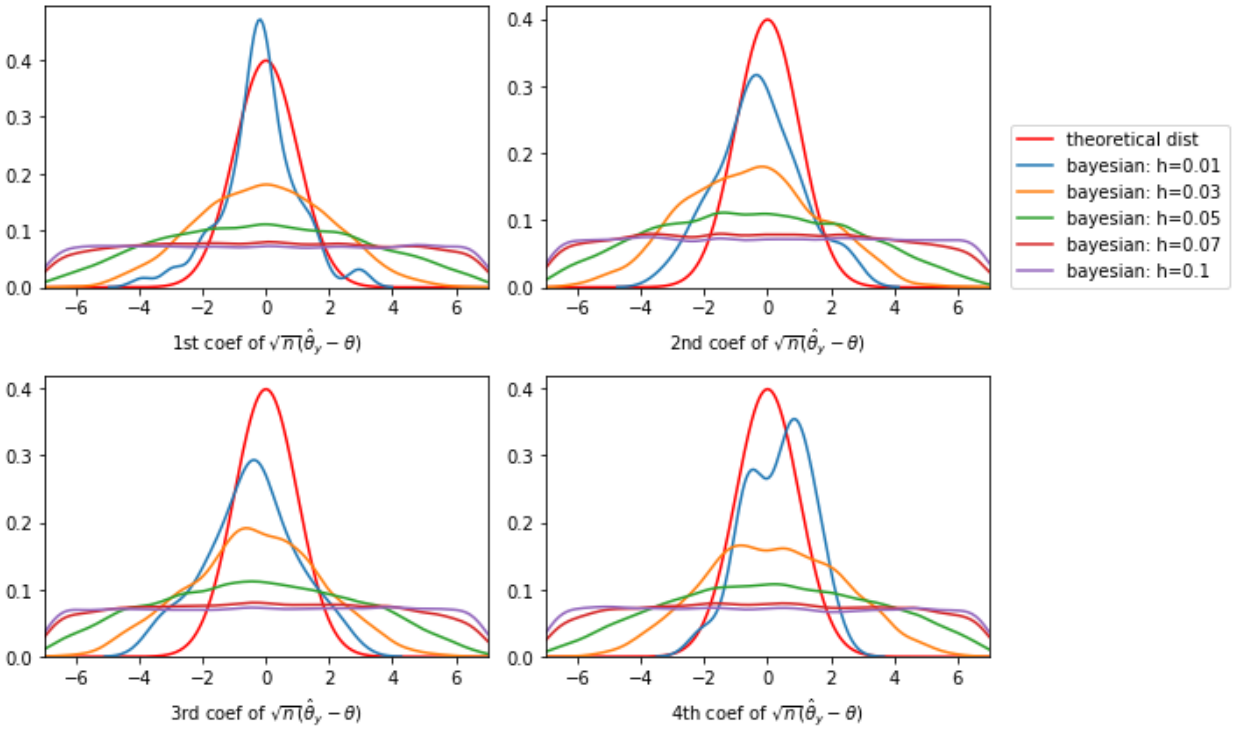


Figure 4.4: From top left to bottom right, it shows the density plots for four coefficients of $\sqrt{n}(\widehat{\theta}_Y - \theta)|\widehat{\theta}_Y$, with θ drawn from Bayesian distribution using pseudo ABC and $\widehat{\theta}_Y$ is given. In each plot, red line represents the theoretical density, and other lines in different colors represent different bandwidth h.

From Figure 4.4, we can see that when bandwidth is smaller, Bayesian distribution better recovers the true distribution, as the shape of blue curve is the closest to the red curve. However, when dimension is getting larger, it gets harder to accept samples with small bandwidth. It remains as future work to modify the pseudo ABC algorithm such that the proposal distribution gets updated during the process to be closer to the true distribution, and hence more samples can be accepted in the end.

4.5 Ancillary Analysis

Before ending this chapter, we present the following ancillary lemma and theorem for pseudo ABC algorithm, but the condition is hard to check.

Assumption 4.5.1. *Suppose there exists some positive definite constant matrix I_0, some fixed kernel function $\mathscr{K}$, and some sufficiently small constant δ, such we have*

$$\sup_{\theta,\widehat{\theta}_Y \in B(\theta_0,\delta)} \frac{\mathbb{E}_{X\sim P_\theta^n}\left[\mathscr{K}\left(\frac{\tau_n S_n(X,\widehat{\theta}_Y)}{\varepsilon_n}\right)\right]}{\mathbb{E}_{X\sim P_\theta^n}\left[\mathscr{K}\left(\frac{I_0\sqrt{n}(\widehat{\theta}_X-\widehat{\theta}_Y)}{\varepsilon_n'}\right)\right]} = 1+o(1),$$

for some sequences $\varepsilon_n, \varepsilon_n' \to 0$, where $1/\tau_n$ is the rate of $S_n(X,\widehat{\theta}_Y)$.

Theorem 4.5.1. *Suppose $\theta_{PABC}|\widehat{\theta}_Y \sim P_\varepsilon$ generated by pseudo ABC, where the density takes the form as*

$$p_\varepsilon(\theta) \propto \pi(\theta)\int \mathscr{K}\left(\frac{\tau_n S_n(X,\widehat{\theta}_Y)}{\varepsilon_n}\right)\prod_{i=1}^n p_\theta(X_i)dX \cdot \mathbb{I}\left\{\sqrt{n}\|\theta-\widehat{\theta}_Y\| \le M_n\right\}.$$

Under Assumptions (4.2.1, 4.2.2', 4.2.3, 4.2.4, 4.2.5,4.5.1), if $\varepsilon_n \to 0$, then we have $\sqrt{n}(\widehat{\theta}_Y - \theta_{ABCP})|\widehat{\theta}_Y \rightsquigarrow F(\theta_0,\cdot)$ with high probability.

The proofs are deferred to Appendix. The above theorem provides some intuition that it is reasonable to replace Algorithm 3 with pseudo ABC algorithm under some conditions.

4.6 Proof

4.6.1 Proof of Theorem 4.2.1 and Theorem 4.2.2

Proof of Theorem 4.2.1. For the sake of clear presentation, we write

$$T(\theta, M) = \mathbb{P}\left\{\sqrt{n}\left\|\widehat{\theta}_X - \theta\right\| \geq M \;\middle|\; X \sim P_\theta^n\right\} \tag{7}$$

to simplify the notation.

For some M to be specified later, denote event $\mathscr{E} = \left\{\sqrt{n}\left\|\widehat{\theta}_Y - \theta_0\right\| \leq M\right\}$, and

$$\mathbb{E}_{\theta_0}\left[Q\left(\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \geq M'\right)\right] \leq \mathbb{E}_{\theta_0}\left[Q\left(\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \geq M'\right)\mathbb{I}\{\mathscr{E}\}\right] + \mathbb{P}_{\theta_0}\left\{\mathscr{E}^C\right\}.$$

The second term can be trivially bounded by Assumption 4.2.2, and we proceed to bound the first term. Let random variable $T = \sqrt{n}(\widehat{\theta}_Y - \theta)$ with distribution $\tilde{Q}$, and its pdf $\tilde{q}(t)$ is defined as

$$\tilde{q}(t) = \frac{1}{n^{d/2}} q(\widehat{\theta}_Y - t/\sqrt{n}) = \frac{\pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t)}{\int \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t) dt}.$$

Thus, we have

$$\tilde{Q}(\|T\| \geq M') = \frac{\int_{\|t\| \geq M'} \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t) dt}{\int_t \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t) dt} = \frac{(A)}{(B)}.$$

Under event $\mathscr{E}$, it follows that

$$\begin{aligned}
(B) &\geq \int_{\|t\| \leq M_1} \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t) dt \\
&\geq \left(1 - (M_1 + M) L_\pi / \sqrt{n}\right) \pi(\theta_0) \int_{\|t\| \leq M_1} f_n(\widehat{\theta}_Y - t/\sqrt{n}, t) dt \qquad \text{(By Assumption 4.2.5)} \\
&= (1 - o(1)) \pi(\theta_0) \left(\int_{\|t\| \leq M_1} f_n(\theta_0, t) dt + \int_{\|t\| \leq M_1} \left(f_n(\widehat{\theta}_Y - t/\sqrt{n}, t) - f_n(\theta_0, t) \right) dt \right),
\end{aligned}$$

where

$$\int_{\|t\|\le M_1} f_n(\theta_0,t)dt = \mathbb{P}_{\theta_0}\left\{\sqrt{n}\left\|\widehat{\theta}_X-\theta_0\right\|\le M_1\right\} = 1-T(\theta_0,M_1).$$

Furthermore, under the event $\mathscr{E}$ and by Assumption 4.2.3, it follows that

$$\begin{aligned}
&\int_{\|t\|\le M_1}\left(f_n(\widehat{\theta}_Y-t/\sqrt{n},t)-f_n(\theta_0,t)\right)dt\\
\ge&-\int_{\|t\|\le M_1}\left|\frac{f_n(\widehat{\theta}_Y-t/\sqrt{n},t)}{f_n(\theta_0,t)}-1\right| f_n(\theta_0,t)dt\\
=&-\int_{\|t\|\le M_1}\left|\frac{f_n(\widehat{\theta}_Y-t/\sqrt{n},t)}{f(\widehat{\theta}_Y-t/\sqrt{n},t)}\frac{f(\theta_0,t)}{f_n(\theta_0,t)}\frac{f(\widehat{\theta}_Y-t/\sqrt{n},t)}{f(\theta_0,t)}-1\right| f_n(\theta_0,t)dt\\
\ge&-C(M+M_1)L_1/\sqrt{n},
\end{aligned}$$

for some constant C, where L_1 depends on constant M_1. Thus, to conclude, under event $\mathscr{E}$, we have

$$(B)\ge(1-C(M_1+M)L_\pi/\sqrt{n})\pi(\theta_0)\left(1-T(\theta_0,M_1)-C(M+M_1)L_1/\sqrt{n}\right), \tag{8}$$

for some constant L_1 depending on M_1. The above (8) holds for any large constant M_1, and there exists some $M_1\to\infty$ such that $(B)\ge(1-o(1))\pi(\theta_0)$.

We now proceed to lower bound (A). We can always decompose (A) as

$$(A)=\int_{\|t\|\ge\sqrt{n}\delta} f_n d\pi+\int_{M'\le\|t\|\le\sqrt{n}\delta} f_n d\pi=(A_1)+(A_2), \tag{9}$$

for any sufficiently small constant δ (to be specified later). Write variable $S=\sqrt{n}(\widehat{\theta}_Y-\theta_0)$, and $S|\theta_0\sim F_n(\theta_0,\cdot)$ with pdf as $f_n(\theta_0,\cdot)$. Then, the first term in the above equation is bounded in

expectation by

$$\begin{aligned}
\mathbb{E}_{\theta_0}[A_1\mathbb{I}\{\mathscr{E}\}] &= \mathbb{E}_{\theta_0}\left[\int_{\|t\|\geq\sqrt{n}\delta}\pi\left(\widehat{\theta}_Y - t/\sqrt{n}\right) f_n\left(\widehat{\theta}_Y - t/\sqrt{n}, t\right)\mathbb{I}\{\mathscr{E}\}\,dt\right] \\
&\leq \int_{\|s\|\leq M}\int_{\|t\|\geq\sqrt{n}\delta}\pi\left(\theta_0 + \frac{s-t}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s-t}{\sqrt{n}}, t\right) f_n(\theta_0, s)\,dtds \\
&= \int_{\|s\|\leq M}\int_{\|t\|\geq\sqrt{n}\delta}\pi\left(\theta_0 + \frac{s-t}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s-t}{\sqrt{n}}, t\right)\frac{f_n(\theta_0, s)}{f(\theta_0, s}\,dtds \\
&\leq (1+o(1))\int_{s'}\int_{\|t\|\geq\sqrt{n}\delta}\pi\left(\theta_0 + \frac{s'}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s'}{\sqrt{n}}, t\right) f(\theta_0, s'+t)\,dtds' \\
&\qquad\qquad \text{(By change of variable, and Assumption 4.2.2')} \\
&\leq (1+o(1))\sup_t f(\theta_0, t)\cdot\sup_{\theta\in\Theta}\mathbb{P}_\theta\left\{\left\|\widehat{\theta}_X - \theta\right\| \geq \delta\right\}. \\
&\qquad\qquad \text{(Assumption 4.2.1 and Assumption 4.2.4)}
\end{aligned}$$

Similarly, the second term in (9) is bounded by

$$\begin{aligned}
\mathbb{E}_{\theta_0}[A_2\mathbb{I}\{\mathscr{E}\}] =& \mathbb{E}_{\theta_0}\left[\int_{M'\leq\|t\|\leq\sqrt{n}\delta}\pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t)\mathbb{I}\{\mathscr{E}\}\,dt\right] \\
&\leq (1+o(1))\int_{\|s\|\leq M}\int_{M'\leq\|t\|\leq\sqrt{n}\delta}\pi\left(\theta_0 + \frac{s-t}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s-t}{\sqrt{n}}, t\right) f(\theta_0, s)\,dtds \\
&\leq (1+o(1))\int_{\|s'\|\leq M+\sqrt{n}\delta}\int_{M'\leq\|t\|\leq\sqrt{n}\delta}\pi\left(\theta_0 + \frac{s'}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s'}{\sqrt{n}}, t\right) f(\theta_0, s'+t)\,dtds' \\
&\leq \sup_t f(\theta_0, t)\cdot\sup_{\theta\in B(\theta_0, \delta + M/\sqrt{n})}\mathbb{P}_\theta\left\{\sqrt{n}\left\|\widehat{\theta}_X - \theta\right\| \geq M'\right\}, \\
&\qquad\qquad \text{(By Assumption 4.2.2 implied by Assumption 4.2.2', and Assumption 4.2.4)}
\end{aligned}$$

where we choose δ sufficiently small such that $B(\theta_0, \delta + M/\sqrt{n}) \subset U$ as defined in Assumption 4.2.2'. Combining all bounds and we have

$$\mathbb{E}_{\theta_0} Q\left(\sqrt{n}\left\|\theta - \widehat{\theta}_Y\right\| \geq M'\right) \leq T(\theta_0, M) + (1+o(1))\frac{\sup_t f(\theta_0, t)}{\pi(\theta_0)}\left(\sup_\Theta T(\theta, \sqrt{n}\delta) + \sup_{\theta\in B(\theta_0, \tilde{\delta})} T(\theta, M')\right).$$

Choosing M and M' to be large constants directly leads to the result. □

Proof of Theorem 4.2.3. Define $\mathscr{U} = \{t : \|t\| \leq M'\}$ for some large constant M', and define event

$\mathscr{E} = \left\{ \sqrt{n} \left\| \widehat{\theta}_Y - \theta_0 \right\| \leq M \right\}$. Let $\tilde{Q}^{\mathscr{U}}$ be the distribution of variable $T = \sqrt{n}(\widehat{\theta}_Y - \theta)$ restricted to region $\mathscr{U}$. Then, we have

$$\begin{aligned}
\left\| \tilde{Q} - F_n(\theta_0, \cdot) \right\|_{\mathsf{TV}} &\leq \left\| \tilde{Q} - \tilde{Q}^{\mathscr{U}} \right\|_{\mathsf{TV}} + \left\| \tilde{Q}^{\mathscr{U}} - F_n^{\mathscr{U}}(\theta_0, \cdot) \right\|_{\mathsf{TV}} + \left\| F_n^{\mathscr{U}}(\theta_0, \cdot) - F_n(\theta_0, \cdot) \right\|_{\mathsf{TV}} \\
&\leq \left\| \tilde{Q}^{\mathscr{U}} - F_n^{\mathscr{U}}(\theta_0, \cdot) \right\|_{\mathsf{TV}} + 2\tilde{Q}(\mathscr{U}^C) + 2F_n(\theta_0, \mathscr{U}^C),
\end{aligned}$$

where the latter two terms are both negligible, and

$$\begin{aligned}
&\left\| \tilde{Q}^{\mathscr{U}} - F_n^{\mathscr{U}}(\theta_0, \cdot) \right\|_{\mathsf{TV}} \\
=&\frac{1}{2} \int_{\mathscr{U}} \left| \frac{\tilde{q}(t)}{\int_{\mathscr{U}} \tilde{q}(h) dh} - \frac{f_n(\theta_0, t)}{\int_{\mathscr{U}} f_n(\theta_0, h) dh} \right| dt \\
=&\frac{1}{2} \int_{\mathscr{U}} \left| \frac{\pi(\widehat{\theta}_Y - t/\sqrt{n})}{\pi(\theta_0)} \frac{f_n(\widehat{\theta}_Y - t/\sqrt{n}, t)}{f_n(\theta_0, t)} \cdot \frac{\pi(\theta_0) \int_{\mathscr{U}} f_n(\theta_0, h) dh}{\int_{\mathscr{U}} \pi(\widehat{\theta}_Y - h/\sqrt{n}) f_n(\widehat{\theta}_Y - h/\sqrt{n}, h) dh} - 1 \right| f_n^{\mathscr{U}}(\theta_0, t) dt.
\end{aligned}$$

Under event $\mathscr{E}$, it follows that for $\|t\| \leq M'$,

$$\frac{\pi(\widehat{\theta}_Y - t/\sqrt{n})}{\pi(\theta_0)} = e^{\log \pi(\widehat{\theta} - t/\sqrt{n}) - \log \pi(\theta_0)} = 1 + \mathscr{O}\left(L_\pi (M + M')/\sqrt{n} \right),$$

(By Assumption 4.2.5)

$$\frac{f_n(\widehat{\theta}_Y - t/\sqrt{n}, t)}{f_n(\theta_0, t)} = \frac{f(\widehat{\theta}_Y - t/\sqrt{n}, t)}{f(\theta_0, t)}(1 + o(1)) = e^{\log f(\widehat{\theta}_Y - t/\sqrt{n}, t) - \log f(\theta_0, t)} = 1 + \mathscr{O}\left(L_1 (M + M')/\sqrt{n} \right).$$

(By Assumption 4.2.2' and Assumption 4.2.3)

Thus, $\left\| \tilde{Q}^{\mathscr{U}} - F_n^{\mathscr{U}}(\theta_0, \cdot) \right\|_{\mathsf{TV}} \mathbb{I}\{\mathscr{E}\} = o(1)$, and we have

$$\mathbb{E}_{\theta_0} \left\| \tilde{Q} - F_n(\theta_0, \cdot) \right\|_{\mathsf{TV}} \leq \mathbb{E}_{\theta_0} \left[\left\| \tilde{Q} - F_n(\theta_0, \cdot) \right\|_{\mathsf{TV}} \mathbb{I}\{\mathscr{E}\} \right] + \mathbb{P}_{\theta_0} \left\{ \mathscr{E}^C \right\} \leq \varepsilon$$

holds for any sufficiently small constant ε. By assumption, $\sqrt{n}(\widehat{\theta}_Y - \theta_0) \rightsquigarrow F(\theta_0, \cdot)$, and it directly leads to the result. □

Proof of Theorem 4.2.4. Denote $\mathscr{E} = \left\{\sqrt{n}\|\widehat{\theta}_Y - \theta_0\| \le M\right\}$, and $T = \sqrt{n}(\widehat{\theta}_Y - \theta)$. Follow the same argument in the proof of Theorem 4.2.1, it suffices to upper bound

$$\tilde{Q}_h\left(\|T\| \ge M'\right) = \frac{\int_{\|t\|\ge M'}\int_s f_n \mathscr{K}_{\sqrt{n}h} ds d\pi}{\int_t\int_s f_n \mathscr{K}_{\sqrt{n}h} ds d\pi} = \frac{(A)}{(B)},$$

under event $\mathscr{E}$ in expectation w.r.t. P_{θ_0}. Under $\mathscr{E}$, it follows that

$$\begin{aligned}
(B) &\ge \int_{\|t\|\le M_1} \pi(\widehat{\theta}_Y - t/\sqrt{n}) \int_s f_n(\widehat{\theta}_Y - t/\sqrt{n}, s)\mathscr{K}_{\sqrt{n}h}(s-t)\, ds dt \\
&= \int_{\|t\|\le M_1} \pi(\widehat{\theta}_Y - t/\sqrt{n}) \int_u f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu)\mathscr{K}(u)\, du dt \\
&= \int_u \int_{\|t\|\le M_1} \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu) dt \mathscr{K}(u)\, du.
\end{aligned}$$

There always exists some large constant R such that $\mathscr{K}\{u : \|u\| > R\} \le \varepsilon_R$ is negligible. Define $K_R = \mathscr{K}\{u : \|u\| \le R\}$. Thus, we have

$$\begin{aligned}
(B) &\ge \int_{\|u\|\le R}\int_{\|t\|\le M_1} \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu) dt \frac{\mathscr{K}(u)}{K_R} du \cdot K_R \\
&\ge C\pi(\theta_0) \int_{\|u\|\le R}\int_{\|t\|\le M_1} f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu) dt \frac{\mathscr{K}(u)}{K_R} du \qquad \text{(By Assumption 4.2.5)} \\
&= C\pi(\theta_0) \int_{\|u\|\le R}\int_{\|t\|\le M_1} f_n(\theta_0, t+\sqrt{n}hu) + \\
&\quad \left(f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu) - f_n(\theta_0, t+\sqrt{n}hu)\right) dt \frac{\mathscr{K}(u)}{K_R} du.
\end{aligned}$$

It follows that

$$\int_{\|u\|\le R}\int_{\|t\|\le M_1} f_n(\theta_0, t+\sqrt{n}hu) dt \frac{\mathscr{K}(u)}{K_R} du \ge \mathbb{P}_{\theta_0}\left\{\sqrt{n}\left\|\widehat{\theta}_X - \theta_0\right\| \le M_1 - \sqrt{n}hR\right\}.$$

The above quantity is close to 1 if we choose M_1 to be sufficiently large. We also have

$$\begin{aligned}
&\int_{\|u\|\le R}\int_{\|t\|\le M_1} f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu) - f_n(\theta_0, t+\sqrt{n}hu)dt\frac{\mathscr{K}(u)}{K_R}du \\
\ge &-\int_{\|u\|\le R}\int_{\|t\|\le M_1}\left|\frac{f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu)}{f_n(\theta_0, t+\sqrt{n}hu)} - 1\right| f_n(\theta_0, t+\sqrt{n}hu)dt\frac{\mathscr{K}(u)}{K_R}du \\
\ge &-OL_1(M+M_1)/\sqrt{n},
\end{aligned}$$

the last inequality follows Assumption 4.2.3 with some constant L_1 depending on $(M+\sqrt{n}hR)$. Then, We proceed to upper bound (A). For any small constant δ, we can decompose (A) into

$$(A) = \int_{\|t\|\ge\sqrt{n}\delta}\int_s f_n \mathscr{K}_{\sqrt{n}h} ds d\pi + \int_{M'\le\|t\|\le\sqrt{n}\delta}\int_s f_n \mathscr{K}_{\sqrt{n}h} ds d\pi = (A_1) + (A_2).$$

By some calculations, we have

$$\begin{aligned}
&\mathbb{E}_{\theta_0}[A_1 \mathbb{I}\{\mathscr{E}\}] \\
\le &(1+o(1))\int_u\int_{\|s\|\le M}\int_{\|t\|\ge\sqrt{n}\delta} \pi\left(\theta_0 + \frac{s-t}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s-t}{\sqrt{n}}, t+\sqrt{n}hu\right) f(\theta_0, s) dt ds \mathscr{K}(u) du \\
\le &(1+o(1))\int_u\int_{s'}\int_{\|t\|\ge\sqrt{n}\delta} \pi\left(\theta_0 + \frac{s'}{\sqrt{n}}\right) f_n\left(\theta_0 + \frac{s'}{\sqrt{n}}, t+\sqrt{n}hu\right) f(\theta_0, s'+t) dt ds' \mathscr{K}(u) du \\
\le &(1+o(1))\sup_t f(\theta_0, t)\left(1 - K_R + K_R \cdot \sup_{\theta\in\Theta}\mathbb{P}_\theta\left\{\left\|\widehat{\theta}_X - \theta\right\| \ge \delta - hR\right\}\right),
\end{aligned}$$

where the last inequality follows by dividing the domain of U into $\{\|u\| \le R\}$ and its complement.

Similarly, we have

$$
\begin{aligned}
&\mathbb{E}_{\theta_0}[A_2\mathbb{I}\{\mathscr{E}\}]\\
=&\mathbb{E}_{\theta_0}\left[\int_{M'\le\|t\|\le\sqrt{n}\delta}\int_u \pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu)\mathscr{K}(u)\mathbb{I}\{\mathscr{E}\}\,dudt\right]\\
\lesssim&\int_u\int_{\|s\|\le M}\int_{M'\le\|t\|\le\sqrt{n}\delta}\pi\left(\theta_0+\frac{s-t}{\sqrt{n}}\right)f_n\left(\theta_0+\frac{s-t}{\sqrt{n}}, t+\sqrt{n}hu\right)f(\theta_0,s)dtds\mathscr{K}(u)du\\
\lesssim&\int_u\int_{\|s'\|\le M+\sqrt{n}\delta}\int_{M'\le\|t\|\le\sqrt{n}\delta}\pi\left(\theta_0+\frac{s'}{\sqrt{n}}\right)f_n\left(\theta_0+\frac{s'}{\sqrt{n}}, t+\sqrt{n}hu\right)f(\theta_0,s'+t)dtds'\mathscr{K}(u)du\\
\lesssim&\sup_t f(\theta_0,t)\left(1-K_R+K_R\cdot\sup_{\theta\in B(\theta_0,\delta+M/\sqrt{n})}\mathbb{P}\left\{\sqrt{n}\left\|\widehat{\theta}_X-\theta\right\|\ge M'-\sqrt{n}hR\right\}\right).
\end{aligned}
$$

Again, we choose δ sufficiently small such that $B(\theta_0, M/\sqrt{n}+\delta)\subset U$ defined in Assumption 4.2.2. We choose $M'-\sqrt{n}hR$ to be a large constant such that the the latter part is negligible. To conclude, for any $h\lesssim\mathscr{O}(1/\sqrt{n})$, for any sufficiently small ε, we can choose R, M' to be sufficiently large correspondingly, such that the result holds. □

Proof of Theorem 4.2.4. We first define some quantities and distributions for the clear presentation of proof. Let $\mathscr{E}=\left\{\sqrt{n}(\widehat{\theta}_Y-\theta_0)\le M\right\}$ for some large constant M. We consider the joint variables $(U,T)\sim\tilde{Q}_{U,T}$, with pdf

$$\tilde{q}_{U,T}(u,t)\propto\pi(\widehat{\theta}_Y-t/\sqrt{n})f_n(\widehat{\theta}_Y-t/\sqrt{n},t+\sqrt{n}hu)\mathscr{K}(u).$$

Let $\tilde{Q}_T$ be marginal distribution of variable T with density function given by

$$\tilde{q}_T(t)=\int_u\tilde{q}_{U,T}(u,t)du,$$

and it suffices to show that

$$\mathbb{E}_{\theta_0}\left\|\tilde{Q}_T-F_n(\theta_0,\cdot)\right\|_{\mathsf{TV}}\to 0.$$

As proved in Theorem 4.2.2, for any large constant R with $K_R=\mathscr{K}(\|U\|\le R)\ge 1-\varepsilon_R$, we

have

$$\mathbb{E}_{\theta_0}\tilde{Q}_{U,T}(\|U\| \geq R) \leq \mathbb{E}_{\theta_0}\tilde{Q}_{U,T}(\|U\| \geq R)\mathbb{I}\{\mathscr{E}\} + \mathbb{P}_{\theta_0}\{\mathscr{E}\} \leq C\frac{\sup_t f(\theta_0,t)}{\pi(\theta_0)}(1-K_R) + \mathbb{P}_{\theta_0}\{\mathscr{E}\}. \tag{10}$$

Define region $\mathscr{U} = \{(t,u) : \|t\| \leq M', \|u\| \leq R\}$ for some large constants M', R, and $\mathbb{E}_{\theta_0}\tilde{Q}_{U,T}(\mathscr{U}^C)$ is negligible by Theorem 4.2.2 and (10). Then, we define the truncated distribution of (U,T) restricted to region $\mathscr{U}$ as $\tilde{Q}^{\mathscr{U}}_{U,T}$.

Define another joint distribution of (U,T) as $\tilde{F}_{U,T}$ with pdf

$$\tilde{f}_{U,T}(u,t) = f_n(\theta_0, t+\sqrt{n}hu)\mathscr{K}(u),$$

and the marginal distribution of T as $\tilde{F}_T$ with pdf

$$\tilde{f}_U(t) = \int_u f_n(\theta_0, t+\sqrt{n}hu)\mathscr{K}(u)du.$$

It follows that

$$\tilde{F}_{U,T}(\mathscr{U}^c) \leq \mathscr{K}(\|U\| > R) + \mathbb{P}_{\theta_0}\left\{\sqrt{n}\|\widehat{\theta}_X - \theta_0\| \geq M' - \sqrt{n}hR\right\},$$

which is negligle as well if R and $M' - \sqrt{n}hR$ is large. We also define the distribution of (U,T) restricted to region $\mathscr{U}$ as $\tilde{F}^{\mathscr{U}}_{U,T}$. Now we proceed to prove the theorem.

We first upper bound the total variation distance between $\tilde{Q}_T$ and $F_n(\theta_0,\cdot)$ by

$$\left\|\tilde{Q}_T - F_n(\theta_0,\cdot)\right\|_{\mathsf{TV}} \leq \left\|\tilde{Q}_T - \tilde{F}_T\right\|_{\mathsf{TV}} + \left\|\tilde{F}_T - F_n(\theta_0,\cdot)\right\|_{\mathsf{TV}} \tag{11}$$

where the second term on the RHS is bounded by

$$
\begin{aligned}
&2\left\|\tilde{F}_T - F_n(\theta_0,\cdot)\right\|_{\mathsf{TV}} \\
=&\int_t \left| \int_u f_n(\theta_0, t+\sqrt{n}hu)\mathscr{K}(u)du - f_n(\theta_0,t)\right| dt \\
\leq&\int_{\|t\|\leq M'} \left| \int_{\|u\|\leq R} f_n(\theta_0, t+\sqrt{n}hu)\mathscr{K}(u)du - f_n(\theta_0,t)\right| dt + \tilde{F}_{U,T}(\mathscr{U}^c) + T(\theta_0, M'),
\end{aligned}
$$

($T(\theta, M')$ is defined in (7) for simplicity.)

where

$$
\begin{aligned}
&\int_{\|t\|\leq M'} \left| \int_{\|u\|\leq R} f_n(\theta_0, t+\sqrt{n}hu)\mathscr{K}(u)du - f_n(\theta_0,t)\right| dt \\
\leq&\int_{\|t\|\leq M}\int_{\|u\|\leq R} \left| f_n(\theta_0, t+\sqrt{n}hu) - f_n(\theta_0,t)\right| \mathscr{K}(u)dudt \\
\leq&CL_2\sqrt{n}hR, \qquad \text{(By Assumption 4.2.3)}
\end{aligned}
$$

where L_2 is some constant depending on the quantity $(M' + \sqrt{n}hR)$. The first term in (11) on the RHS is bounded by

$$
\left\|\tilde{Q}_T - \tilde{F}_T\right\|_{\mathsf{TV}} \leq \left\|\tilde{Q}_{U,T} - \tilde{F}_{U,T}\right\|_{\mathsf{TV}} \leq \left\|\tilde{Q}^{\mathscr{U}}_{U,T} - \tilde{F}^{\mathscr{U}}_{U,T}\right\|_{\mathsf{TV}} + 2\tilde{Q}_{U,T}(\mathscr{U}^c) + 2\tilde{F}_{U,T}(\mathscr{U}^c),
$$

where $\tilde{Q}_{U,T}(\mathscr{U}^c)$ and $\tilde{F}_{U,T}(\mathscr{U}^c)$ are negligible by choosing M', R to be sufficiently large, and the first term can be further bounded by

$$
\left\|\tilde{Q}^{\mathscr{U}}_{U,T} - \tilde{F}^{\mathscr{U}}_{U,T}\right\|_{\mathsf{TV}} = \int\int_{(u,t)\in\mathscr{U}} \left| \frac{\tilde{q}^{\mathscr{U}}_{U,T}(u,t)}{\tilde{f}^{\mathscr{U}}_{U,T}(u,t)} - 1\right| \tilde{f}^{\mathscr{U}}_{U,T}(u,t)dudt,
$$

where

$$
\begin{aligned}
&\frac{\tilde{q}^{\mathscr{U}}_{U,T}(u,t)}{\tilde{f}^{\mathscr{U}}_{U,T}(u,t)} \\
=&\frac{\pi(\widehat{\theta}_Y - t/\sqrt{n}) f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu)\mathscr{K}(u)}{f_n(\theta_0, t+\sqrt{n}hu)\mathscr{K}(u)} \cdot \frac{\int\int_{(s,v)\in\mathscr{U}} f_n(\theta_0, s+\sqrt{n}hv)\mathscr{K}(v)dvds}{\int\int_{(s,v)\in\mathscr{U}} \pi(\widehat{\theta}_Y - s/\sqrt{n}) f_n(\widehat{\theta}_Y - s/\sqrt{n}, s+\sqrt{n}hv)\mathscr{K}(v)dvds} \\
=&\frac{\pi(\widehat{\theta}_Y - t/\sqrt{n})}{\pi(\theta_0)} \cdot \frac{f_n(\widehat{\theta}_Y - t/\sqrt{n}, t+\sqrt{n}hu)}{f_n(\theta_0, t+\sqrt{n}hu)} \cdot \frac{\int\int_{(s,v)\in\mathscr{U}} \pi(\theta_0) f_n(\theta_0, s+\sqrt{n}hv)\mathscr{K}(v)dvds}{\int\int_{(s,v)\in\mathscr{U}} \pi(\widehat{\theta}_Y - s/\sqrt{n}) f_n(\widehat{\theta}_Y - s/\sqrt{n}, s+\sqrt{n}hv)\mathscr{K}(v)dvds}.
\end{aligned}
$$

Thus, under the event $\mathscr{E}$, by Assumption 4.2.5 and Assumption 4.2.3, we have

$$
\left\| \tilde{Q}^{\mathscr{U}}_{U,T} - \tilde{F}^{\mathscr{U}}_{U,T} \right\|_{\mathsf{TV}} \mathbb{I}\{\mathscr{E}\} \leq C(L_\pi + L_1)(M' + M)/\sqrt{n} = o(1),
$$

where L_1 depends on $(M' + \sqrt{n}hR)$.

Hence, if $h = o(1/\sqrt{n})$, then for any small constant ε, we have

$$
\mathbb{E}_{\theta_0} \left\| \tilde{Q}_T - F(\theta_0, \cdot) \right\|_{\mathsf{TV}} \leq \varepsilon,
$$

hold for n sufficiently large. □

Lemma 4.6.1. *Let $X_n(\theta) = \sqrt{n}(\widehat{\theta}_X - \theta)$ for $X \sim P^n_\theta$ for simplicity. Suppose there exists some sufficiently small constant $\tilde{\delta}$, such that for all $\theta \in B(\theta_0, \tilde{\delta})$, $X_n(\theta) = X(\theta) + W_n(\theta)$, where $X(\theta) \sim F(\theta, \cdot)$ and $W_n(\theta) = o_P(1)$. If $\sup_{\theta \in B(\theta_0, \tilde{\delta})} \mathbb{E}\|W_n(\theta)\| = o(1)$, and Assumption 4.2.3 and Assumption 4.2.4 hold, then Assumption 4.2.2' holds.*

Proof of Lemma 4.2.1. By definition, we have $f_n(\theta, t) = \int_w f(\theta, t-w) p_{\theta,n}(w) dw$, where $p_{\theta,n}(w)$ is the density for $W_n(\theta)$ depending on θ and n. For any large constant M, define $B(0, M) = \{t : \|t\| \leq M\}$. For all $\theta \in B(\theta_0, \tilde{\delta})$ and all $t \in B(0, M)$, there exists some corresponding small

constant η such that

$$\begin{aligned}\left|\frac{f_n(\theta,t)}{f(\theta,t)}-1\right| &= \left|\int_w \left(\frac{f(\theta,t-w)}{f(\theta,t)}-1\right) p(w)dw\right| \\ &\leq \int_w \left|\frac{f(\theta,t-w)}{f(\theta,t)}-1\right| \mathbb{I}\{\|w\|\leq\eta\}\, p(w)dw + \int_w \left|\frac{f(\theta,t-w)}{f(\theta,t)}-1\right| \mathbb{I}\{\|w\|>\eta\}\, p(w)dw, \\ &= (A)+(B),\end{aligned}$$

where the first term can be bounded by

$$\begin{aligned}(A) &\leq \int_w \left|e^{L_2\|w\|}-1\right| \mathbb{I}\{\|w\|\leq\eta\}\, p(w)dw && \text{(By Assumption 4.2.3} \\ &\leq \int_w \frac{e^{L_2\eta}-1}{\eta}\|w\|\mathbb{I}\{\|w\|\leq\eta\}\, p(w)dw \\ &\leq \frac{e^{L_2\eta}-1}{\eta}\mathbb{E}\|W_n(\theta)\|,\end{aligned}$$

and the second term is bounded by

$$\begin{aligned}(B) &\leq \left(\frac{\sup_t f(\theta,t)}{\inf_{\|t\|\leq M} f(\theta,t)}-1\right)\mathbb{P}\{\|W_n(\theta)\|\geq\eta\} \\ &\leq \left(\frac{\sup_t f(\theta,t)e^{L_2M}}{f(\theta,0)}-1\right)\mathbb{P}\{\|W_n(\theta)\|\geq\eta\}\end{aligned}$$

The constant L_2 depends on $M+\eta$. Let $n\to\infty$ and the result directly follows. □

CHAPTER 5

DISCUSSION

In this thesis we proposed and studied the sampling algorithm that works in the area of Bayesian community detection, Bayesian sparse linear regression, and Bayesian generative model. We now summarize the results and discuss potential future research directions.

In Chapter 2 we discussed the Metropolis-Hasting algorithm to sample discrete community label from posterior distribution. We provided posterior strong consistency result under the minimal signal-to-noise ratio condition in the literature and showed that the algorithm converges to stationary distribution polynomially fast. This is the first convergence analysis in Bayesian community detection area. It remains interesting to study the convergence without artificial temperature parameter in Algorithm 1.

Chapter 3 proposed TDLMC algorithm to sample from some non-convex and even unbounded distribution. We provided non-asymptotic convergence analysis indicating that the algorithm converges to stationary distribution polynomially fast in some cases. We also compared our algorithm with others in [28, 15], and showed our algorithm outperformed their algorithms in theoretical and practical aspects. It might be interesting to study the theoretical convergence of TDLMC algorithm sampling from some non-convex distribution, when the distribution contracts to small area. Possible applications to high dimensional Bayesian sparse regression are also interesting.

In Chapter 4 considered a special case where a complex loss function is given, and we aims to add prior information. The Bayesian distribution proposed in Chapter 4 is natural and yields good asymptotic properties. Though itself is intractable, we provided ABC algorithm and showed the approximate Bayesian distribution induced by ABC algorithm also shares asymptotic contraction and Bernstein-von Mises type of properties. In most cases, ABC algorithm is not efficient to run optimization for all synthetic data, and hence we provided pseudo ABC algorithm. It remains interesting to accelerate the pseudo ABC algorithm such that more samples are accepted during the process.

APPENDIX A

PROOFS

A.1 Proofs in Chapter 2

A.1.1 Proof of Lemma 2.4.4

For any state $Z \in S_\alpha$, we define

$$\begin{aligned}
&\mathscr{N}(Z) = \left\{Z' : H(Z', Z) = 1\right\}, \\
&\mathscr{A}(Z) = \left\{Z' \in \mathscr{N}(Z) : d(Z', Z^*) = d(Z, Z^*) + 1\right\}, \\
&\mathscr{B}(Z) = \left\{Z' \in \mathscr{N}(Z) : d(Z', Z^*) = d(Z, Z^*) - 1\right\},
\end{aligned} \tag{1}$$

where $\mathscr{N}(Z)$ denotes the neighborhood states of Z with only one sample classified differently, and $\mathscr{A}(Z)$ (resp. $\mathscr{B}(Z)$) denotes the set of states with more mistakes (resp. fewer mistakes) in the neighbor. We further define

$$\begin{aligned}
p_m(Z) = P(Z, \mathscr{A}(Z)) &= \frac{1}{2(K-1)n} \sum_{Z' \in \mathscr{A}(Z) \cap S_\alpha} \min\left\{1, \frac{\widetilde{\Pi}_g(Z'|A)}{\widetilde{\Pi}_g(Z|A)}\right\}, \\
q_m(Z) = P(Z, \mathscr{B}(Z)) &= \frac{1}{2(K-1)n} \sum_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \min\left\{1, \frac{\widetilde{\Pi}_g(Z'|A)}{\widetilde{\Pi}_g(Z|A)}\right\},
\end{aligned} \tag{2}$$

where $p_m(Z), q_m(Z)$ are the probabilities of Z jumping to states with the number of mistakes equal to $m+1$, $m-1$ respectively. We have the following lemma to bound the ratio of $p_m(Z)$ and $q_m(Z)$ for any $Z \in \mathscr{G}$. Recall that $\mathscr{G} = \mathscr{G}(\gamma_0)$ is defined in (30).

Lemma A.1.1. *Suppose γ_0 satisfies Condition 2.2, 2.4, or 2.5. Let τ be any sufficiently small constant τ, and denote $\mathscr{G}^* = \{Z : n^{-\tau} \le \ell(Z, Z^*) \le \max\{\gamma_0, n^{-\tau}\} + (\log n)^2/n\}$. Then, we have*

$$\mathbb{P}\left\{\max_{Z \in \mathscr{G}^*} \frac{p_m(Z)}{q_m(Z)} \ge \eta\right\} \le \exp\left(-n^{1-\tau}\right)$$

for some small constant $4\tau < \eta < 1$. *The probability is with respect to the data-generating process, i.e., the adjacency matrix A.*

The proof of Lemma A.1.1 is deferred to Section A.1.6. We take the τ in Lemma A.1.1 to be the same as τ defined in $\mathscr{G}$. In order to show that the Markov chain will stay in $\mathscr{G}$ with high probability, we transform the original problem into an one dimensional random walk problem. Lemma A.1.1 shows that the probability ratio of the one dimensional random walk on the region $\mathscr{G}^*$ can be bounded with high probability. All the following analysis is conditioning on the adjacency matrix A such that the event

$$\mathscr{E}(A) = \left\{ \max_{Z \in \mathscr{G}^*} \frac{p_m(Z)}{q_m(Z)} \leq \eta \right\}$$

happens. We construct the following three types of Markov chains in order to prove Lemma 2.4.4.

Type I MarKov chain. Consider a particle starting at the initial position u on the x-axis where $0 < u < b$ at time $t = 0$, and it moves one unit to the left, to the right, or stay at the current position at time $t = 1, 2, \ldots$ with probability q_t, p_t, or $1 - q_t - p_t$, where $\eta > p_t/q_t$ for all time t. It stops once it reaches the left or the right boundary, and we are interested in the probability of its stopping at the boundary b or the boundary 0.

Suppose the position of the particle at time t is X_t, and $X_{t+1} = X_t + \xi_t$, where ξ_t follows the distribution

$$\mathbb{P}\{\xi_t = 1\} = p_t, \ \ \mathbb{P}\{\xi_t = -1\} = q_t, \ \ \mathbb{P}\{\xi_t = 0\} = 1 - p_t - q_t.$$

We define $Y_t = \exp((b - X_t)\log\eta)$. It is easy to verify that the stochastic process Y_t is a super-

martingale, due to the fact that

$$\begin{aligned}\mathbb{E}(Y_{t+1}|Y_t) &= \exp((b-X_t)\log\eta)\cdot\mathbb{E}(\exp(-\xi_t\log\eta)) \\ &= \exp((b-X_t)\log\eta)\cdot(1-p_t-q_t+p_t/\eta+q_t\cdot\eta) \\ &\leq Y_t.\end{aligned}$$

Let $\tau = \min\{t \geq 1 : X_t = b \text{ or } X_t = 0\}$, and τ is *stopping time* of this random walk. It is evident that $|Y_{t\wedge\tau}| \leq 1$ since $X_\tau < b$ for all time t. By Doob's optional stopping time theorem, it follows that $\mathbb{E}(Y_\tau) \leq \mathbb{E}(Y_0)$, i.e.,

$$\mathbb{P}\{X_\tau = b\} \leq \mathbb{P}\{X_\tau = 0\}\cdot\eta^b + \mathbb{P}\{X_\tau = b\} \leq \eta^{b-u}, \tag{3}$$

where the first inequality holds since $\mathbb{P}\{X_\tau = 0\} \geq 0$, and the second inequality is due to Doob's optional stopping time theorem. By (3), the probability of the particle reaching boundary b first is upper bounded by η^{b-u}, where u is the starting position. Let $P_u^b = \mathbb{P}\{X_0 = u, X_\tau = b\}$ for $u \in (0,b)$ denote the probability of starting at u and stopping at b. Then, we have $P_1^b \leq \eta^{b-1}$.

Now suppose a particle starts at 0, i.e., $X_0 = 0$. It moves to the right or stay at the current position with some fixed probability p_0 or $1-p_0$. Let $P_0^0 = \mathbb{P}\{X_0 = 0,\, X_\tau = 0\}$, and $P_0^b = \mathbb{P}\{X_0 = 0,\, X_\tau = b\}$, where τ is the stopping time as defined before. Then, we have that

$$P_0^0 = p_0\cdot P_1^0 + (1-p_0), \quad P_0^b = p_0\cdot P_1^b, \quad P_0^0 + P_0^b = 1. \tag{4}$$

Type II Markov chain. We now define another Markov chain that is similar to the previous one. Consider a particle starting at position 0 at time 0, and follows the same updating rule as the previous chain but different stopping rule. We use W_t to denote the position of the particle at time t. The particle will only stop when it reaches the boundary b. When it is at the position 0, it still moves to the right or stay at 0 with fixed probability p_0 or $1-p_0$ (the same probability as defined

in Type I Markov chain). Thus, this newly defined Markov chain is a *reflected random walk*.

It is worth noting that the Type II Markov chain can always be decomposed into several Type I Markov chains. We use τ_W to denote the stopping time of Type II Markov chain, defined by $\tau_W = \min\{t \geq 1 : W_t = b\}$.

Type III Markov chain. Now return to our original problem and construct Type III Markov chain. Let $m_0 = n\ell(Z_0, Z^*)$, and we use $H_t = n\ell(Z_t, Z^*) - m_0$ to denote the position of the particle, where Z_t is the label assignment after t steps. The state space is all integers between $-m_0$ and b, where we take $b = \max\{0, n^{1-\tau} - m_0\} + \log^2 n$. When $H_t \in (0, b)$, the particle moves to the left, to the right, or stay at the current position with probability q_t, p_t, or $1 - q_t - p_t$, which is the same as the Type II chains. The particle will only stop when it reaches the boundary b. The stopping time of Type III Markov chain is defined by $\tau_H = \min\{t \geq 1 : H_t = b\}$.

Proof of Lemma 2.4.4. Recall that $b = \max\{0, n^{1-\tau} - m_0\} + \log^2 n$. In order to prove Lemma 2.4.4, it is equivalent to show that, for any T that is a polynomial of n, the event $\{H_t < b,\ t \leq T\}$ happens with high probability, i.e., $\{\tau_H > T\}$ happens with high probability. By the definition of Type II and III Markov chains we have that

$$\mathbb{P}\{\tau_H \leq T\} \leq \mathbb{P}\{\tau_W \leq T\}.$$

The above inequality holds since $H_0 = W_0$, $H_t \geq 0$ for all time t, and the updating rule of H_t and W_t are exactly the same when $W_t, H_t \in (0, b)$.

We now connect the Type II Markov chain with multiple Type I chains. The event $\{\tau_W \leq T\}$ means that the particle starts at 0 and reaches the boundary b within T steps, and it can be written as

$$\{\tau_W \leq T\} = \bigcup_{k=1}^{T} \left\{ \left[\bigcap_{i=1}^{k-1} \left\{ X_0^{(i)} = 0,\ X_{\tau_i}^{(i)} = 0 \right\} \right] \bigcap \left\{ X_0^{(k)} = 0,\ X_{\tau_i}^{(k)} = b \right\} \bigcap \left\{ \sum_{i=1}^{k} \tau_i \leq T \right\} \right\}, \tag{5}$$

where we use $X^{(i)}$ to denote the ith Type I Markov chain, and τ_i is the stopping time of $X^{(i)}$. Note

$X^{(i)}$ is independent with $X^{(j)}$ for $i \neq j$. The right hand side of (5) can be interpreted as that the particle reaches the boundary 0 for $k-1$ times with $k \leq T$ before reaching the boundary b, and the total number of steps is less than T. Therefore, it follows directly that

$$\begin{aligned}
\mathbb{P}\{\tau_W \leq T\} &= \mathbb{P}\left\{ \bigcup_{k=1}^{T} \left\{ \left[\bigcap_{i=1}^{k-1} \left\{ X_0^{(i)} = 0,\, X_{\tau_i}^{(i)} = 0 \right\} \right] \bigcap \left\{ X_0^{(k)} = 0,\, X_{\tau_i}^{(k)} = b \right\} \bigcap \left\{ \sum_{i=1}^{k} \tau_i \leq T \right\} \right\} \right\} \\
&\leq \sum_{k=1}^{T} \mathbb{P}\left\{ \left[\bigcap_{i=1}^{k-1} \left\{ X_0^{(i)} = 0,\, X_{\tau_i}^{(i)} = 0 \right\} \right] \bigcap \left\{ X_0^{(k)} = 0,\, X_{\tau_i}^{(k)} = b \right\} \bigcap \left\{ \sum_{i=1}^{k} \tau_i \leq T \right\} \right\} \\
&\leq \sum_{k=1}^{T} \mathbb{P}\left\{ \left[\bigcap_{i=1}^{k-1} \left\{ X_0^{(i)} = 0,\, X_{\tau_i}^{(i)} = 0 \right\} \right] \bigcap \left\{ X_0^{(k)} = 0,\, X_{\tau_i}^{(k)} = b \right\} \right\} \\
&= \sum_{k=1}^{T} (P_0^0)^{k-1} P_0^b.
\end{aligned}$$

The first inequality holds by a union bound. The third inequality is by the independence. By (4), we have that

$$\begin{aligned}
\sum_{k=1}^{T} P_0^b (P_0^0)^{k-1} &= 1 - (P_0^0)^T \\
&\leq -T \log P_0^0 \leq T \cdot \frac{1 - P_0^0}{P_0^0} = T \cdot \frac{p_0 \cdot P_1^b}{p_0 \cdot P_1^0 + 1 - p_0} \\
&\leq T \cdot \frac{P_1^b}{P_1^0} \leq \exp\left(\log \frac{\eta^{b-1}}{1 - \eta^{b-1}} + \log T \right).
\end{aligned}$$

Since T is a polynomial of n, and $b \gg \log n$, then it follows that

$$\mathbb{P}\{\tau_H \leq T\} \leq \mathbb{P}\{\tau_W \leq T\} \leq \exp\left(-(1 - o(1)) b \log \frac{1}{\eta} \right).$$

Thus, based on the result of Lemma A.1.1, for any given initial label assignment Z_0 with $\ell(Z_0, Z^*) \leq \gamma_0$ where γ_0 satisfies Condition 2.2, 2.4, or 2.5, we have that in any polynomial running time, the number of mistakes is upper bounded by

$$m \leq m_0 + b = \max\{m_0, n^{1-\tau}\} + \log^2 n \leq n \max\{\gamma_0, n^{-\tau}\} + \log^2 n,$$

with probability at least $1-\exp(-\log^2 n)$. □

A.1.2 Some preparations before the proofs of Lemma 2.4.2 and Lemma 2.4.6

In this section, we will define some events and introduce some quantities to simplify the main proof.

Basic events

For any $Z \in S_\alpha$, recall that $O_{ab}(Z) = \sum_{i,j} A_{ij} \mathbb{I}\{Z_i = a, Z_j = b\}$, and define $X_{ab}(Z) = O_{ab}(Z) - \mathbb{E}[O_{ab}(Z)]$ for any $a, b \in [K]$. Let $X(Z)$ denote a $K \times K$ matrix with its (a,b)th element equal to $X_{ab}(Z)$ for any $a, b \in [K]$. For any positive sequences $\bar{\varepsilon} = \bar{\varepsilon}_n$, $\gamma = \gamma_n$, and $\theta = \theta_n$ satisfying that $\bar{\varepsilon}, \gamma, \theta \to 0, \bar{\varepsilon}^2 nI \to \infty$, and $\theta^2 \gamma n I \to \infty$, consider the following events:

$$
\begin{aligned}
\mathscr{E}_1(\bar{\varepsilon}) &= \left\{ \max_{Z \in S_\alpha} \|X(Z)\|_\infty \le \bar{\varepsilon} n^2 (p-q) \right\}, \\
\mathscr{E}_2 &= \left\{ \max_{Z \in S_\alpha} \|X(Z) - X(Z^*)\|_\infty - (\alpha+\beta) mn(p-q)/K \le 0 \right\}, \\
\mathscr{E}_3(\bar{\varepsilon}) &= \left\{ \max_{Z \in S_\alpha} \|X(Z) - X(Z^*)\|_\infty \le \bar{\varepsilon} n^2 (p-q) \right\}, \\
\mathscr{E}_4(\gamma, \theta) &= \left\{ \max_{Z \in S_\alpha : m > \gamma n} \|X(Z) - X(Z^*)\|_\infty \le \theta mn(p-q) \right\}, \\
\mathscr{E}_5 &= \left\{ \max_{Z \in S_\alpha} \frac{1}{|\mathscr{B}(Z) \cap S_\alpha|} \sum_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \sum_{a \le a'} \left| X_{aa'}(Z) - X_{aa'}(Z') \right| \le 10n(p-q) \right\}, \\
\mathscr{E}_6(\gamma, \theta) &= \left\{ \max_{Z \in S_\alpha : m > \gamma n} \frac{1}{|\mathscr{B}(Z) \cap S_\alpha|} \sum_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \sum_{a \le a'} \left| X_{aa'}(Z) - X_{aa'}(Z') \right| \le \theta n(p-q) \right\},
\end{aligned}
\tag{6}
$$

where $\mathscr{B}(Z)$ is defined in (1). We may also use $\mathscr{E}_1$ to denote $\mathscr{E}_1(\bar{\varepsilon})$ for simplicity, and such simplification also applies to any other event. Denote

$$
\mathscr{E} = \mathscr{E}(\bar{\varepsilon}, \gamma, \theta) = \mathscr{E}_1 \cap \mathscr{E}_2 \cap \mathscr{E}_3 \cap \mathscr{E}_4 \cap \mathscr{E}_5 \cap \mathscr{E}_6. \tag{7}
$$

By Lemmas A.1.8, A.1.10, A.1.11, A.1.12, A.1.13, and A.1.14, it follows that for any $\bar{\varepsilon}, \gamma, \theta$ satisfying the conditions,

$$\mathbb{P}\{\mathscr{E}(\bar{\varepsilon}, \gamma, \theta)\} \geq 1 - n\exp(-(1-o(1))\bar{n}I).$$

Likelihood modularity

The posterior distribution is hard to deal with directly. Hence, we first analyze the performance of the likelihood modularity function, and then bound the difference between the likelihood modularity function and the posterior distribution to simplify the proof.

Likelihood modularity is first introduced in [11], which takes the form as

$$Q_{LM}(Z,A) = \sum_{a \leq b} n_{ab}(Z)\tau\left(\frac{O_{ab}(Z)}{n_{ab}(Z)}\right), \tag{8}$$

where $\tau(x) = x\log(x) + (1-x)\log(1-x)$. This criterion replaces the connectivity probabilities by maximum likelihood estimates. Instead of comparing the direct difference between the Bayesian expression and the likelihood modularity as in [79], we have the following lemma to bound the relative difference.

Lemma A.1.2. *Under the event $\mathscr{E}(\bar{\varepsilon}, \gamma, \theta)$ defined in* (7)*, we have*

$$\max_{Z \in S_\alpha} \left| \log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} - (Q_{LM}(Z,A) - Q_{LM}(Z^*,A)) \right| \leq C_{LM}$$

for some constant C_{LM} only depending on K, α, β.

The above lemma is rephrased and proved in Lemma A.1.19.

Discrepancy matrix

For any label assignment $Z \in S_\alpha$, let R_Z be a *discrepancy matrix*, which takes the form of

$$R_Z(a,b) = \sum_{i=1}^{n} \mathbb{I}\left\{Z_i = a, Z_j^* = b\right\}, \quad a,b \in [K], \tag{9}$$

where $R_Z(a,b)$ is the number of samples misclassified to group a but actually from group b based on the true label assignment. Note that the true label assignment is only unique up to a label permutation, and thus we always permute the rows of R_Z to minimize the off diagonal sum. Later we write $R_Z(k,l)$ as R_{kl} for simplicity.

Using the discrepancy matrix R_Z, we have

$$\begin{aligned}
\mathbb{E}\left[O_{ab}(Z)\right] &= \mathbb{E}\left[\sum_{i,j} A_{ij}\mathbb{I}\left\{Z_i = a, Z_j = b\right\}\right] = (RBR^T)_{ab}, \quad \text{for } a \neq b \in [K], \\
\mathbb{E}\left[O_{aa}(Z)\right] &= \mathbb{E}\left[\sum_{i<j} A_{ij}\mathbb{I}\left\{Z_i = Z_j = a\right\}\right] = \frac{1}{2}\left((RBR^T)_{aa} - \sum_k B_{kk}R_{ak}\right), \quad \text{for } a \in [K].
\end{aligned} \tag{10}$$

A.1.3 Proof of Lemma 2.4.2

Before proving the lemma, we need to present some notations that will be frequently used:

$$\begin{aligned}
O_s(Z) &= \sum_{a\in[K]} O_{aa}(Z) = \sum_{i<j} A_{ij}\mathbb{I}\left\{Z_i = Z_j\right\}, \\
n_s(Z) &= \sum_{a\in[K]} n_{aa}(Z) = \sum_{i<j} \mathbb{I}\left\{Z_i = Z_j\right\}, \\
\Delta\widetilde{O}_{ab} &= O_{ab}(Z) - O_{ab}(Z^*), \;\; \Delta\widetilde{O}_s = \sum_{a\in[K]} \Delta\widetilde{O}_{aa}, \\
\Delta\widetilde{n}_{ab} &= n_{ab}(Z) - n_{ab}(Z^*), \;\; \Delta\widetilde{n}_s = \sum_{a\in[K]} \Delta\widetilde{n}_{aa},
\end{aligned}$$

and we may write $n_{ab}(Z^*)$, $O_{ab}(Z^*)$ as n_{ab}, O_{ab} for simplicity.

Proof of Lemma 2.4.2. For any positive sequences $\gamma = \gamma_n$ and $\theta = \theta_n$ such that $\gamma \to 0$, $\gamma^2 nI \to \infty$,

and $\theta^2\gamma nI \to \infty$, we can construct the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$ defined in (7) by setting $\bar{\varepsilon}=\gamma$, and perform analysis on a large mistake region and a small mistake region separately.

Small mistake region. For $m \le \gamma n$, by some calculations, it follows that

$$\begin{aligned}
&Q_{LM}(Z,A) - Q_{LM}(Z^*,A) \\
=&\log\Pi_0(Z|A) - \log\Pi_0(Z^*|A) - \sum_{a\le b} n_{ab}\cdot D\left(\frac{O_{ab}}{n_{ab}}\middle\|\frac{O_{ab}(Z)}{n_{ab}(Z)}\right) + \\
&\underbrace{\sum_{a\le b}\Delta\widetilde{O}_{ab}\left(\log\frac{O_{ab}(Z)}{n_{ab}(Z)} - \log B_{ab}\right) + (\Delta\widetilde{n}_{ab} - \Delta\widetilde{O}_{ab})\left(\log\frac{n_{ab}(Z)-O_{ab}(Z)}{n_{ab}(Z)} - \log(1-B_{ab})\right)}_{(Error)},
\end{aligned} \tag{11}$$

where by (24)

$$\log\Pi_0(Z|A) - \log\Pi_0(Z^*|A) = 2t^*\left(\Delta\widetilde{O}_s - \lambda^*\Delta\widetilde{n}_s\right), \tag{12}$$

and

$$\lambda^* = \log\frac{1-q}{1-p}\Big/\log\frac{p(1-q)}{q(1-p)}, \quad t^* = \frac{1}{2}\log\frac{p(1-q)}{q(1-p)}. \tag{13}$$

Under the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$, by Lemma A.1.22, we have $(Error) \le C\gamma mnI$ for some constant C. Hence, for any fixed $Z \in S_\alpha$, by Lemma A.1.2, we have that

$$\begin{aligned}
&\log\frac{\Pi(Z|A)}{\Pi(Z^*|A)} - \log\frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} \\
\le& Q_{LM}(Z,A) - Q_{LM}(Z^*,A) + C_{LM} - \log\frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} \\
\le& C\gamma mnI + C_{LM}.
\end{aligned}$$

Thus, there exists some constant C_2 such that under the event $\mathscr{E}$,

$$\max_{Z\in S_\alpha : m<\gamma n} \log\frac{\Pi(Z|A)}{\Pi(Z^*|A)} - \log\frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} - C_2\gamma mnI \le 0,$$

which proves the second statement in Lemma 2.4.2.

Large mistake region. For $m > \gamma n$, we have that

$$\begin{aligned} Q_{LM}(Z,A) - Q_{LM}(Z^*,A) &= \sum_{a\le b} n_{ab}(Z)\tau\left(\frac{O_{ab}(Z)}{n_{ab}(Z)}\right) - n_{ab}(Z^*)\tau\left(\frac{O_{ab}(Z^*)}{n_{ab}(Z^*)}\right) \\ &= (G(Z)+\Delta(Z)) - (G(Z^*)+\Delta(Z^*))\,, \end{aligned} \tag{14}$$

where we write

$$\begin{aligned} G(\cdot) &= \sum_{a\le b} n_{ab}(\cdot)\tau\left(\frac{\mathbb{E}\left[O_{ab}(\cdot)\right]}{n_{ab}(\cdot)}\right), \\ \Delta(\cdot) &= \sum_{a\le b} n_{ab}(\cdot)\left(\tau\left(\frac{O_{ab}(\cdot)}{n_{ab}(\cdot)}\right) - \tau\left(\frac{\mathbb{E}\left[O_{ab}(\cdot)\right]}{n_{ab}(\cdot)}\right)\right). \end{aligned}$$

Let $\widetilde{B}_{ab} = \mathbb{E}\left[O_{ab}(Z)\right]/n_{ab}(Z)$ for any $a,b \in [K]$. By (10), it follows that

$$\begin{aligned} 2G(Z) =& 2\sum_{a\le b} n_{ab}(Z)\tau\left(\widetilde{B}_{ab}\right) \\ =& 2\sum_{a\le b} \mathbb{E}\left[O_{ab}(Z)\right]\log\widetilde{B}_{ab} + (n_{ab}(Z) - \mathbb{E}\left[O_{ab}(Z)\right])\log(1-\widetilde{B}_{ab}) \\ =& \sum_{a,b,k,l} R_{ak}R_{bl}(B_{kl}\log\widetilde{B}_{ab} + (1-B_{kl})\log(1-\widetilde{B}_{ab})) - \sum_a n_a(Z)(p\log\widetilde{B}_{aa} + (1-p)\log(1-\widetilde{B}_{aa})). \end{aligned}$$

Then, we have that

$$\begin{aligned} & 2G(Z) - 2G(Z^*) \\ =& \sum_{a,b,k,l} R_{ak}R_{bl}(B_{kl}\log\widetilde{B}_{ab} + (1-B_{kl})\log(1-\widetilde{B}_{ab})) - \sum_{a,b,k,l} R_{ak}R_{bl}(B_{kl}\log B_{kl} + (1-B_{kl})\log(1-B_{kl})) \\ -& \sum_a n_a(Z)\left((p\log\widetilde{B}_{aa} + (1-p)\log(1-\widetilde{B}_{aa})) - (p\log p + (1-p)\log(1-p))\right) \\ =& -\sum_{a,b,k,l} R_{ak}R_{bl}D\left(B_{kl}\middle\|\widetilde{B}_{ab}\right) + \sum_a n_a(Z)D\left(p\middle\|\widetilde{B}_{aa}\right). \end{aligned} \tag{15}$$

By Lemma A.1.23 and Lemma A.1.25 that bound the above two terms separately, we have

$$2G(Z) - 2G(Z^*) \leq -CmnI,$$

for some constant C. Under the events $\mathscr{E}_1(\bar{\varepsilon}), \mathscr{E}_3(\bar{\varepsilon}), \mathscr{E}_4(\gamma,\theta)$ in (6), by Lemma A.1.19 and Lemma A.1.26, we have that

$$\max_{Z\in S_\alpha} \left| \log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} - (Q_{LM}(Z,A) - Q_{LM}(Z^*,A)) \right| \leq C_{LM},$$

and

$$\max_{Z\in S_\alpha : m>\gamma n} |\Delta(Z) - \Delta(Z^*)| \leq \varepsilon mnI,$$

for some $\varepsilon \to 0$. Hence, it follows that there exists some constant C_1 such that

$$\mathbb{P}\left\{ \max_{Z\in S_\alpha : m>\gamma n} \log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} > -C_1 mnI \right\} \leq 4\exp(-n).$$

Combining the result of two regions directly gives Lemma 2.4.2.

□

A.1.4 Proof of Lemma 2.4.6 with known connectivity probabilities

In order to distinguish from the case where probabilities are unknown, we use $\Pi_0(\cdot|A)$ to denote the posterior distribution in this case, and define $\widetilde{\Pi}_0(\cdot|A)$ as the scaled distribution proportional to $\Pi_0^{\xi}(\cdot|A)$. It suffices to prove the following lemma.

Lemma A.1.3. *Recall that $g(Z)$ is the next state of Z. Suppose γ_0 satisfies Condition 2.2, 2.4, or 2.5. Then, there exists some positive sequence $\gamma \to 0$ such that, with probability at least $1 - C_1 n^{-C_2}$,*

$$\frac{\Pi_0(Z|A)}{\Pi_0(g(Z)|A)} \leq \begin{cases} \exp\left(-\varepsilon \bar{n} I\right), & \text{if } m \leq \gamma n, \\ \exp\left(-4\left(1/K\alpha - \gamma_0\right) nI(1-o(1))\right), & \text{if } m > \gamma n, \end{cases}$$

holds uniformly for all $Z \in \mathscr{G}(\gamma_0)$ defined in (30)*. Here, ε is any constant satisfying $\varepsilon < 2\varepsilon_0$ with ε_0 defined in Condition 2.5, and C_1, C_2 are two constants depending on ε. Furthermore, if ξ satisfies Condition 2.5, then by choosing $\varepsilon \in ((1-\varepsilon_0)/\xi, 2\varepsilon_0)$, we have*

$$\max_{Z \in \mathscr{G}(\gamma_0)} \frac{\widetilde{\Pi}_0(Z|A)}{\widetilde{\Pi}_0(g(Z)|A)} \leq \exp(-C\bar{n}I)$$

for some constant $C > 1-\varepsilon_0$ with probability at least $1 - C_3 n^{-C_4}$.

Proof of Lemma A.1.3. Recall the definitions of $\mathscr{A}(Z)$, $\mathscr{B}(Z)$, and $\mathscr{N}(Z)$ in (1). We introduce some notations first to simplify the proof. For any $a, b \in [K]$ and any two label assignments Z, Z', we write

$$\begin{aligned}
&\Delta O_{ab} = O_{ab}(Z) - O_{ab}(Z'), && \Delta O_s = \sum_a \Delta O_{aa}, \\
&\Delta n_{ab} = n_{ab}(Z) - n_{ab}(Z'), && \Delta n_s = \sum_a \Delta n_{aa}, \\
&\Delta_n(Z, Z') = \Delta O_s - \lambda^* \Delta n_s, && \lambda^* = \log \frac{1-q}{1-p} \Big/ \log \frac{p(1-q)}{q(1-p)}.
\end{aligned}$$

Suppose the current state is Z, and we randomly choose one misclassified sample from group a and move to its true group b. Denote the new state as Z'. It follows that $R_{Z'}(a,b) = R_Z(a,b) - 1$, and $R_{Z'}(b,b) = R_Z(b,b) + 1$. Write $R_Z(a,b)$ as R_{ab} for simplicity. Furthermore, let $\{x_l\}_{l \geq 1}, \{\widetilde{x}_l\}_{l \geq 1}$ be i.i.d. copies of Bernoulli(q) and $\{y_l\}_{l \geq 1}, \{\widetilde{y}_l\}_{l \geq 1}$ be i.i.d. copies of Bernoulli(p). We have that

$$\Delta_n(Z, Z') = \sum_{l=1}^{R_{aa} + \sum_{k \neq a,b} R_{ak}} (x_l - \lambda^*) - \sum_{l=1}^{R_{bb}} (y_l - \lambda^*) + \sum_{l=1}^{R_{ab}-1} (\widetilde{y}_l - \lambda^*) - \sum_{l=1}^{\sum_{k \neq b} R_{bk}} (\widetilde{x}_l - \lambda^*). \quad (16)$$

By (24), it directly follows that

$$\log \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)} = 2t^* \Delta_n(Z, Z'), \;\; t^* = \frac{1}{2} \log \frac{p(1-q)}{q(1-p)}. \quad (17)$$

For some positive sequence $\gamma = \gamma_n \to 0$ to be specified later, we can again divide S_α into two

regions.

Small mistake region. For $m \leq \gamma n$, we have

$$\begin{aligned}
&\mathbb{E}\left[\sqrt{\frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)}}\right] = \mathbb{E}\left[e^{t^*\Delta_n(Z,Z')}\right] \\
=&\mathbb{E}\exp\left[t^*\left(\sum_{l=1}^{R_{aa}}(x_l-\lambda^*)-\sum_{l=1}^{R_{bb}}(y_l-\lambda^*)\right)\right] \qquad (18)\\
\cdot&\mathbb{E}\exp\left[t^*\left(\sum_{l=1}^{R_{ab}-1}(\widetilde{y}_l-\lambda^*)-\sum_{l=1}^{\Sigma_{k\neq b}R_{bk}}(\widetilde{x}_l-\lambda^*)+\sum_{l=1}^{\Sigma_{k\neq a,b}R_{ak}}(x_l-\lambda^*)\right)\right]. \qquad (19)
\end{aligned}$$

Based on Lemma 6.1 in [89], for any positive integers n_1, n_2, we have

$$\mathbb{E}\exp\left[t^*\left(\sum_{i=1}^{n_1}(x_i-\lambda^*)-\sum_{i=1}^{n_2}(y_i-\lambda^*)\right)\right] = \exp\left(-\frac{(n_1+n_2)I}{2}\right),$$

and it leads to

$$(18) = \exp\left(-\frac{(R_{aa}+R_{bb})I}{2}\right) \leq \exp\left(-\frac{(n_a+n_b-m)I}{2}\right) \leq \exp\left(-(1-c\gamma)\bar{n}I\right),$$

for some constant c. By Lemma A.1.17, we have

$$(19) \leq (\exp(C_0 I))^m = \exp(C_0 m I) \leq \exp(C_0\gamma n I),$$

for some constant C_0. It follows that

$$\mathbb{E}\left[\sqrt{\frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)}}\right] \leq \exp\left(-(1-C_1\gamma)\bar{n}I\right) \qquad (20)$$

for some constant C_1 depending on C_0. Let ε be any small constant satisfying $\varepsilon < 2\varepsilon_0$, and

$w = \varepsilon\bar{n}I/2t^*$. By Lemma A.1.27, since $\gamma < c_{\alpha,\beta}$, we have $\mathscr{B}(Z) \subset S_\alpha$. Then,

$$\begin{aligned}
\mathbb{P}\left\{\min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \Delta_n(Z,Z') \geq -w\right\} &\leq \mathbb{P}\left\{\sum_{Z'\in\mathscr{B}(Z)} \Delta_n(Z,Z') \geq -mw\right\} \\
&= \mathbb{P}\left\{\exp\left(t^* \sum_{Z'\in\mathscr{B}(Z)} \Delta_n(Z,Z')\right) \geq \exp\left(-t^* mw\right)\right\} \\
&\leq \mathbb{E}\left[\exp\left(t^* \sum_{Z'\in\mathscr{B}(Z)} \Delta_n(Z,Z')\right) \cdot \exp\left(t^* mw\right)\right] \\
&= \mathbb{E}\left[\exp\left(t^* \sum_{Z'\in\mathscr{B}(Z)} \Delta_n(Z,Z')\right) \cdot \exp\left(\varepsilon m\bar{n}I/2\right)\right], \qquad (21)
\end{aligned}$$

The first inequality holds because the minimum is smaller than the average. The second inequality is due to Markov's inequality. We now proceed to bound (21) by $\exp(-(1-o(1))m\bar{n}I)$.

We first define set $\mathscr{C}(Z) = \{i : Z_i = Z_i^*\}$, which is the set of samples that are correctly classified. Thus, we have $|\mathscr{C}(Z)| = \sum_{a\in[K]} R_{aa} = n - m$. Suppose $Z' \in \mathscr{B}(Z)$ corrects kth sample from a misclassified group a, where $Z_k = a$, to its true group b, where $Z'_k = b$. Then, we must have $k \in [n] \setminus \mathscr{C}(Z)$, and by Lemma A.1.28, we can rewrite

$$\begin{aligned}
\Delta_n(Z,Z') &= \sum_{i\in[n]} \left(A_{ik}\mathbb{I}\left\{Z'_i = Z_k\right\} - \lambda^*\right) - \sum_{i\in[n]} \left(A_{ik}\mathbb{I}\left\{Z_i = Z'_k\right\} - \lambda^*\right) \\
&= \underbrace{\sum_{i\in\mathscr{C}(Z)} \left(A_{ik}\mathbb{I}\left\{Z'_i = Z_k\right\} - \lambda^*\right) - \sum_{i\in\mathscr{C}(Z)} \left(A_{ik}\mathbb{I}\left\{Z_i = Z'_k\right\} - \lambda^*\right)}_{(A_k)} \\
&\quad + \underbrace{\sum_{i\notin\mathscr{C}(Z)} \left(A_{ik}\mathbb{I}\left\{Z'_i = Z_k\right\} - \lambda^*\right) - \sum_{i\notin\mathscr{C}(Z)} \left(A_{ik}\mathbb{I}\left\{Z_i = Z'_k\right\} - \lambda^*\right)}_{(B_k)}.
\end{aligned}$$

Here, A_k, B_k correspond to summations in (18) and (19) respectively. We further have

$$\sum_{Z'\in\mathscr{B}(Z)} \Delta_n(Z,Z') = \sum_{k\in[n]\setminus\mathscr{C}(Z)} \left(A_k + B_k\right).$$

It is obvious that $A_k \perp A_j$ for $k, j \in [n] \setminus \mathscr{C}(Z)$, and $\sum_{k\in[n]\setminus\mathscr{C}(Z)} A_k$ can be written as the independent sum of the random variable A_{ij} for some $i \in [n] \setminus \mathscr{C}(Z)$ and $j \in C(Z)$. As for $\sum_{k\in[n]\setminus\mathscr{C}(Z)} B_k$, it is the summation of A_{ij} for some $i, j \in [n] \setminus \mathscr{C}(Z)$. For each random variable A_{ij}, the coefficient is at most 2 (since it can only be added twice or canceled out), and the total number of random variables is at most $\binom{m}{2}$. Hence, by the argument from (18) to (20), we can bound (21) by

$$(21) \leq \exp\left(-\left(1 - C\gamma - \varepsilon/2\right) m\bar{n}I\right),$$

for some constant C. By the definition of $g(Z)$ defined in (32), we have

$$\begin{aligned}
\mathbb{P}\left\{\max_{Z\in\mathscr{G}(\gamma_0):m\leq\gamma n} \frac{\Pi_0(Z|A)}{\Pi_0(g(Z)|A)} \geq \exp(-\varepsilon\bar{n}I)\right\} &= \mathbb{P}\left\{\max_{Z\in\mathscr{G}(\gamma_0):m\leq\gamma n} \min_{Z'\in\mathscr{B}(Z)} \Delta_n(Z,Z') \geq -w\right\} \\
&\leq \sum_{m=1}^{\gamma n} \binom{n}{m} (K-1)^m \exp\left(-\left(1 - C\gamma - \varepsilon/2\right) m\bar{n}I\right) \\
&\leq \sum_{m=1}^{\gamma n} \left(enK \exp\left(-\left(1 - C\gamma - \varepsilon/2\right)\bar{n}I\right)\right)^m \\
&= n \exp\left(-\left(1 - \varepsilon/2\right)\bar{n}I(1 - o(1))\right),
\end{aligned} \tag{22}$$

where we require $\varepsilon < 2\varepsilon_0$ in order for the last equation going to 0 as n tends to infinity.

Large mistake region. For $m > \gamma n$, recall that $\mathscr{G}(\gamma_0) \subset S_\alpha$, and S_α is defined in (4). If Z' corrects one sample from group a to group b, by (16), we have $\Delta n_s = n'_a - n'_b - 1$. By 16, we have $\Delta_n(Z,Z') = \Delta O_s - \lambda^* \Delta n_s$. Let $\lambda = (p+q)/2$, $m_b = \sum_{k\neq b} R_{kb}$, $n'_a = n_a(Z)$, and $n'_b = n_b(Z)$ for

simplicity. Thus, it follows that

$$\begin{aligned}
\mathbb{E}\left[\Delta_n(Z,Z')\right] &= \mathbb{E}\left[\Delta O_s\right] - \lambda^* \Delta n_s = \mathbb{E}\left[\Delta O_s\right] - \lambda \Delta n_s + (\lambda - \lambda^*)\Delta n_s \\
&= -\frac{p-q}{2}(n'_a + n'_b - 2m_b - 2R_{ab}) + \frac{p+q-2\lambda^*}{2}(n'_a - n_b) - (p-\lambda^*) \\
&\leq -\frac{p-q}{2}(n'_a + n'_b - 2m) + \frac{p+q-2\lambda^*}{2}(n'_a - n'_b) \\
&= -\frac{p-q}{2}(n'_a + n'_b - 2m - C_\lambda(n'_a - n'_b)) \\
&= -\frac{p-q}{2}(n'_b + C_\lambda n'_b + (1-C_\lambda)n'_a - 2m) \\
&\leq -\left(\frac{n}{K\alpha} - m\right)(p-q),
\end{aligned}$$

where $C_\lambda = 2(\lambda - \lambda^*)/(p-q)$. It is easy to verify that $C_\lambda \in (0,1)$, and C_λ tends to 1 (resp. tends to 0) when $(p-q)/p$ tends to 1 (resp. tends to 0). If γ_0 satisfies that $(1-K\alpha\gamma_0)^2 nI \to \infty$, it follows that

$$\max_{Z \in \mathscr{G}(\gamma_0): m > \gamma n} \max_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \left(\mathbb{E}\left[\Delta O_s\right] - \lambda^* \Delta n_s\right) \leq -\left(\frac{1}{K\alpha} - \gamma_0\right) n(p-q)(1-o(1)). \tag{23}$$

Denote $\delta_0 = 1/K\alpha - \gamma_0$ for simplicity. Then, it follows that (23) $\leq -\delta_0 n(p-q)(1-o(1))$. Since $\delta_0^2 nI \to \infty$, there exist some positive sequences γ and θ such that $\gamma, \theta \to 0$, $\theta^2 \gamma nI \to \infty$, and $\theta \ll \delta_0$. To be specific, we may take $\gamma = 1/(\delta_0\sqrt{nI})$, $\theta = \delta/(\delta\sqrt{nI})^{1/4}$. Hence, we can construct the event $\mathscr{E}_6(\gamma,\theta)$ as defined in (6). Note that $X(Z) - X(Z^*) = \Delta O(Z) - \mathbb{E}\left[\Delta O(Z)\right]$. Under the event $\mathscr{E}_6$, we

have

$$
\begin{aligned}
&\max_{Z\in S_\alpha: m>\gamma n} \min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} |\Delta O_s - \mathbb{E}[\Delta O_s]| \\
\leq &\max_{Z\in S_\alpha: m>\gamma n} \frac{1}{|\mathscr{B}(Z)\cap S_\alpha|} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} |\Delta O_s - \mathbb{E}[\Delta O_s]| \\
\leq &\max_{Z\in S_\alpha: m>\gamma n} \frac{1}{|\mathscr{B}(Z)\cap S_\alpha|} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \sum_{a\in[K]} |\Delta O_{aa} - \mathbb{E}[\Delta O_{aa}]| \\
\leq &\max_{Z\in S_\alpha: m>\gamma n} \frac{1}{|\mathscr{B}(Z)\cap S_\alpha|} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \sum_{a\leq a'} \left|X_{aa'}(Z) - X_{aa'}(Z')\right| \\
\leq &\theta n(p-q).
\end{aligned}
$$

Then, it follows that

$$
\begin{aligned}
&\max_{Z\in\mathscr{G}(\gamma_0): m>\gamma n} \min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \Delta_n(Z,Z') \\
= &\max_{Z\in\mathscr{G}(\gamma_0): m>\gamma n} \min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} (\Delta O_s - \lambda^* \Delta n_s) \\
\leq &\max_{Z\in\mathscr{G}(\gamma_0): m>\gamma n} \min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} (\Delta O_s - \mathbb{E}[\Delta O_s]) + \max_{Z\in S_\alpha: m>\gamma n} \max_{Z'\in\mathscr{B}(Z)\cap S_\alpha} (\mathbb{E}[\Delta O_s] - \lambda^* \Delta n_s) \qquad (24) \\
\leq &\theta n(p-q) - \delta_0 n(p-q)(1-o(1)) \\
= &-\delta_0(1-o(1))n(p-q).
\end{aligned}
$$

By the definition of $g(Z)$ and by (17), we have that

$$
\begin{aligned}
\max_{Z\in\mathscr{G}(\gamma_0): m>\gamma n} \log \frac{\Pi_0(Z|A)}{\Pi_0(g(Z)|A)} &= \log \frac{p(1-q)}{q(1-p)} \left[\max_{Z\in\mathscr{G}(\gamma_0): m>\gamma n} \min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \Delta_n(Z,Z')\right] \\
&\leq -\log\left(\frac{p}{q}\right) \delta_0 n(p-q)(1-o(1)).
\end{aligned}
$$

Furthermore, since

$$
I = (\sqrt{p}-\sqrt{q})^2(1+o(1)), \quad \frac{(p-q)\log\left(\frac{p}{q}\right)}{\left(\sqrt{p}-\sqrt{q}\right)^2} \geq 4,
$$

we have

$$\log\left(\frac{p}{q}\right)(p-q) \geq 4(\sqrt{p}-\sqrt{q})^2 = 4I(1-o(1)).$$

Then, it follows directly

$$\max_{Z\in\mathscr{G}(\gamma_0):m>\gamma n} \frac{\Pi_0(Z|A)}{\Pi_0(g(Z)|A)} \leq \exp\left(-4\delta_0 nI(1-o(1))\right).$$

By Lemma A.1.14, we have that $\mathscr{E}_6$ happens with probability at least $1-\exp(-n)$.

Combining results of two regions directly gives the result of Lemma A.1.3. □

A.1.5 Proof of Lemma 2.4.6 with unknown connectivity probabilities

It suffices to prove the following lemma when the connectivity probabilities are unknown.

Lemma A.1.4. *Suppose γ_0 satisfies Condition 2.2, 2.4, or 2.5. Then, there exists some positive sequence $\gamma \to 0$ such that, with probability at least $1-C_1 n^{-C_2}$,*

$$\frac{\Pi(Z|A)}{\Pi(g(Z)|A)} \leq \begin{cases} \exp\left(-\varepsilon\bar{n}I(1-o(1))\right), & \text{if } m \leq \gamma n, \\ \exp\left(-\dfrac{(1-K\gamma_0)^4 nI}{2\alpha^2}(1-o(1))\right), & \text{if } m > \gamma n, \end{cases}$$

holds uniformly for all $Z \in \mathscr{G}(\gamma_0)$ defined in (30)*. Here, ε is any constant satisfying $\varepsilon < \varepsilon_0$, and C_3, C_4 are constants depending on ε. Furthermore, if ξ satisfies Condition 2.3 or 2.4 correspondingly, then by choosing $\varepsilon \in ((1-\varepsilon_0)/\xi, 2\varepsilon_0)$, we have*

$$\max_{Z\in\mathscr{G}(\gamma_0)} \frac{\widetilde{\Pi}(Z|A)}{\widetilde{\Pi}(g(Z)|A)} \leq \exp(-C\bar{n}I),$$

for some constant $C > 1-\varepsilon_0$ with probability at least $1-C_3 n^{-C_4}$.

Note that when the connectivity probabilities are unknown, the initial conditions are different for the case of two communities and the case of more than two communities. In order to prove

Lemma A.1.4, we again divide S_α into a small mistake region and a large mistake region, according to whether $m > \gamma n$, where $\gamma \to 0$ is a positive sequence to be specified later. It is worth noting that we always start from the likelihood modularity, and then bound the exact posterior distribution.

Proof of Lemma A.1.4. Under the conditions of Lemma A.1.4, let $\varepsilon_{\gamma_0} = 1 - K\gamma_0$ for simplicity, and we have $\varepsilon_{\gamma_0}^4 nI \to \infty$, $\varepsilon_{\gamma_0}(1 - K\beta\gamma_0)n \to \infty$. Then, for any positive sequences $\bar{\varepsilon}, \gamma, \theta \to 0$ satisfying that $\bar{\varepsilon}^2 nI \to \infty$, $\theta^2 \gamma nI \to \infty$, $\varepsilon_{\gamma_0}^2 \gg \bar{\varepsilon}$, and $\varepsilon_{\gamma_0}^3 \gg \theta\bar{\varepsilon}$. To be specific, we can set $\bar{\varepsilon} = \varepsilon_{\gamma_0}^2/(\varepsilon_{\gamma_0}^4 nI)^{1/4}$, $\gamma = \varepsilon_{\gamma_0}^2$, $\theta = 1/\sqrt{\gamma nI}$.

All the following analyses are based on the event $\mathscr{E}(\bar{\varepsilon}, \gamma, \theta)$.

Small mistake region. We write $\mathscr{M}_s = \{Z \in \mathscr{G}(\gamma_0) : m \le \gamma n\}$. By Lemma A.1.27, since $\gamma < c_{\alpha,\beta}$, we have that for any $Z \in \mathscr{M}_s$, $\mathscr{B}(Z) \subset S_\alpha$. By Lemma A.1.20, under the event $\mathscr{E}_1(\bar{\varepsilon})$, we have

$$\begin{aligned}
&\max_{Z\in M_s} \min_{Z'\in\mathscr{B}(Z)} \log \frac{\Pi(Z|A)}{\Pi(Z'|A)} \\
\le &\max_{Z\in M_s} \frac{1}{m} \sum_{Z'\in\mathscr{B}(Z)} \log \frac{\Pi(Z|A)}{\Pi(Z'|A)} \\
\le &\max_{Z\in M_s} \frac{1}{m} \sum_{Z'\in\mathscr{B}(Z)} \big(Q_{LM}(Z,A) - Q_{LM}(Z',A)\big) + \varepsilon_{LM}.
\end{aligned} \tag{25}$$

Thus, we proceed to upper bound $Q_{LM}(Z,A) - Q_{LM}(Z',A)$. By some calculations, we have

$$\begin{aligned}
&Q_{LM}(Z,A) - Q_{LM}(Z',A) \\
=&\log \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)} - \sum_{a\le a'} n_{aa'} \cdot D\left(\frac{O_{aa'}}{n_{aa'}} \middle\| \frac{O_{aa'}(Z)}{n_{aa'}(Z)}\right) + \\
&\underbrace{\sum_{a\le a'} \Delta O_{aa'} \left(\log \frac{O_{aa'}(Z)}{n_{aa'}(Z)} - \log B_{aa'}\right) + (\Delta n_{aa'} - \Delta O_{aa'}) \left(\log \frac{n_{aa'}(Z) - O_{aa'}(Z)}{n_{aa'}(Z)} - \log(1 - B_{aa'})\right)}_{Err(Z,Z')},
\end{aligned} \tag{26}$$

where $\log\big[\Pi_0(Z|A)/\Pi_0(Z'|A)\big]$ is calculated in (17). Now suppose we correct one sample from a

misclassified group b to its true group b'. Then, by Lemma A.1.28, we have

$$\Delta O_{b'b'} + \Delta O_{bb} + \Delta O_{bb'} = 0, \;\; \Delta O_s = \Delta O_{bb} + \Delta O_{b'b'},$$
$$\Delta O_{ab} + \Delta O_{ab'} = 0, \;\; \Delta O_{aa'} = 0, \quad \text{for any } a, a' \in [K] \setminus \{b, b'\}.$$

Denote $\widetilde{B}_{aa'} = \mathbb{E}[O_{aa'}(Z)]/n_{aa'}(Z)$ and $\widehat{B}_{aa'} = O_{aa'}(Z)/n_{aa'}(Z)$ for any $a, a' \in [K]$. By Lemma A.1.21, we have

$$\frac{\left|\widetilde{B}_{aa'} - B_{aa'}\right|}{p-q} = \begin{cases} \dfrac{\sum_{k \neq l} R_{ak} R_{al}}{n'_a(n'_a - 1)}, & \text{if } a = a', \\ \dfrac{\sum_k R_{ak} R_{a'k}}{n'_a n'_{a'}}, & \text{if } a \neq a', \end{cases} \tag{27}$$

and $\left\|\widetilde{B} - B\right\|_\infty \leq 2K\alpha m(p-q)/n$. Under the event $\mathscr{E}_1(\bar{\varepsilon})$ defined in (6), by Lemma A.1.9, we have

$$\left\|\widehat{B} - B\right\|_\infty \leq \left\|\widehat{B} - \widetilde{B}\right\|_\infty + \left\|\widetilde{B} - B\right\|_\infty \leq \left(C\bar{\varepsilon} + \frac{2K\alpha m}{n}\right)(p-q) \lesssim (\gamma + \bar{\varepsilon})(p-q). \tag{28}$$

We then bound $Err(Z, Z')$ in (26) under the event $\mathscr{E}(\bar{\varepsilon}, \gamma, \theta)$. Since $p \asymp q$, by some calculations, we have

$$\sum_{Z' \in \mathscr{B}(Z)} Err(Z, Z') = \sum_{a \leq a'} \log \frac{\widehat{B}_{aa'}(1 - B_{aa'})}{B_{aa'}(1 - \widehat{B}_{aa'})} \sum_{Z' \in \mathscr{B}(Z)} \left(\Delta O_{aa'} - \lambda^*_{aa'} \Delta n_{aa'}\right) \tag{29}$$

$$\leq \frac{C}{p}\left\|\widehat{B} - B\right\|_\infty \underbrace{\sum_{a \leq a'} \left| \sum_{Z' \in \mathscr{B}(Z)} \left(\Delta O_{aa'} - \lambda^*_{aa'} \Delta n_{aa'}\right) \right|}_{(A)} \tag{30}$$

for some constant C, where

$$\lambda^*_{aa'} = \log \frac{1 - B_{aa'}}{1 - \widehat{B}_{aa'}} \Big/ \log \frac{\widehat{B}_{aa'}(1 - B_{aa'})}{B_{aa'}(1 - \widehat{B}_{aa'})} \in \left[B_{aa'} \wedge \widehat{B}_{aa'}, B_{aa'} \vee \widehat{B}_{aa'}\right].$$

Under the event $\mathscr{E}(\bar{\varepsilon}, \gamma, \theta)$, for any $Z \in \mathscr{M}_s$, we bound the above term (A) by the following three

terms separately,

$$\sum_{a\le a'}\left|\sum_{Z'\in\mathscr{B}(Z)}(\Delta O_{aa'}-\mathbb{E}[\Delta O_{aa'}])\right|\lesssim mn(p-q), \tag{31}$$

$$\sum_{a\le a'}\left|\sum_{Z'\in\mathscr{B}(Z)}(\mathbb{E}[\Delta O_{aa'}]-B_{aa'}\Delta n_{aa'})\right|\le\sum_{Z'\in\mathscr{B}(Z)}\sum_{a\le a'}|\mathbb{E}[\Delta O_{aa'}]-B_{aa'}\Delta n_{aa'}|\le 2Kmn(p-q), \tag{32}$$

$$\sum_{a\le a'}\left|\sum_{Z'\in\mathscr{B}(Z)}(B_{aa'}-\lambda^*_{aa'})\Delta n_{aa'}\right|\le\sum_{Z'\in\mathscr{B}(Z)}\sum_{a\le a'}|\Delta n_{aa'}|\cdot\left\|\widehat{B}-B\right\|_\infty\lesssim(\gamma+\bar{\varepsilon})mn(p-q). \tag{33}$$

The first inequality directly follows by Lemma A.1.13. The second inequality is due to that for each fixed Z', there are at most $2K$ pairs of groups contributing to the summations of the absolute values, and for each summation, there are at most n random variables associated. The third inequality follows by (28) and Lemma A.1.28. Hence, under the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$, we have

$$\sum_{Z'\in\mathscr{B}(Z)}Err(Z,Z')\le C(\gamma+\bar{\varepsilon})m\bar{n}I,$$

for some constant C where $\gamma, \bar{\varepsilon} \to 0$ as defined in the beginning of the proof. Hence, it follows that

$$
\begin{aligned}
&\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}\min_{Z'\in\mathscr{B}(Z)}\log\frac{\Pi(Z|A)}{\Pi(Z'|A)} > -\varepsilon\bar{n}I\right\}\\
\leq&\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}\frac{1}{m}\sum_{Z'\in\mathscr{B}(Z)}\log\frac{\Pi(Z|A)}{\Pi(Z'|A)} > -\varepsilon\bar{n}I\right\}\\
\leq&\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}\frac{1}{m}\sum_{Z'\in\mathscr{B}(Z)}Q_{LM}(Z,A)-Q_{LM}(Z',A) > -\varepsilon\bar{n}I-\varepsilon_{LM},\mathscr{E}\right\}+\mathbb{P}\{\mathscr{E}^c\}\\
\leq&\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}\frac{1}{m}\sum_{Z'\in\mathscr{B}(Z)}\left(\log\frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)}+Err(Z,Z')\right) > -\varepsilon\bar{n}I-\varepsilon_{LM},\mathscr{E}\right\}+\mathbb{P}\{\mathscr{E}^c\}\\
\leq&\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}\frac{1}{m}\sum_{Z'\in\mathscr{B}(Z)}\log\frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)} > -\varepsilon\bar{n}I-C(\gamma+\bar{\varepsilon})\bar{n}I-\varepsilon_{LM}\right\}+\mathbb{P}\{\mathscr{E}^c\}\\
=&\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}t^*\sum_{Z'\in\mathscr{B}(Z)}\Delta_n(Z,Z') > -m\varepsilon\bar{n}I/2-Cm(\gamma+\bar{\varepsilon})\bar{n}I/2-m\varepsilon_{LM}/2\right\}+\mathbb{P}\{\mathscr{E}^c\}.
\end{aligned}
\tag{34}
$$

where $\varepsilon_{LM} \to 0$ is defined in (51), and ε is any small constant satisfying $\varepsilon < 2\varepsilon_0$. A simple union bound and following the argument from (21) to (22) lead to that

$$
\mathbb{P}\left\{\max_{Z\in\mathscr{M}_s}\min_{Z'\in\mathscr{B}(Z)}\log\frac{\Pi(Z|A)}{\Pi(Z'|A)} > -\varepsilon\bar{n}I\right\} \leq C'n\exp\left(-(1-\varepsilon/2)\bar{n}I(1-o(1))\right),
$$

for some constant C'.

Remark A.1.1. Before performing the analysis for the large mistake region, it is worth noting that for the small mistake region, the proof works for any sequence $\gamma \to 0$. Thus, in the case of more than two communities ($K \geq 3$), if $\gamma_0 \to 0$, by Lemma 2.4.4, $\mathscr{G}(\gamma_0) \subset \mathscr{M}_s$ for some sequence $\gamma \to 0$, and then the proof is complete. Therefore, we only need to analyze the large mistake region for $K = 2$.

Large mistake region. We write $\mathscr{M}_l = \{Z \in \mathscr{G}(\gamma_0) : m > \gamma n\}$. By the same argument in (25),

we start with $Q_{LM}(Z,A) - Q_{LM}(Z',A)$. For $K = 2$ and $Z \in \mathscr{M}_l$, by some calculations, we have

$$
\begin{aligned}
& Q_{LM}(Z,A) - Q_{LM}(Z',A) \\
&= \sum_{a \le a'} n_{aa'}(Z)\tau\left(\frac{O_{aa'}(Z)}{n_{aa'}(Z)}\right) - \sum_{a \le a'} n_{aa'}(Z')\tau\left(\frac{O_{aa'}(Z')}{n_{aa'}(Z')}\right) \\
&= \sum_{a \le a'} \Delta O_{aa'} \log \frac{O_{aa'}(Z)}{n_{aa'}(Z)} + (\Delta n_{aa'} - \Delta O_{aa'}) \log\left(1 - \frac{O_{aa'}(Z)}{n_{aa'}(Z)}\right) - \sum_{a \le a'} n_{aa'}(Z') D\left(\frac{O_{aa'}(Z')}{n_{aa'}(Z')} \middle\| \frac{O_{aa'}(Z)}{n_{aa'}(Z)}\right) \\
&\le \underbrace{\sum_{a \le a'} \Delta O_{aa'} \log \widetilde{B}_{aa'} + (\Delta n_{aa'} - \Delta O_{aa'}) \log\left(1 - \widetilde{B}_{aa'}\right)}_{P(Z,Z') + Err_1(Z,Z')} + \\
&\quad \underbrace{\sum_{a \le a'} \Delta O_{aa'} \left(\log \frac{O_{aa'}(Z)}{n_{aa'}(Z)} - \log \widetilde{B}_{aa'}\right) + (\Delta n_{aa'} - \Delta O_{aa'}) \left(\log\left(1 - \frac{O_{aa'}(Z)}{n_{aa'}(Z)}\right) - \log\left(1 - \widetilde{B}_{aa'}\right)\right)}_{Err_2(Z,Z')},
\end{aligned}
\tag{35}
$$

where we write

$$
P(Z,Z') = \sum_{a \le a'} \mathbb{E}[\Delta O_{aa'}] \log \widetilde{B}_{aa'} + (\Delta n_{aa'} - \mathbb{E}[\Delta O_{aa'}]) \log\left(1 - \widetilde{B}_{aa'}\right), \tag{36}
$$

$$
Err_1(Z,Z') = \sum_{a \le a'} (\Delta O_{aa'} - \mathbb{E}[\Delta O_{aa'}]) \log \frac{\widetilde{B}_{aa'}}{1 - \widetilde{B}_{aa'}}. \tag{37}
$$

Recall $g(Z)$ and $\mathscr{B}(Z)$ are defined in (32) and (1) respectively. Let $N = |B(Z) \cap S_\alpha|$. By Lemma A.1.27, we have $N \ge \min\{cn, m\}$ for some constant c.

Step 1: bound $P(Z,Z')$. By Lemma A.1.32, for $Z \in \mathscr{M}_l$ with $m = d(Z,Z^*)$ and any $Z' \in \mathscr{B}(Z) \cap S_\alpha$, we have

$$
-P(Z,Z') \ge \frac{nI}{2\alpha^2} \varepsilon_m^3 \max\{1 - \beta + 1/\alpha, \varepsilon_m\}(1 - o(1)) \gtrsim \varepsilon_m^3 nI,
$$

where $\varepsilon_m = 1 - Km/n$. Note that $m < n/K\beta$ under the condition. The last inequality holds because $\varepsilon_m \to 0$ only if $\beta \to 1$, and $m/n \to 1/K$. Thus, $1 - \beta + 1/\alpha \gg \varepsilon_m$.

Step 2: bound $Err_1(Z,Z')$. Recall that $N = |\mathscr{B}(Z) \cap S_\alpha|$. Under the event $\mathscr{E}_6(\gamma,\theta)$ defined in (6), we have

$$\max_{Z\in S_\alpha : m>\gamma n} \frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \sum_{a\le a'} |\Delta O_{aa'} - \mathbb{E}[\Delta O_{aa'}]| \le \theta n(p-q),$$

for the positive sequence $\theta \to 0$ defined in the beginning of the proof. When $K=2$, and Z' corrects one sample from group b to b', by Lemma A.1.28, $Err_1(Z,Z')$ in (37) can be expressed as

$$Err_1(Z,Z') = (\Delta O_{bb} - \mathbb{E}[\Delta O_{bb}]) \log \frac{\widetilde{B}_{bb}(1-\widetilde{B}_{b'b})}{\widetilde{B}_{b'b}(1-\widetilde{B}_{bb})} + (\Delta O_{b'b'} - \mathbb{E}[\Delta O_{b'b'}]) \log \frac{\widetilde{B}_{b'b'}(1-\widetilde{B}_{bb'})}{\widetilde{B}_{bb'}(1-\widetilde{B}_{b'b'})}.$$

By $p \asymp q$ and $(n-2m)(n-2\beta m)/n \to \infty$, it follows by Lemma A.1.29 that

$$\begin{aligned}
\frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} Err_1(Z,Z') &\le C_1 \frac{\theta n(p-q)}{p} \cdot \left(\left|\widetilde{B}_{b'b'} - \widetilde{B}_{bb'}\right| + \left|\widetilde{B}_{bb} - \widetilde{B}_{b'b}\right| \right) \\
&\le C_2 \frac{\theta n(p-q)^2 \det(R)}{p} \left(\frac{2\alpha}{n}\right)^3 (|R_{b'b'} - R_{b'b}| + |R_{bb} - R_{bb'}|) \\
&\le C_3 \frac{\theta I \det(R)(n-2m)}{n^2},
\end{aligned}$$

for some constants C_1, C_2, C_3. Hence, under the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$, where the sequences are defined in the beginning of the proof, if $(n-2m)(n-2\beta m)/n \to \infty$ and $(n-2m)/n \gg \theta$, then by Lemma A.1.33 we have

$$\frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} Err_1(Z,Z') \ll \frac{\theta I \det(R)(n-2m)}{n^2} \ll -\frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} P(Z,Z'),$$

for any $Z \in \mathscr{M}_l$.

Step 3: bound $Err_2(Z,Z')$. Recall $N = |\mathscr{B}(Z) \cap S_\alpha|$. By (35), we have

$$\begin{aligned}\frac{1}{N}\sum_{Z' \in \mathscr{B}(Z)\cap S_\alpha} Err_2(Z,Z') &= \frac{1}{N}\sum_{a \le a'} \log\frac{\widehat{B}_{aa'}(1-\widetilde{B}_{aa'})}{\widetilde{B}_{aa'}(1-\widehat{B}_{aa'})} \sum_{Z' \in \mathscr{B}(Z)\cap S_\alpha} \left(\Delta O_{aa'} - \lambda^*_{aa'}\Delta n_{aa'}\right) \\ &\le \frac{C}{p}\left\|\widehat{B}-\widetilde{B}\right\|_\infty \sum_{a \le a'} \left|\frac{1}{N}\sum_{Z' \in \mathscr{B}(Z)\cap S_\alpha} \left(\Delta O_{aa'} - \lambda^*_{aa'}\Delta n_{aa'}\right)\right|,\end{aligned}$$

where

$$\lambda^*_{aa'} = \log\frac{1-\widetilde{B}_{aa'}}{1-\widehat{B}_{aa'}} \Big/ \log\frac{\widehat{B}_{aa'}(1-\widetilde{B}_{aa'})}{\widetilde{B}_{aa'}(1-\widehat{B}_{aa'})} \in \left[\widetilde{B}_{aa'} \wedge \widehat{B}_{aa'}, \widetilde{B}_{aa'} \vee \widehat{B}_{aa'}\right].$$

Under the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$ defined in (6), by Lemma A.1.9, we have $\max_{Z\in S_\alpha}\left\|\widehat{B}-\widetilde{B}\right\|_\infty \lesssim \bar{\varepsilon}(p-q)$. By the same argument from (28) to (33), we have

$$\begin{aligned}\frac{1}{N}\sum_{a\le a'}\left|\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}(\Delta O_{aa'} - \mathbb{E}[\Delta O_{aa'}])\right| &\le \theta n(p-q), \\ \frac{1}{N}\sum_{a\le a'}\left|\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\left(\mathbb{E}[\Delta O_{aa'}] - \widetilde{B}_{aa'}\Delta n_{aa'}\right)\right| &\le C(n-2m)(p-q) \lesssim \varepsilon_m n(p-q), \\ \frac{1}{N}\sum_{a\le a'}\left|\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}(\widetilde{B}_{aa'} - \lambda_{aa'})\Delta n_{aa'}\right| &\le K^2 n\left\|\widehat{B}-\widetilde{B}\right\|_\infty \lesssim \bar{\varepsilon} n(p-q).\end{aligned}$$

It follows that for any $Z \in \mathscr{M}_l$ with $d(Z,Z^*) = m$,

$$\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} Err_2(Z,Z') \lesssim \frac{\bar{\varepsilon}n(p-q)^2}{p}\left(\theta + \varepsilon_m + \bar{\varepsilon}\right),$$

where $\varepsilon_m = 1 - 2m/n$. Hence, under the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$, where the sequences are defined in the beginning of the proof. If $\varepsilon_m^3 \gg \theta\bar{\varepsilon}$, $\varepsilon_m^2 \gg \bar{\varepsilon}$, and $\varepsilon_m^3 \gg \bar{\varepsilon}^2$, then we have

$$\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} Err_2(Z,Z') \ll -\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} P(Z,Z'),$$

for any $Z \in \mathscr{M}_l$.

By combining all three steps, we require that for all $Z \in \mathscr{G}(\gamma_0)$ with $d(Z,Z^*) = m$,

$$(1-2m/n)^4 nI \to \infty, \;\; (n-2m)(n-2\beta m)/n \to \infty \text{ (for Lemma A.1.32).} \tag{38}$$

By the definition of $\mathscr{G}(\gamma_0)$ in (30) and by Lemma 2.4.4, it suffices to require

$$(1-2\gamma_0)^4 nI \to \infty, \quad (1-2\gamma_0)(1-2\beta\gamma_0)n \to \infty.$$

Recall that $\varepsilon_{\gamma_0} = 1-2\gamma_0$ in the beginning of the proof. Then, under the event $\mathscr{E}(\bar{\varepsilon},\gamma,\theta)$ defined in (6), by the conclusions of three steps and by Lemma A.1.20, we have

$$\begin{aligned}
&\max_{Z\in\mathscr{M}_l} \min_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \log\frac{\Pi(Z|A)}{\Pi(Z'|A)} \\
\le &\max_{Z\in\mathscr{M}_l} \frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \log\frac{\Pi(Z|A)}{\Pi(Z'|A)} \\
\le &\max_{Z\in\mathscr{M}_l} \frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} (Q_{LM}(Z,A) - Q_{LM}(Z',A) + \varepsilon_{LM}) \\
= &\max_{Z\in\mathscr{M}_l} \frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \left\{P(Z,Z') + Err_1(Z,Z') + Err_2(Z,Z')\right\} + \varepsilon_{LM} \\
= &\max_{Z\in\mathscr{M}_l} \frac{1}{N} \sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha} P(Z,Z')(1-o(1)) + \varepsilon_{LM} \\
\le &-\frac{nI}{2\alpha^2}\varepsilon_{\gamma_0}^3 \max\{1-\beta+1/\alpha, \varepsilon_{\gamma_0}\}(1-o(1)) \\
\le &-\frac{nI}{2\alpha^2}\varepsilon_{\gamma_0}^4 (1-o(1)).
\end{aligned} \tag{39}$$

Combine the results of two regions, and Lemma A.1.4 directly follows. □

A.1.6 Proof of Lemma A.1.1

For the simplicity of presentation, we first introduce some notations that will be used in the proof. Denote

$$\widetilde{m} = n\max\{\gamma_0, n^{-\tau}\} + \log^2 n, \quad m^* = n^{1-\tau},$$

where τ is a sufficiently small constant defined in (30). For any $Z \in \mathscr{G}(\gamma_0)$, we define the following set

$$\mathscr{S}(Z,\eta) = \left\{ S' \subset \mathscr{N}(Z) \cap S_\alpha : |S'| \geq \eta \, |\mathscr{B}(Z) \cap S_\alpha| \right\}, \tag{40}$$

where η is a small constant satisfying $2\tau < \eta < 1/2$, and $\mathscr{A}(Z), \mathscr{B}(Z), \mathscr{N}(Z)$ are defined in (1). Then, we have the following lemma.

Lemma A.1.5. *Suppose γ_0, ξ satisfy all conditions for Theorem 2.2.1 or those for Theorem 2.2.3, with probability at least $1 - \exp(-n^{1-\tau})$, we have*

$$\max_{Z \in S_\alpha : m^* \leq m \leq \widetilde{m}} \max_{S' \in \mathscr{S}(Z,\eta)} \min \left\{ \min_{Z' \in S' \cap \mathscr{B}(Z)} \frac{\widetilde{\Pi}(Z|A)}{\widetilde{\Pi}(Z'|A)}, \min_{Z' \in S' \cap \mathscr{A}(Z)} \frac{\widetilde{\Pi}(Z'|A)}{\widetilde{\Pi}(Z|A)} \right\} \leq \exp(-C\bar{n}I),$$

for some constant $C > 1 - \varepsilon_0$, and $2\tau < \eta < 1/2$.

To understand Lemma A.1.5, we say that $Z' \in \mathscr{N}(Z)$ is *making a mistake* if $Z' \in \mathscr{B}(Z)$ but rejected, or $Z' \in \mathscr{A}(Z)$ but accepted. Lemma A.1.5 implies that under all the conditions, for any current state Z with $m \in [m^*, \widetilde{m}]$, if we make at least $\eta|\mathscr{B}(Z) \cap S_\alpha|$ different choices of Z', then there is at least one Z' such that it is not making a mistake with high probability. In other words, it holds with high probability that Z' will make less than $\eta|\mathscr{B}(Z) \cap S_\alpha|$ mistakes among all possible choices.

Though Lemma A.1.5 seems very similar to Lemma A.1.3, Lemma A.1.3 works for all $Z \in \mathscr{G}(\gamma_0)$, and focuses on the posterior ratio of the current state and the next possible state in set $\mathscr{B}(Z)$, while Lemma A.1.5 works for $Z \in \mathscr{G}(\gamma_0)$ with $m \in [m^*, \widetilde{m}]$, and also bounds the probability of updating to $\mathscr{A}(Z)$.

Proof of Lemma A.1.1. By Lemma A.1.5, with probability at least $1-\exp(-n^{1-\tau})$, for any $Z \in \mathscr{G}(\gamma_0)$ with $m \in [m^*, \widetilde{m}]$, by (2), we have that

$$p_m(Z) = p(Z, \mathscr{A}(Z)) = \frac{1}{2(K-1)n} \sum_{Z' \in \mathscr{A}(Z) \cap S_\alpha} \min\left\{1, \frac{\widetilde{\Pi}_g(Z'|A)}{\widetilde{\Pi}_g(Z|A)}\right\} \leq \frac{\eta |\mathscr{B}(Z) \cap S_\alpha| + \varepsilon}{2(K-1)n},$$
$$q_m(Z) = p(Z, \mathscr{B}(Z)) = \frac{1}{2(K-1)n} \sum_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \min\left\{1, \frac{\widetilde{\Pi}_g(Z'|A)}{\widetilde{\Pi}_g(Z|A)}\right\} \geq \frac{(1-\eta)|\mathscr{B}(Z) \cap S_\alpha|}{2(K-1)n},$$

where $\varepsilon = n\exp(-C\bar{n}I) \to 0$ for some $C > 1-\varepsilon_0$. It follows that with probability at least $1-\exp(-n^{1-\tau})$, $p_m(Z)/q_m(Z) \leq 2\eta$ holds for any $Z \in \mathscr{G}(\gamma_0)$ with $m \in [m^*, \widetilde{m}]$. □

In order to prove Lemma A.1.5, we first state two lemmas according to whether the connectivity probabilities are known or not. In the case of known connectivity probabilities, we use $\Pi_0(\cdot|A)$ to denote the posterior distribution.

Lemma A.1.6. *When p, q are both known, given τ sufficiently small and η satisfying $2\tau < \eta < 1/2$, if $(1-K\alpha\gamma_0)^2 nI \to \infty$, then we have*

$$\max_{S' \in \mathscr{S}(Z,\eta)} \min\left\{\min_{Z' \in S' \cap \mathscr{B}(Z)} \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)}, \min_{Z' \in S' \cap \mathscr{A}(Z)} \frac{\Pi_0(Z'|A)}{\Pi_0(Z|A)}\right\}$$
$$\leq \begin{cases} \exp\left(-\varepsilon\bar{n}I\right), & \text{if } m^* \leq m \leq \gamma n, \\ \exp\left(-4(1/K\alpha - \gamma_0)nI(1-o(1))\right), & \text{if } \gamma n < m \leq \widetilde{m}, \end{cases}$$

holds uniformly for all $Z \in \mathscr{G}(\gamma_0)$ with $m \in [m^*, \widetilde{m}]$ and some sequence $\gamma \to 0$, with probability at least $1-\exp\left(-n^{1-\tau}\right)$. Here, ε is any small constant satisfying $\varepsilon < 2\varepsilon_0$.

Lemma A.1.7. *Given τ sufficiently small and η satisfying $2\tau < \eta < 1/2$. Suppose γ_0 satisfies Condition 2.2 or 2.4, there exists some positive sequence $\gamma \to 0$ such that with probability at least*

$1-\exp\left(-n^{1-\tau}\right)$,

$$\max_{S'\in\mathscr{S}(Z,\eta)}\min\left\{\min_{Z'\in S'\cap\mathscr{B}(Z)}\frac{\Pi(Z|A)}{\Pi(Z'|A)},\min_{Z'\in S'\cap\mathscr{A}(Z)}\frac{\Pi(Z'|A)}{\Pi(Z|A)}\right\}$$
$$\leq\begin{cases}\exp\left(-\varepsilon\bar{n}I(1-o(1))\right), & \text{if } m^*\leq m\leq\gamma n,\\ \exp\left(-\dfrac{(1-K\gamma_0)^4nI}{2\alpha^3}(1-o(1))\right), & \text{if } \gamma n<m\leq\widetilde{m},\end{cases}$$

holds uniformly for all $Z\in\mathscr{G}(\gamma_0)$ *with* $m\in[m^*,\widetilde{m}]$. *Here,* ε *is any constant satisfying* $\varepsilon<2\varepsilon_0$.

The proofs of Lemma A.1.6 and Lemma A.1.7 will be presented in the sequel. We first proceed to prove Lemma A.1.5 based on these two lemmas.

Proof of Lemma A.1.5. The result directly follows Lemma A.1.6 and Lemma A.1.7 by choosing ξ properly in Theorem 2.2.1 or Theorem 2.2.3. Then, by choosing $\varepsilon\in((1-\varepsilon_0)/\xi,2\varepsilon_0)$, we have that

$$\max_{Z\in S_\alpha:m^*\leq m\leq\widetilde{m}}\max_{S'\in\mathscr{S}(Z,\eta)}\min\left\{\min_{Z'\in S'\cap\mathscr{B}(Z)}\frac{\widetilde{\Pi}(Z|A)}{\widetilde{\Pi}(Z'|A)},\min_{Z'\in S'\cap\mathscr{A}(Z)}\frac{\widetilde{\Pi}(Z'|A)}{\widetilde{\Pi}(Z|A)}\right\}\leq\exp(-C\bar{n}I),$$

for some constant $C>1-\varepsilon_0$ with probability at least $1-\exp(-n^{1-\tau})$ for the sufficiently small constant τ. □

We finally present the proofs of Lemma A.1.6 and Lemma A.1.7 to complete this section.

Proof of Lemma A.1.6. We consider any positive sequences γ,θ satisfying $\gamma,\theta\to0$, $\theta^2\gamma nI\to\infty$, and $\theta\ll1-K\alpha\gamma_0$. Suppose $\gamma n\in[m^*,\widetilde{m}]$, and we perform analyses for the following cases.

Case 1: $m^* \le m \le \gamma n$**.** Since the minimum is upper bounded by the average, it follows that

$$\begin{aligned}
&\log \max_{Z: m^* \le m \le \gamma n} \max_{S' \in \mathscr{S}(Z,\eta)} \min\left\{ \min_{Z' \in S' \cap \mathscr{B}(Z)} \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)}, \min_{Z' \in S' \cap \mathscr{A}(Z)} \frac{\Pi_0(Z'|A)}{\Pi_0(Z|A)} \right\} \qquad (41)\\
\le &\max_{Z: m^* \le m \le \gamma n} \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \left(\sum_{Z' \in S' \cap \mathscr{B}(Z)} \log \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)} + \sum_{Z' \in S' \cap \mathscr{A}(Z)} \log \frac{\Pi_0(Z'|A)}{\Pi_0(Z|A)} \right)\\
= &\max_{Z: m^* \le m \le \gamma n} \max_{S' \in \mathscr{S}(Z,\eta)} \frac{2t^*}{|S'|} \left(\sum_{Z' \in S' \cap \mathscr{B}(Z)} \Delta_n(Z,Z') + \sum_{Z' \in S' \cap \mathscr{A}(Z)} \Delta_n(Z',Z) \right),
\end{aligned}$$

where $\Delta_n(Z,Z')$ is defined in (16). By a similar argument from (21) to (22), we have

$$\mathbb{E}\left[t^* \left(\sum_{Z' \in S' \cap \mathscr{B}(Z)} \Delta_n(Z,Z') + \sum_{Z' \in S' \cap \mathscr{A}(Z)} \Delta_n(Z',Z) \right) \right] \le \exp(-(1-C\gamma)|S'|\bar{n}I).$$

We also have $|\mathscr{B}(Z) \cap S_\alpha| = m$ by Lemma A.1.27. For any small constant $\varepsilon < 2\varepsilon_0$, write $C_{\gamma,\varepsilon} = 1 - C\gamma - \varepsilon/2$ for simplicity. It follows that

$$\begin{aligned}
\mathbb{P}\{(41) \ge -\varepsilon\bar{n}I\} &\le \underbrace{\sum_{m^* \le m \le \gamma n} \binom{n}{m} (K-1)^m}_{\text{all possible } Z} \underbrace{\sum_{|S'| \ge \eta m} \binom{Kn}{|S'|}}_{\text{all possible } S'} \underbrace{\exp\left(-C_{\gamma,\varepsilon}|S'|\bar{n}I\right)}_{\text{bound for each given } Z \text{ and } S'}\\
&\lesssim \sum_{m^* \le m \le \gamma n} \binom{n}{m} (K-1)^m \binom{Kn}{\eta m} \exp\left(-C_{\gamma,\varepsilon}\eta m \bar{n}I\right)\\
&\lesssim \binom{n}{m^*} (K-1)^{m^*} \binom{Kn}{\eta m^*} \exp\left(-C_{\gamma,\varepsilon}\eta m^* \bar{n}I\right)\\
&\lesssim \exp\left(-C_{\gamma,\varepsilon}\eta m^* \bar{n}I + (\eta+1) m^* \log \frac{eKn}{\eta m^*} \right). \qquad (42)
\end{aligned}$$

For any sufficiently small τ, when $\eta > 2\tau$ and $m^* = n^{1-\tau}$, it is easy to check that

$$(42) \lesssim \exp\left(-m^* \left(C_{\gamma,\varepsilon}\eta\bar{n}I - (\eta+1)\log\frac{eKn^\tau}{\eta} \right) \right) \le e^{-m^*}.$$

Case 2: $\gamma n < m \le \widetilde{m}$**.** Recall that $\Delta O_s(Z,Z') = O_s(Z) - O_s(Z')$ for any label assignments Z and

Z'. By (17) and (23), we have that

$$\log \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)} = 2t^* \Delta_n(Z,Z'),$$
$$\max_{Z\in\mathscr{G}(\gamma_0)} \max_{Z'\in\mathscr{B}(Z)\cap S_\alpha} \mathbb{E}\left[\Delta_n(Z,Z')\right] \le -\left(\frac{1}{K\alpha} - \gamma_0\right)(p-q)(1-o(1)).$$

Since $|S'| \ge \eta |\mathscr{B}(Z)\cap S_\alpha|$, by the same proof in Lemma A.1.14, we have that with probability at least $1-\exp(-n)$,

$$\begin{aligned}
&\max_{Z\in S_\alpha:\gamma n<m\le\widetilde{m}} \max_{S'\in S(Z,\alpha)} \left\{ \frac{1}{|S'|} \sum_{Z'\in S'} \left|\Delta_n(Z,Z') - \mathbb{E}\left[\Delta_n(Z,Z')\right]\right| \right\} \\
=&\max_{Z\in S_\alpha:\gamma n<m\le\widetilde{m}} \max_{S'\in S(Z,\alpha)} \left\{ \frac{1}{|S'|} \sum_{Z'\in S'} \left|\Delta O_s(Z,Z') - \mathbb{E}\left[\Delta O_s(Z,Z')\right]\right| \right\} \\
\le&\theta n(p-q),
\end{aligned}$$

where the positive sequence θ is defined in the beginning of the proof. Thus, by a similar argument from (23) to (24), the result directly follows.

Combining two cases gives Lemma A.1.6 directly. Note that in the case of $m^* \ge \gamma n$ or $\widetilde{m} < \gamma n$, the result trivially follows. □

Proof of Lemma A.1.7. Consider any positive sequences $\bar{\varepsilon}, \gamma, \theta \to 0$ satisfying that $\bar{\varepsilon}^2 nI \to \infty$, $\theta^2\gamma nI \to \infty$, $(1-K\gamma_0)^2 \gg \bar{\varepsilon}$, and $(1-K\gamma_0)^3 \gg \theta\bar{\varepsilon}$. Note that the second case in Lemma A.1.7 is only for the case of $K=2$. When $K \ge 3$, we require $\gamma_0 \to 0$, and thus there exists some $\gamma \to 0$, such that for all $Z \in \mathscr{G}(\gamma_0)$, $m \le \gamma n$ always holds.

The following proof is similar with those of Lemma A.1.4 and Lemma A.1.6. Denote $\Delta Q(Z,Z') = Q_{LM}(Z,A) - Q_{LM}(Z',A)$ for any two label assignments $Z, Z' \in S_\alpha$, where $Q_{LM}(Z,A)$ is the likelihood modularity function defined in (8). Under the event $\mathscr{E}_1(\bar{\varepsilon})$, by Lemma A.1.20, there exists

some sequence $\varepsilon_{LM} \to 0$ such that

$$\begin{aligned}
&\log \max_{S' \in \mathscr{S}(Z,\eta)} \min \left\{ \min_{Z' \in S' \cap \mathscr{B}(Z)} \frac{\Pi(Z|A)}{\Pi(Z'|A)}, \min_{Z' \in S' \cap \mathscr{A}(Z)} \frac{\Pi(Z'|A)}{\Pi(Z|A)} \right\} \\
\leq & \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \left\{ \sum_{Z' \in S' \cap \mathscr{B}(Z)} \log \frac{\Pi(Z|A)}{\Pi(Z'|A)} + \sum_{Z' \in S' \cap \mathscr{A}(Z)} \log \frac{\Pi(Z'|A)}{\Pi(Z|A)} \right\} \\
\leq & \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \left\{ \sum_{Z' \in S' \cap \mathscr{B}(Z)} \Delta Q(Z,Z') + \sum_{Z' \in S' \cap \mathscr{A}(Z)} \Delta Q(Z',Z) \right\} + \varepsilon_{LM}. \qquad (43)
\end{aligned}$$

The first inequality is because minimum is smaller than the average.

Case 1: $m^* \leq m \leq \gamma n$**.** In this case, $B(Z) \subset S_\alpha$ by Lemma A.1.27. By (26), we have that under the event $\mathscr{E}_1(\bar{\varepsilon})$,

$$(43) \leq \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \left\{ \sum_{Z' \in S' \cap \mathscr{B}(Z)} \log \frac{\Pi_0(Z|A)}{\Pi_0(Z'|A)} + \sum_{Z' \in S' \cap \mathscr{A}(Z)} \log \frac{\Pi_0(Z'|A)}{\Pi_0(Z|A)} + \sum_{Z' \in S'} \left| Err(Z,Z') \right| \right\} + \varepsilon_{LM}.$$

By a similar argument from (26) to (34), in order to prove Lemma A.1.7, it suffices to show that

$$\max_{Z \in \mathscr{G}(\gamma_0): m^* \leq m \leq \gamma n} \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \left| Err(Z,Z') \right| = o(\bar{n}I). \qquad (44)$$

Recall that $Err(Z,Z')$ is defined in (26). It follows by (30) that under the event $\mathscr{E}_1(\bar{\varepsilon})$, for any $Z \in \mathscr{G}(\gamma_0)$ with $m \in [m^*, \gamma n]$,

$$\begin{aligned}
&\max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \left| Err(Z,Z') \right| \\
\lesssim & (\gamma + \bar{\varepsilon}) \frac{p-q}{p} \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \sum_{a \leq a'} \left| \Delta O_{aa'} - \lambda^*_{aa'} \Delta n_{aa'} \right| \\
\lesssim & (\gamma + \bar{\varepsilon}) \frac{p-q}{p} \left\{ (A) + (B) + (C) \right\},
\end{aligned}$$

where

$$(A) = \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \sum_{a \leq a'} |\Delta O_{aa'} - \mathbb{E}[\Delta O_{aa'}]| \lesssim n(p-q),$$
$$(B) = \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \sum_{a \leq a'} |\mathbb{E}[\Delta O_{aa'}] - B_{aa'} \Delta n_{aa'}| \lesssim n(p-q),$$
$$(C) = \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \sum_{a \leq a'} \left|(B_{aa'} - \lambda^*_{aa'}) \Delta n_{aa'}\right| \ll n(p-q).$$

The first inequality holds with probability at least $1 - e^{-m^*}$ by the same proof of Lemma A.1.6 and Lemma A.1.13. The second and the third inequalities hold due to the same arguments for (32) and (33). Hence, the proof is complete for the small mistake region.

Case 2: $\gamma n < m \leq \widetilde{m}$**.** We only analyze this case for $K = 2$. By (35), we have that under the event $\mathscr{E}_1(\bar{\varepsilon})$,

$$(43) \leq \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \left\{ \sum_{Z' \in S' \cap \mathscr{B}(Z)} P(Z,Z') + \sum_{Z' \in S' \cap \mathscr{A}(Z)} P(Z',Z) \right\} +$$
$$\max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \left(\left|Err_1(Z,Z')\right| + \left|Err_2(Z,Z')\right| \right) + \varepsilon_{LM},$$

where $Err_1(Z,Z')$ and $Err_2(Z,Z')$ are defined in (35) and (37). According to arguments in Lemma 2.4.6, we only need to bound $Err_1(Z,Z')$ and $Err_2(Z,Z')$ in order to upper bound bound (43) as well as the posterior ratio. The only term inside $Err_1(Z,Z')$ and $Err_2(Z,Z')$ needed to be treated specially is

$$\max_{Z \in S_\alpha : \gamma n < m \leq \widetilde{m}} \max_{S' \in \mathscr{S}(Z,\eta)} \frac{1}{|S'|} \sum_{Z' \in S'} \sum_{a \leq a'} \left| \Delta O_{aa'}(Z,Z') - \mathbb{E}\left[\Delta O_{aa'}(Z,Z')\right] \right|,$$

denoted by D for simplicity. By the same proof of Lemma A.1.14, we have that

$$\mathbb{P}\{D \geq \theta n(p-q)\} \leq e^{-n},$$

for the positive sequence θ defined in the beginning of the proof. Hence, following the same arguments from (35) to (39), the proof of Lemma A.1.7 is completed for the large mistake region.

Combining two cases gives Lemma A.1.7 directly. Note that in the case of $m^* \geq \gamma n$ or $\widetilde{m} < \gamma n$, the result trivially follows. □

A.1.7 Proof of Lemma 2.2.1

In this section, we proceed to lower bound the posterior distribution.

When connectivity probabilities are known

Let $\{x_i\}_{i\geq 1}, \{y_j\}_{j\geq 1}$ be i.i.d. copies of Bernoulli(q) and Bernoulli(p). According to (24) and (12), for any $Z \in S_\alpha$, we have that

$$\log \frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} = \log \frac{p(1-q)}{q(1-p)} \left(\Delta\widetilde{O}_s - \mathbb{E}\left[\Delta\widetilde{O}_s\right] + \mathbb{E}\left[\Delta\widetilde{O}_s\right] - \lambda^* \Delta\widetilde{n}_s \right),$$

where $\Delta\widetilde{O}_s = \sum_{i=1}^{N_d} x_i - \sum_{i=1}^{N_s} y_i$, and

$$N_d = \sum_{i<j} \mathbb{I}\left\{ Z_i = Z_j, Z_i^* \neq Z_j^* \right\}, \;\; N_s = \sum_{i<j} \mathbb{I}\left\{ Z_i^* = Z_j^*, Z_i \neq Z_j \right\}. \tag{45}$$

Recall that $m_k = \sum_{i=1}^n \mathbb{I}\{Z_i = k,\ Z_i^* \neq k\}$, and it follows that $m = \sum_{k\in[K]} m_k$. Write

$$\beta_k = \sum_{i<j} \mathbb{I}\left\{ Z_i = Z_j = k, Z_i^* \neq Z_j^* \right\}$$

for simplicity, and we have

$$\beta_k = \sum_{a<b} R_{ka}R_{kb} \le R_{kk}\cdot m_k + m_k^2 = n_k' m_k \le \frac{\alpha n m_k}{K}.$$

It follows that $N_d = \sum_{k=1}^K \beta_k \le \alpha mn/K$. Similarly, we have $N_s \le \beta mn/K$, and

$$\begin{aligned}
\mathbb{E}\left[\Delta\widetilde{O}_s\right] - \lambda^* \Delta\widetilde{n}_s &= N_d q - N_s p - \lambda^*(N_d - N_s) \\
&= -(N_d\cdot(\lambda^* - q) + N_s\cdot(p-\lambda^*)) \\
&\ge -(N_d + N_s)\cdot(p-q) \ge -(\alpha+\beta)mn(p-q)/K.
\end{aligned}$$

Furthermore, by Lemma A.1.10, with probability at least $1 - n\exp(-(1-o(1))\bar{n}I)$, for any $Z \in S_\alpha$ with $m = d(Z, Z^*)$, we have

$$\begin{aligned}
\left|\Delta\widetilde{O}_s - \mathbb{E}\left[\Delta\widetilde{O}_s\right]\right| &\le \sum_{a\in[K]} \left|\Delta\widetilde{O}_{aa} - \mathbb{E}\left[\Delta\widetilde{O}_{aa}\right]\right| = \sum_{a\in[K]} |X_{aa}(Z) - X_{aa}(Z^*)| \\
&\le K\,\|X(Z) - X(Z^*)\|_\infty \le (\alpha+\beta)mn(p-q).
\end{aligned}$$

Hence, it follows that

$$\begin{aligned}
\log\frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)} &\ge -\log\frac{p(1-q)}{q(1-p)}\cdot Cmn(p-q) \\
&\ge -Cmn\cdot\frac{(p-q)^2}{q(1-p)} \\
&\ge -C'nmI\cdot(1+o(1)),
\end{aligned}$$

for some constants C and C' with probability at least $1 - n\exp(-(1-o(1))\bar{n}I)$.

When connectivity probabilities are unknown

In order to simplify the proof, we first define some events, and all the following analysis are conditioning on the given events. For any positive sequences $\gamma, \theta \to 0$ with $\gamma^2 nI \to \infty$, and $\theta^2 \gamma nI \to \infty$, let $\bar{\varepsilon} = \gamma$ for simplicity. Consider events $\mathscr{E}_1(\bar{\varepsilon})$, $\mathscr{E}_2$, $\mathscr{E}_3(\bar{\varepsilon})$, $\mathscr{E}_4(\gamma, \theta)$ defined in (6). Under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_3(\bar{\varepsilon})$, by Lemma A.1.19, we have that for any $Z \in S_\alpha$,

$$\log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} - \Delta Q(Z, Z^*) \geq -C_{LM},$$

where $\Delta Q(Z, Z^*) = Q_{LM}(Z, A) - Q_{LM}(Z^*, A)$. Thus, it suffices to lower bound $\Delta Q(Z, Z^*)$.

Case 1: $m \leq \gamma n$**.** By (11), we have

$$\begin{aligned}
&\Delta Q(Z, Z^*) \\
&= \underbrace{\log \frac{\Pi_0(Z|A)}{\Pi_0(Z^*|A)}}_{(A)} - \underbrace{\sum_{a \leq b} n_{ab}(Z^*) \cdot D\left(\frac{O_{ab}}{n_{ab}} \middle\| \frac{O_{ab}(Z)}{n_{ab}(Z)}\right)}_{(B)} \\
&\quad + \underbrace{\sum_{a \leq b} \Delta\widetilde{O}_{ab} \left(\log \frac{O_{ab}(Z)}{n_{ab}(Z)} - \log B_{ab}\right) + (\Delta\widetilde{n}_{ab} - \Delta\widetilde{O}_{ab}) \left(\log \frac{n_{ab}(Z) - O_{ab}(Z)}{n_{ab}(Z)} - \log(1 - B_{ab})\right)}_{(C)},
\end{aligned}$$

where $\Pi_0(\cdot|A)$ is defined in (12). We proceed to bound each term above separately. Under the event $\mathscr{E}_2$ and by the same argument in Section A.1.7, we have

$$A \gtrsim -mnI.$$

By Lemma A.1.36, under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_2$, we have

$$B \lesssim m^2 I.$$

By Lemma A.1.22, under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_2$, we have

$$C \gtrsim -(\bar{\varepsilon}+\gamma)mnI.$$

Hence, under the events $\mathscr{E}_1(\bar{\varepsilon}), \mathscr{E}_2, \mathscr{E}_3(\bar{\varepsilon})$, we have that for any $Z \in S_\alpha$ with $m \leq \gamma n$,

$$\log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} \geq \Delta Q(Z,Z^*) - C_{LM} \geq -CmnI,$$

for some constant C.

Case 2: $m > \gamma n$**.** By (14), we have

$$\Delta Q(Z,Z^*) = G(Z) - G(Z^*) + \Delta(Z) - \Delta(Z^*),$$

where by (15) and Lemma A.1.35, we have

$$G(Z) - G(Z^*) \geq -\frac{1}{2} \sum_{a,b,k,l} R_{ak} R_{bl} D\left(B_{kl} \middle\| \widetilde{B}ab\right) \gtrsim -mnI.$$

By Lemma A.1.26, under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_4(\gamma,\theta)$, for any $Z \in S_\alpha$ with $m > \gamma n$, we have

$$\Delta(Z) - \Delta(Z^*) \geq -\varepsilon mnI$$

for some sequence $\varepsilon \to 0$. Hence, under the events $\mathscr{E}_1(\bar{\varepsilon})$, $\mathscr{E}_3(\bar{\varepsilon})$, and $\mathscr{E}_4(\gamma,\theta)$, we have that for any $Z \in S_\alpha$ with $m > \gamma n$,

$$\log \frac{\Pi(Z|A)}{\Pi(Z^*|A)} \geq \Delta Q(Z,Z^*) - C_{LM} \geq -C'mnI,$$

for some constant C'.

Combining two cases, we have that with probability at least $1-n\exp(-(1-o(1))\bar{n}I)$,

$$\min_{Z\in S_\alpha}\left(\log\frac{\Pi(Z|A)}{\Pi(Z^*|A)}+CmnI\right)\geq 0,$$

for some constant C. By Theorem 4.2.1, we have $\log\Pi(Z^*|A)\geq C$ for some constant C with high probability. To conclude, there exists some constants C_3, C_4 and C_5 such that for any $Z\in S_\alpha$,

$$\min_{Z\in S_\alpha}\left(\log\Pi(Z|A)+C_3mnI\right)\geq 0.$$

A.1.8 Bounding probability of events

We first introduce some notations. For any $Z\in S_\alpha$ and any $a\in[K]$, we use n'_a to denote $n_a(Z)$, m_a to denote $n_a(Z)-R_{aa}$ for simplicity.

Lemma A.1.8. *Denote $X_{ab}(Z)=O_{ab}(Z)-\mathbb{E}\left[O_{ab}(Z)\right]$ for any $a,b\in[K]$, then for a general $K\geq 2$, we have*

$$\mathbb{P}\left\{\max_{Z\in S_\alpha}\|X(Z)\|_\infty\geq\bar{\varepsilon}n^2(p-q)\right\}\leq\exp(-n),$$

as long as $\bar{\varepsilon}^2nI\to\infty$.

Proof. For $K\geq 2$, we have $\mathrm{Var}\left(O_{ab}(Z)\right)\leq n_{ab}(Z)p\leq\alpha^2n^2p/K^2$. Then, by a union bound and Bernstein inequality, it follows that

$$\begin{aligned}\mathbb{P}\left\{\max_{Z\in S_\alpha}\|X(Z)\|_\infty\geq\bar{\varepsilon}n^2(p-q)\right\}&\leq 2K^2K^n\left(\exp\left(-\frac{\bar{\varepsilon}^2n^2(p-q)^2K^2}{2\alpha^2p}\right)+\exp\left(-\frac{3\bar{\varepsilon}n^2(p-q)}{2}\right)\right)\\&\leq 2K^{n+2}\exp\left(-(1-o(1))\frac{\bar{\varepsilon}^2K^2n^2I}{2\alpha^2}\right)\\&\leq\exp(-n),\end{aligned}$$

for any $\bar{\varepsilon}$ satisfying that $\bar{\varepsilon}^2nI\to\infty$. □

The conclusion of Lemma A.1.8 directly leads to the following lemma.

Lemma A.1.9. *For $K \geq 2$ and any $a, b \in [K]$, let $\widehat{B}_{ab} = O_{ab}(Z)/n_{ab}(Z)$, and $\widetilde{B}_{ab} = \mathbb{E}[O_{ab}(Z)]/n_{ab}(Z)$. Under the event $\mathscr{E}$ defined in* (7)*, we have*

$$\max_{Z \in S_\alpha} \left\| \widehat{B} - \widetilde{B} \right\|_\infty = \max_{Z \in S_\alpha} \max_{a,b \in [K]} \frac{X_{ab}(Z)}{n_{ab}(Z)} \leq \frac{\bar{\varepsilon} n^2 (p-q)}{(K\alpha/n)^2} = C\bar{\varepsilon}(p-q),$$

for some constant C depending on K and α.

We state the following lemmas whose proofs will be given together.

Lemma A.1.10.

$$\mathbb{P}\left\{ \max_{Z \in S_\alpha} \|X(Z) - X(Z^*)\|_\infty - (\alpha+\beta)mn(p-q)/K \geq 0 \right\} \leq n \exp(-(1-o(1))nI/K).$$

Lemma A.1.11.

$$\mathbb{P}\left\{ \max_{Z \in S_\alpha} \|X(Z) - X(Z^*)\|_\infty \geq \bar{\varepsilon} n^2 (p-q) \right\} \leq \exp(-n),$$

as long as $\bar{\varepsilon}^2 nI \to \infty$.

Lemma A.1.12. *For any positive sequences $\gamma \to 0$, $\theta \to 0$ satisfying $\theta^2 \gamma nI \to \infty$, we have*

$$\mathbb{P}\left\{ \max_{Z \in S_\alpha : m > \gamma n} \|X(Z) - X(Z^*)\|_\infty \geq \theta mn(p-q) \right\} \leq \exp(-n).$$

Proofs of Lemmas A.1.10, A.1.11, and A.1.12. Recall that $X_{ab}(Z) - X_{ab}(Z^*) = \Delta\widetilde{O}_{ab} - \mathbb{E}\left[\Delta\widetilde{O}_{ab}\right]$ for $a, b \in [K]$. We first consider the case where $a \neq b$, and it follows that $\Delta\widetilde{O}_{ab} = \sum_{i,j \in S_1} A_{ij} - \sum_{i,j \in S_2} A_{ij}$, where

$$|S_1| = \sum_{i,j \in [n]} \left(\mathbb{I}\left\{ Z_i = a, Z_j = b \right\} - \mathbb{I}\left\{ Z_i = Z_i^* = a, Z_j = Z_j^* = b \right\} \right),$$

$$|S_2| = \sum_{i,j\in[n]} \left(\mathbb{I}\left\{ Z_i^* = a, Z_j^* = b \right\} - \mathbb{I}\left\{ Z_i = Z_i^* = a, Z_j = Z_j^* = b \right\} \right).$$

Therefore, we have

$$|S_1| = n'_a n'_b - R_{aa} R_{bb} \le m_a n'_b + m_b n'_a \le \frac{\alpha mn}{K}.$$

Similarly, we have $|S_2| \le \beta mn/K$. Thus, $\mathrm{Var}\left(\Delta\widetilde{O}_{ab}\right) \le (\alpha+\beta)mnp/K$. A similar argument gives that for any $a \in [K]$, $\mathrm{Var}\left(\Delta\widetilde{O}_{ab}\right) \le (\alpha+\beta)mnp/K$ also holds. Then, by a union bound and Bernstein inequality, we have

$$\begin{aligned}
&\mathbb{P}\left\{ \max_{Z\in S_\alpha} \|X(Z) - X(Z^*)\|_\infty - (\alpha+\beta)mn(p-q)/K \ge 0 \right\} \\
\le& 2K^2 \sum_{Z: m < cn/K} \binom{n}{m} (K-1)^m \exp\left(-(1-o(1))mnI/K\right) \\
\le& n \exp(-(1-o(1))nI/K),
\end{aligned}$$

for some constant c, which leads to Lemma A.1.10.

Since $\mathrm{Var}\left(\Delta\widetilde{O}_{ab}\right) \le (\alpha+\beta)n^2 p/K$ for any $a, b \in [K]$, we also have

$$\begin{aligned}
&\mathbb{P}\left\{ \max_{Z\in S_\alpha} \|X(Z) - X(Z^*)\|_\infty \ge \bar{\varepsilon} n^2 (p-q) \right\} \\
\le& 2K^2 K^n \exp\left(-(1-o(1)) \frac{\bar{\varepsilon}^2 K n^2 (p-q)^2}{2(\alpha+\beta)^2 p} \right) \\
\le& \exp(-n),
\end{aligned}$$

as long as $\bar{\varepsilon}^2 nI \to \infty$, which leads to Lemma A.1.11.

For any sequences $\gamma, \theta \to 0$ satisfying $\theta^2 \gamma nI \to \infty$, we also have

$$\mathbb{P}\left\{ \|X(Z) - X(Z^*)\|_\infty \ge \theta mn(p-q) \right\} \le 2K^2 \left(\exp\left(-\frac{\theta^2 m^2 n^2 (p-q)^2}{2(\alpha+\beta)mnp/K} \right) + \exp\left(-\frac{3\theta mn(p-q)}{2} \right) \right).$$

It follows that

$$\begin{aligned}
&\mathbb{P}\left\{\max_{Z\in S_\alpha: m>\gamma n} \|X(Z)-X(Z^*)\|_\infty \le \theta mn(p-q)\right\}\\
\le& C'\sum_{m>\gamma n}\binom{n}{m}K^m\exp(-C\theta^2 mnI)\\
\le& C'\sum_{m>\gamma n}\left(\frac{Ke}{\gamma}\right)^m\exp(-C\theta^2 mnI) \qquad (46)\\
\le& C'\sum_{m>\gamma n}\exp\left(-m\left(C\theta^2 nI-\log\frac{Ke}{\gamma}\right)\right)\\
\le& \exp(-n).
\end{aligned}$$

The last inequality holds since $\theta^2\gamma nI\to\infty$ and thus $\theta^2 nI \gg 1/\gamma \gg \log(1/\gamma)$. It directly leads to Lemma A.1.12. □

Recall that for any $a,a'\in[K]$,

$$X_{aa'}(Z)=O_{aa'}(Z)-\mathbb{E}\left[O_{aa'}(Z)\right],\ \ \Delta O_{aa'}=O_{aa'}(Z)-O_{aa'}(Z'),$$

it follows that

$$X_{aa'}(Z)-X_{aa'}(Z')=\Delta O_{aa'}-\mathbb{E}\left[\Delta O_{aa'}\right].$$

We state the following two lemmas and the proofs will be presented together.

Lemma A.1.13. *Let $N=|\mathscr{B}(Z)\cap S_\alpha|$, and we have*

$$\mathbb{P}\left\{\max_{Z\in S_\alpha}\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}\left[\Delta O_{aa'}\right]|\ge 10n(p-q)\right\}\le n\exp(-2\bar{n}I).$$

Lemma A.1.14. *Let $N=|\mathscr{B}(Z)\cap S_\alpha|$. For any positive sequence $\gamma\to 0$, $\theta\to 0$, and $\theta^2\gamma nI\to\infty$, we have*

$$\mathbb{P}\left\{\max_{Z\in S_\alpha: m>\gamma n}\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}\left[\Delta O_{aa'}\right]|\ge \theta n(p-q)\right\}\le \exp(-n).$$

Proofs of Lemma A.1.13 and Lemma A.1.14. In order to apply Bernstein inequality, we proceed to eliminate the absolute function. Note that $\Delta O_{aa'}$ depends on both the current state Z and the next state Z'. For any $Z \in S_\alpha$, we can rewrite

$$
\begin{aligned}
&\sum_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \sum_{a \le a'} |\Delta O_{aa'} - \mathbb{E}[\Delta O_{aa'}]| \\
&= \max_{h \in \{-1,1\}^{(2K-1)\times N}} \sum_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \left\{ \sum_{a \in [K] \setminus \{b'\}} h_a(Z') (\Delta O_{ab} - \mathbb{E}[\Delta O_{ab}]) + \right. \\
&\quad \left. \sum_{a \in [K] \setminus \{b\}}^{K} h_{a+K-1}(Z') (\Delta O_{ab'} - \mathbb{E}[\Delta O_{ab'}]) + h_{2K-1}(Z') (\Delta O_{bb'} - \mathbb{E}[\Delta O_{bb'}]) \right\} \\
&\triangleq \max_{h \in \{-1,1\}^{(2K-1)\times N}} S(h)
\end{aligned}
\tag{47}
$$

where $h \in \{-1,1\}^{(2K-1)\times N}$ is a matrix whose elements are either 1 or -1. We use $h(Z')$ to denote a column vector of h corresponding to Z', and $h_a(Z')$ is the ath element of $h(Z')$. To prove the equality in (47) holds, we suppose the next state Z' updates one sample from a group b to another group b'. Then, $\Delta O_{aa'} = 0$ for any $a, a' \in [K] \setminus \{b, b'\}$. Hence, for any possible choice of Z', $h(Z') \in \{-1,1\}^{2K-1}$, and then the equality holds.

Claim: for any samples i, j, the random variable A_{ij} appears at most four times in (47). This is because A_{ij} can only appear when we update sample i or sample j. Suppose the sample i is corrected, then A_{ij} appears in $\Delta O_{Z_i, Z_j}$ and $\Delta O_{Z_i', Z_j}$. Thus, the claim holds, which means the same Bernoulli variable appears at most four times in (47).

By the above claim, for any matrix h, it trivially follows that

$$V(h) = \mathrm{Var}(S(h)) \le 16Nnp.$$

Hence, for any $Z \in S_\alpha$ and $w > 0$, by a union bound and Bernstein inequality, we have

$$\begin{aligned}
&\mathbb{P}\left\{\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}[\Delta O_{aa'}]|\ge w\right\}\\
=&\mathbb{P}\left\{\max_{h\in\{-1,1\}^{(2K-1)\times N}} S(h)\ge Nw\right\}\\
\le&\sum_{h\in\{-1,1\}^{(2K-1)\times N}}\mathbb{P}\{S(h)\ge Nw\}\\
\le&2^{2KN}\left(\exp\left(-\frac{N^2w^2}{2V}\right)+\exp\left(-\frac{3Nw}{2\cdot 4}\right)\right),
\end{aligned}$$

where $V = 16Nnp$. Thus, by a union bound, there exists some constant C such that for $w = Cn(p-q)$, we have

$$\begin{aligned}
&\mathbb{P}\left\{\max_{Z\in S_\alpha}\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}[\Delta O_{aa'}]|\ge w\right\}\\
\le&\sum_{m\ge 1}\binom{n}{m}(K-1)^m\mathbb{P}\left\{\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}[\Delta O_{aa'}]|\ge w\right\}\\
\le&\sum_{m\ge 1}\binom{n}{m}(K-1)^m 2^{2KN}\left(\exp\left(-\frac{N^2w^2}{2V}\right)+\exp\left(-\frac{3Nw}{2\cdot 4}\right)\right)\\
\le&n\exp(-2\bar{n}I).
\end{aligned}$$

The last inequality holds since $N \ge \min\{m,\ c_{\alpha,\beta}n\}$ by Lemma A.1.27. It is easy to verify that when $C = 10$, the result holds, which leads to Lemma A.1.13.

For any positive sequences $\gamma, \theta \to 0$ satisfying $\gamma^2\theta nI \to \infty$, let $w = \theta n(p-q)$. Then, it follows

that

$$
\begin{aligned}
&\mathbb{P}\left\{\max_{Z\in S_\alpha: m>\gamma n}\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}[\Delta O_{aa'}]|\ge w\right\}\\
\le&\sum_{m>\gamma n}\binom{n}{m}(K-1)^m\mathbb{P}\left\{\frac{1}{N}\sum_{Z'\in\mathscr{B}(Z)\cap S_\alpha}\sum_{a\le a'}|\Delta O_{aa'}-\mathbb{E}[\Delta O_{aa'}]|\ge w\right\}\\
\le&C'\sum_{m\ge 1}\binom{n}{m}(K-1)^m 2^{2KN}\exp\left(-C\theta^2 NnI\right)\\
\le&\exp(-n),
\end{aligned}
$$

where the last inequality holds by the same argument for (46) and Lemma A.1.27. Thus, the proof of Lemma A.1.14 is complete. □

A.1.9 Proofs of technical lemmas

Lemma A.1.15. *$P,\widetilde{P}$ are the probability measures defined in set Ω. Suppose there exists a subset $A\subset\Omega$ such that $\widetilde{P}(B)=P(B\cap A)/P(A)$ for any set $B\subset\Omega$. Then, we have*

$$
\left\|\widetilde{P}-P\right\|_{\mathsf{TV}}\le 2P(A^c).
$$

Proof. It is obvious that $P(B)=P(B\cap A)+P(B\cap A^c)$. Then,

$$
\begin{aligned}
\left\|\widetilde{P}-P\right\|_{\mathsf{TV}}&=\max_B\left|\frac{P(B\cap A)-P(A)P(B)}{P(A)}\right|\\
&=\max_B\left|\frac{P(B\cap A)P(A^c)-P(A)P(B\cap A^c)}{P(A)}\right|\\
&\le 2P(A^c).
\end{aligned}
$$

□

Lemma A.1.16. *For any positive integers x and y, for any constant $\beta > 0$, we have*

$$\log\frac{\Gamma(x+\beta)}{\Gamma(y+\beta)} \le x\log x - y\log y - (x-y) + \frac{\beta^2}{y+\beta} + (\beta+2)\left|\log\frac{y+\beta}{x+\beta}\right|.$$

Proof. For any positive constants a and b, if $b-a$ is a positive integer, then it is easy to verify that

$$\begin{aligned}
\log\frac{\Gamma(b)}{\Gamma(a)} &= \sum_{k=a}^{b-1}\log k \le b\log b - a\log a - (b-a),\\
\log\frac{\Gamma(b)}{\Gamma(a)} &\ge b\log b - a\log a - (b-a) - \left(\frac{1}{a} - \frac{1}{b} + \log\frac{b}{a}\right).\\
&\ge b\log b - a\log a - (b-a) - 2\log\frac{b}{a}
\end{aligned}$$

Then, for any $a, b \ge 1$, if $a-b$ is an integer, then we have

$$\log\frac{\Gamma(a)}{\Gamma(b)} \le a\log a - b\log b - (a-b) + 2\left(\log\frac{b}{a}\right)_+.$$

Now, let $a = x+\beta$ and $b = y+\beta$. It follows that

$$\begin{aligned}
&\log\frac{\Gamma(x+\beta)}{\Gamma(y+\beta)}\\
\le& x\log(x+\beta) - y\log(y+\beta) - (x-y) + \beta(\log(x+\beta) - \log(y+\beta)) + 2\left(\log\frac{y+\beta}{x+\beta}\right)_+\\
\le& x\log x - y\log y - (x-y) + (\beta+2)\left|\log\frac{y+\beta}{x+\beta}\right| + Err,
\end{aligned}$$

where we write

$$\begin{aligned}
Err =& (x\log(x+\beta) - y\log(y+\beta)) - (x\log x - y\log y)\\
=& x\log\left(\frac{x+\beta}{x}\right) + y\log\left(\frac{y}{y+\beta}\right)\\
\le& \beta - \frac{\beta y}{y+\beta}\\
=& \frac{\beta^2}{y+\beta}.
\end{aligned}$$

Hence, the result follows. □

Lemma A.1.17. *Suppose $x \sim Bernoulli(q)$ and $y \sim Bernoulli(p)$ with $p,q \to 0, p \asymp q$. Then, for any constant C, we have that*

$$\max\left\{\mathbb{E}\left[e^{Ct^*(x-\lambda^*)}\right], \mathbb{E}\left[e^{Ct^*(y-\lambda^*)}\right]\right\} \leq e^{C'I}$$

for some constant C', and t^*, λ^* are defined in (13)*.*

Proof. Since $p,q \to 0$ and $p \asymp q$, we have

$$\log\frac{1-q}{1-p} \leq \log\frac{p}{q} \leq \frac{p}{q} - 1, \quad \log\frac{1-q}{1-p} \geq \frac{p-q}{1-q} \geq p-q.$$

Suppose $C > 0$, and then it follows that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(Ct^*(x-\lambda^*)\right)\right] &= \exp(-Ct^*\lambda^*)\left(q\exp(Ct^*) + 1 - q\right) \\
&\leq \exp(-Ct^*\lambda^* + q\exp(Ct^*) - q) \\
&\leq \exp\left(\frac{-\frac{C}{2}(p-q) + q(p/q)^C - q}{(\sqrt{p}-\sqrt{q})^2}(1-o(1))\cdot I\right) \\
&\leq \exp\left(C'\cdot I\right),
\end{aligned}$$

for some constant C' depending on C. The other cases follow by a similar argument. □

Lemma A.1.18. *For any $a,b \in (0,1)$ and $x,y \in \mathbb{R}$, we have*

$$\left|x\log\frac{a}{b} + (y-x)\log\frac{1-a}{1-b}\right| \leq |x-by|\cdot\left|\log\frac{a(1-b)}{b(1-a)}\right| + |y|\cdot D\left(b\|a\right).$$

Proof. It follows that

$$x\log\frac{a}{b}+(y-x)\log\frac{1-a}{1-b}=(x-by+by)\log\frac{a}{b}+(y-by+by-x)\log\frac{1-a}{1-b}$$
$$=(x-by)\log\frac{a(1-b)}{b(1-a)}-y\cdot D(b\|a).$$

Thus, the result directly follows. □

A.1.10 Proofs of auxiliary lemmas

Lemma A.1.19. *When the connectivity probabilities are unknown, under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_3(\bar{\varepsilon})$ defined in* (6)*, if $p \asymp q$, then we have*

$$\max_{Z\in S_\alpha}\left|\log\frac{\Pi(Z|A)}{\Pi(Z^*|A)}-(Q_{LM}(Z,A)-Q_{LM}(Z^*,A))\right|\le C_{LM}$$

for some constant C_{LM}.

Proof. When the connectivity probabilities are unknown, by (5) we have

$$\log\Pi(Z|A)=\sum_{a\le b}\log\mathrm{Beta}(O_{ab}(Z)+\kappa_1,n_{ab}(Z)-O_{ab}(Z)+\kappa_2)+Const.$$

By Lemma A.1.16, we have

$$\log\frac{\Pi(Z|A)}{\Pi(Z^*|A)} \tag{48}$$
$$\le Q_{LM}(Z,A)-Q_{LM}(Z^*,A)+$$
$$(\kappa_1+\kappa_2)^2\cdot\sum_{a\le b}\left(\frac{1}{O_{ab}(Z^*)+\kappa_1}+\frac{1}{n_{ab}(Z^*)-O_{ab}(Z^*)+\kappa_2}+\frac{1}{n_{ab}(Z)+\kappa_1+\kappa_2}\right)+ \tag{49}$$
$$(\kappa_1+\kappa_2+2)\sum_{a\le b}\left(\left|\log\frac{O_{ab}(Z)+\kappa_1}{O_{ab}(Z^*)+\kappa_1}\right|+\left|\log\frac{n_{ab}(Z)-O_{ab}(Z)+\kappa_2}{n_{ab}(Z^*)-O_{ab}(Z^*)+\kappa_2}\right|+\left|\log\frac{n_{ab}(Z)+\kappa_1+\kappa_2}{n_{ab}(Z^*)+\kappa_1+\kappa_2}\right|\right). \tag{50}$$

Recall that Z^* is the true label assignment, and $\Delta\widetilde{O}_{ab} = O_{ab}(Z) - O_{ab}(Z^*)$ for any $a, b \in [K]$. Under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_3(\bar{\varepsilon})$, we have

$$\max_{Z \in S_\alpha} \max_{a,b} |O_{ab}(Z) - \mathbb{E}[O_{ab}(Z)]| \leq \bar{\varepsilon} n^2 (p-q),$$
$$\max_{Z \in S_\alpha} \max_{a,b} \left| \Delta\widetilde{O}_{ab} - \mathbb{E}\left[\Delta\widetilde{O}_{ab}\right] \right| \leq \bar{\varepsilon} n^2 (p-q).$$

It is easy to check that $\sum_{a \leq b} \frac{1}{O_{ab}(Z^*)+\kappa_1}$ is the dominant term in (49). Thus, we have

$$(49) \leq CK^2 \cdot \frac{1}{pn^2/K^2\alpha^2} \asymp \frac{1}{n^2 p},$$

for some constant C. Since $|\log(a/b)| \leq |a-b|/|\min\{a,b\}|$, by a similar argument, we have

$$(50) \leq C'K^2 \cdot \left(\frac{n^2 p + \bar{\varepsilon} n^2 p}{n^2 p/K^2\alpha^2} + \frac{n^2\alpha^2/K^2}{n^2/K^2\alpha^2} \right) \asymp 1,$$

for some constant C'. By symmetry, the same argument also applies to upper bound $\log \Pi(Z^*|A) - \log \Pi(Z|A)$. Hence, the result of Lemma A.1.2 holds. □

Lemma A.1.20. *When the connectivity probabilities are unknown, under the event $\mathscr{E}_1(\bar{\varepsilon})$ defined in* (6)*, if $p \asymp q$, then we have that*

$$\max_{Z \in S_\alpha} \max_{Z' \in \mathscr{N}(Z)} \left| \log \frac{\Pi(Z|A)}{\Pi(Z'|A)} - \big(Q_{LM}(Z,A) - Q_{LM}(Z',A)\big) \right| \leq \varepsilon_{LM}, \tag{51}$$

for some positive sequence $\varepsilon_{LM} \to 0$.

Proof. Recall the definition of $\mathscr{N}(Z)$ in (1). For any label assignment $Z \in S_\alpha$ and any $Z' \in \mathscr{N}(Z)$,

by Lemma A.1.16, we have

$$\log \frac{\Pi(Z|A)}{\Pi(Z'|A)}$$

$$\leq Q_{LM}(Z,A) - Q_{LM}(Z',A) +$$

$$(\kappa_1+\kappa_2)^2 \cdot \sum_{a\leq b}\left(\frac{1}{O_{ab}(Z')+\kappa_1}+\frac{1}{n_{ab}(Z')-O_{ab}(Z')+\kappa_2}+\frac{1}{n_{ab}(Z)+\kappa_1+\kappa_2}\right)+ \tag{52}$$

$$(\kappa_1+\kappa_2+2)\sum_{a\leq b}\left(\left|\log\frac{O_{ab}(Z)+\kappa_1}{O_{ab}(Z')+\kappa_1}\right|+\left|\log\frac{n_{ab}(Z)-O_{ab}(Z)+\kappa_2}{n_{ab}(Z')-O_{ab}(Z')+\kappa_2}\right|+\left|\log\frac{n_{ab}(Z)+\kappa_1+\kappa_2}{n_{ab}(Z')+\kappa_1+\kappa_2}\right|\right). \tag{53}$$

Under the event $\mathscr{E}_1(\bar{\varepsilon})$ defined in (6), we have

$$\max_{Z\in S_\alpha}\max_{a,b\in[K]}|O_{ab}(Z)-\mathbb{E}[O_{ab}(Z)]|\leq \bar{\varepsilon}n^2(p-q).$$

Since $\max_{a,b\in[K]}\mathbb{E}[O_{ab}(Z)]\leq n_{ab}(Z)p=\mathscr{O}\left(n^2p/K^2\right)$, it follows that under the event $\mathscr{E}_1(\bar{\varepsilon})$, $O_{ab}(Z)=\mathscr{O}\left(n^2p\right)$. Since $\sum_{a\leq b}\frac{1}{O_{ab}(Z')+\kappa_1}$ is the dominant term in (52), there exists some constant C such that

$$(52)\leq \frac{C}{n^2/K^2\cdot p}K^2\lesssim\frac{1}{n^2p}.$$

By Lemma A.1.28, $|\Delta O_{ab}|\leq 2n\alpha/K$ and $|\Delta n_{ab}|\leq 2n\alpha/K$ always hold. Since $|\log(x/y)|\leq|x-y|/\min\{x,y\}$ for $x,y>0$, we have that

$$(53)\leq C'K^2\left(\frac{n}{n^2p/K^2}+\frac{n}{n^2/K^2}\right)\lesssim\frac{1}{np},$$

for some constant C'. Hence, under the event $\mathscr{E}_1(\bar{\varepsilon})$, we have that

$$\max_{Z\in S_\alpha}\max_{Z'\in\mathscr{N}(Z)}\left|\log\frac{\Pi(Z|A)}{\Pi(Z'|A)}-\left(Q_{LM}(Z,A)-Q_{LM}(Z',A)\right)\right|\leq\varepsilon_{LM},$$

for some positive sequence $\varepsilon_{LM} \to 0$. The absolute sign is due to the symmetry. □

Lemma A.1.21. *For a general $K \geq 2$, define $\widetilde{B}_{ab} = \mathbb{E}\left[O_{ab}(Z)\right]/n_{ab}(Z)$, and we have*

$$\frac{\left|\widetilde{B}_{aa'} - B_{aa'}\right|}{p-q} = \begin{cases} \frac{\sum_{k\neq l} R_{ak}R_{al}}{n'_a(n'_a-1)}, & \text{if } a = a', \\ \frac{\sum_k R_{ak}R_{a'k}}{n'_a n'_{a'}}, & \text{if } a \neq a'. \end{cases}$$

It follows that

$$\left\|\widetilde{B} - B\right\|_\infty = \max_{a,b\in[K]} \left|\widetilde{B}_{ab} - B_{ab}\right| \leq \frac{2K\alpha m}{n}(p-q)$$

for some constant C depending on α, β.

Proof. We split the proof of Lemma A.1.21 into two cases and calculate the results based on (10).

Case 1: $a = b$**.** It follows that

$$\frac{\left|\widetilde{B}_{aa} - B_{aa}\right|}{p-q} = \left|\frac{(RBR^T)_{aa} - pn'_a}{n'_a(n'_a-1)} - p\right| \Big/ (p-q) = \frac{\sum_{k\neq l} R_{ak}R_{al}}{n'_a(n'_a-1)} \leq \frac{2m_a}{n'_a} \leq \frac{2K\alpha m_a}{n} \leq \frac{2K\alpha m}{n},$$

where the first inequality holds trivially by analyzing the cases with $m_a = 0$, $m_a = 1$, and $m_a \geq 2$, respectively. The second inequality is by the definition of S_α.

Case 2: $a \neq b$**.** In this case, we have

$$\begin{aligned} \frac{\left|\widetilde{B}_{ab} - B_{ab}\right|}{p-q} &= \left|\frac{(RPR^T)_{ab}}{n'_a \cdot n'_b} - q\right| \Big/ (p-q) = \frac{\sum_k R_{ak}R_{bk}}{n'_a \cdot n'_b} \leq \frac{R_{aa}R_{ba} + R_{bb}R_{ab} + m_a m_b}{n'_a \cdot n'_b} \\ &\leq \frac{n'_a \cdot m_b + n'_b \cdot m_a}{n'_a \cdot n'_b} \leq \frac{K\alpha m_a + K\alpha m_b}{n} \leq \frac{K\alpha m}{n}. \end{aligned}$$

The result simply follows by combining two cases. □

Lemma A.1.22. *Let $\bar{\varepsilon}, \gamma$ be any two positive sequences satisfying that $\gamma, \bar{\varepsilon} \to 0$ and $\bar{\varepsilon}^2 nI \to \infty$. Denote $\widehat{B}_{ab} = O_{ab}(Z)/n_{ab}(Z)$. Under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_2$ defined in (6), for any $Z \in S_\alpha$ with $m \leq \gamma n$, we have*

$$\sum_{a \leq b} \left| \Delta\widetilde{O}_{ab} \log\left(\frac{\widehat{B}_{ab}}{B_{ab}}\right) + (\Delta\widetilde{n}_{ab} - \Delta\widetilde{O}_{ab}) \log\left(\frac{1-\widehat{B}_{ab}}{1-B_{ab}}\right) \right| \lesssim (\bar{\varepsilon}+\gamma) mnI. \tag{54}$$

Proof. By Lemma A.1.18, we have

$$(54) \leq \underbrace{\sum_{a \leq b} \left| \Delta\widetilde{O}_{ab} - B_{ab}\Delta\widetilde{n}_{ab} \right| \cdot \left| \log \frac{\widehat{B}_{ab}(1-B_{ab})}{B_{ab}(1-\widehat{B}_{ab})} \right|}_{(A)} + \underbrace{\sum_{a \leq b} |\Delta\widetilde{n}_{ab}| \cdot D\left(B_{ab} \middle\| \widehat{B}_{ab}\right)}_{(B)}.$$

For any $a, b \in [K]$, we have

$$\begin{aligned}
\left| \Delta\widetilde{O}_{ab} - \Delta\widetilde{n}_{ab} \cdot B_{ab} \right| &\leq \left| \Delta\widetilde{O}_{ab} - \mathbb{E}\left[\Delta\widetilde{O}_{ab}\right] \right| + \left| \mathbb{E}\left[\Delta\widetilde{O}_{ab}\right] - \Delta\widetilde{n}_{ab} \cdot B_{ab} \right| \\
&= |X_{ab}(Z) - X_{ab}(Z^*)| + n_{ab}(Z) \left| \widetilde{B}_{ab} - B_{ab} \right| \\
&\leq \|X(Z) - X(Z^*)\|_\infty + \left(\frac{\alpha n}{K}\right)^2 \cdot \left\| \widetilde{B} - B \right\|_\infty.
\end{aligned}$$

Thus, under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_2$, by Lemma A.1.21, it follows that

$$\max_{a,b} \left| \Delta\widetilde{O}_{ab} - \Delta\widetilde{n}_{ab} \cdot B_{ab} \right| \lesssim mn(p-q).$$

Under the events $\mathscr{E}_1(\bar{\varepsilon})$, by Lemma A.1.9 and Lemma A.1.21, we further have that for any $Z \in S_\alpha$,

$$\left\| \widehat{B} - B \right\|_\infty \leq \left\| \widehat{B} - \widetilde{B} \right\|_\infty + \left\| \widetilde{B} - B \right\|_\infty \lesssim (\bar{\varepsilon}+\gamma)(p-q),$$

and thus it follows that

$$(A) \lesssim (\bar{\varepsilon}+\gamma) mnI.$$

For the term (B), under the events $\mathscr{E}_1(\bar{\varepsilon})$, since $|\Delta\widetilde{n}_{ab}| = |n_{ab}(Z) - n_{ab}(Z^*)| \leq 2mn$ trivially holds,

by bounding the Kullback-Leibler divergence with χ^2-divergence, it follows that

$$(B) \le K^2 \max_{a,b\in[K]} \left\{ |\Delta\widetilde{n}_{ab}| \cdot D\left(B_{ab} \middle\| \widehat{B}_{ab}\right) \right\} \lesssim mn \cdot \frac{\left\|\widehat{B}-B\right\|_\infty^2}{p} \lesssim (\bar{\varepsilon}+\gamma)^2 mnI.$$

By combining (A) and (B), the result directly follows. □

Lemma A.1.23. *Suppose $p \asymp q$. Then, we have*

$$\sum_a n_a(Z) D\left(p \middle\| \widetilde{B}_{aa}\right) \le CnI,$$

for some constant C depending on α.

Proof. Recall that $n'_a = n_a(Z)$. By Lemma A.1.21, since $\widetilde{B}_{aa} \le p$ for any $a \in [K]$, then we have

$$p - \widetilde{B}_{aa} = \frac{\sum_{k\neq l} R_{ak}R_{al}}{n'_a(n'_a-1)}(p-q) \le \frac{2m_a}{n'_a}(p-q).$$

We upper bound the Kullback-Leibler divergence by χ^2-divergence, and it follows that

$$\begin{aligned}
\sum_a n_a(Z) D\left(p \middle\| \widetilde{B}_{aa}\right) &\le \sum_a n'_a \frac{(p-\widetilde{B}_{aa})^2}{q(1-q)} \\
&\le \sum_a n'_a \frac{1}{q(1-q)} \frac{4m_a^2(p-q)^2}{(n'_a)^2} \\
&\le CnI,
\end{aligned}$$

for some constant C depending on α. □

Lemma A.1.24. *Suppose $p,q \to 0$, $p \asymp q$. For any $x,y \in [q,p]$, we have*

$$D(x\|y) \ge \frac{(x-y)^2}{2p}.$$

Proof. Suppose we fix x, and construct

$$f(y) = x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y} - \frac{(x-y)^2}{2p},$$

and

$$f'(y) = (y-x)\left(\frac{1}{y(1-y)} - \frac{1}{p}\right),$$

where $1/y(1-y) > 1/p$ always holds. Thus $f(y) \geq f(x) = 0$. □

Lemma A.1.25. *Let $\gamma \to 0$ be any positive sequence with $\gamma^2 nI \to \infty$. For $m \geq \gamma n$, we have that*

$$\sum_{a,b,k,l} R_{ak}R_{bl}D\left(B_{kl}\middle\|\widetilde{B}_{ab}\right) \geq C(\alpha,\beta,K)mnI,$$

for some constant C depending on α,β,K.

Proof. By Lemma A.1.24, $D(x\|y) \geq \frac{(x-y)^2}{2p}$. Then, we have

$$\begin{aligned}
&\sum_{a,b,k,l} R_{ak}R_{bl}D\left(B_{kl}\middle\|\widetilde{B}_{ab}\right) \\
\geq &\sum_{a}\sum_{k,l} R_{ak}R_{al}D\left(B_{kl}\middle\|\widetilde{B}_{aa}\right) \\
\geq &\frac{1}{2p}\sum_{a}\sum_{k} R_{ak}^2(p-\widetilde{B}_{aa})^2 + \frac{1}{2p}\sum_{a}\sum_{k\neq l} R_{ak}R_{al}[(p-q)-(p-\widetilde{B}_{aa})]^2. \qquad (55)
\end{aligned}$$

The first inequality holds since we only keep the terms with $b = a$. The second inequality is using Lemma A.1.24. By Lemma A.1.21, it follows that

$$p - \widetilde{B}_{aa} = \frac{\sum_{k\neq l} R_{ak}R_{al}}{n'_a(n'_a-1)}(p-q).$$

For simplicity, let

$$T_a = n'_a(n'_a - 1), \quad B_a = \sum_{k \neq l} R_{ak} R_{al}, \quad \sum_k R_{ak}^2 = T_a + n'_a - B_a.$$

Then we have that

$$\begin{aligned}
(55) =& \frac{(p-q)^2}{2p} \sum_a \frac{1}{T_a^2} \left((T_a + n'_a - B_a) B_a^2 + B_a (T_a - B_a)^2 \right) \\
=& \frac{(p-q)^2}{2p} \sum_a \frac{1}{T_a^2} \left(T_a (T_a - B_a) B_a + n'_a B_a^2 \right) \\
\geq& \frac{(p-q)^2}{2p} \sum_a \frac{B_a (T_a - B_a)}{T_a}.
\end{aligned}$$

Let $x = \sum_{k \neq a} R_{ak}^2$, and thus $0 \leq x \leq m_a^2$. Then, we can write

$$\begin{aligned}
B_a &= 2R_{aa} m_a + m_a^2 - x, \\
T_a - B_a &= R_{aa}^2 + x - n'_a.
\end{aligned}$$

Since $B_a(T_a - B_a)$ is a quadratic function of x that is concave, it follows that

$$(55) \geq \frac{(p-q)^2}{2p} \sum_a \frac{1}{T_a} \min \left\{ 2R_{aa} m_a \left(R_{aa}^2 + m_a^2 - n'_a \right), \ (2R_{aa} m_a + m_a^2)(R_{aa}^2 - n'_a) \right\}.$$

Claim: there exists an $a' \in [K]$ such that $R_{a'a'} \geq Cn$ and $m_{a'} \geq C'm$ for some constants C and C'. This is because of the following argument. Since $m = \sum_a m_a$, there must exist some a such that $m_a \geq m/K$. Without loss of generality, suppose $m_1 \geq m/K$. Then, there are two cases we need to consider next. Case 1: if $R_{11} \geq \frac{n}{2\beta K^2}$, then we take $a' = 1$. Case 2: if $R_{11} < \frac{n}{2\beta K^2}$, since $n_1 \geq \frac{n}{K\beta}$, it follows that $\sum_{i \neq 1} R_{i1} \geq \frac{(2K-1)n}{2K^2\beta}$. Then, there must exists some $a \neq 1$ such that $R_{a1} \geq \frac{(2K-1)n}{2K^2(K-1)\beta}$. Without loss of generality, suppose $R_{21} \geq \frac{(2K-1)n}{2K^2(K-1)\beta} \geq \frac{n}{K^2\beta}$. Then, we have $m_2 \geq R_{21} \geq \frac{n}{K^2\beta} \geq \frac{m}{K^2\beta}$, and thus $R_{22} \geq R_{21} - R_{11} > \frac{n}{2K^2\beta}$ by the definition of discrepancy matrix R. Then, we take $a' = 2$. Hence, the claim always holds.

Based on the above claim, we have that

$$(55) \geq \frac{(p-q)^2}{2p}\frac{1}{T_{a'}}\min\left\{2R_{a'a'}m_a\left(R^2_{a'a'}+m^2_{a'}-n'_{a'}\right),\ (2R_{a'a'}m_{a'}+m^2_{a'})(R^2_{a'a'}-n'_{a'})\right\}$$
$$\gtrsim mnI.$$

The proof is complete. □

Lemma A.1.26. *Recall that*

$$\Delta(\cdot)=\sum_{a\leq b}n_{ab}(\cdot)\left(\tau\left(\frac{O_{ab}(\cdot)}{n_{ab}(\cdot)}\right)-\tau\left(\frac{\mathbb{E}\left[O_{ab}(\cdot)\right]}{n_{ab}(\cdot)}\right)\right),$$

where $\tau(x)=x\log x+(1-x)\log(1-x)$. *For any positive sequences* $\bar{\varepsilon}=\gamma\to 0,\ \theta\to 0$ *with* $\gamma^2 nI\to\infty$ *and* $\theta^2\gamma nI\to\infty$, *under the events* $\mathscr{E}_1(\bar{\varepsilon})$ *and* $\mathscr{E}_4(\gamma,\theta)$ *defined in* (6), *we have that for any* $Z\in S_\alpha$ *with* $m>\gamma n$,

$$|\Delta(Z)-\Delta(Z^*)|\leq\varepsilon mnI,$$

for some positive sequence $\varepsilon\to 0$.

Proof. Recall that for any $a,b\in[K]$, $\widehat{B}_{ab}(Z)=O_{ab}(Z)/n_{ab}(Z)$ and $\widetilde{B}_{ab}(Z)=\mathbb{E}\left[O_{ab}(Z)\right]/n_{ab}(Z)$. Then, $\widetilde{B}_{ab}(Z^*)=B_{ab}$. Note that $\tau'(x)=\log\frac{x}{1-x}$, and $\tau''(x)=\frac{1}{x(1-x)}$. By Taylor expansion, it follows that for any $Z\in S_\alpha$ and any $a,b\in[K]$, there exists some $\xi_{ab}(Z)\in[\widehat{B}_{ab}(Z),\widetilde{B}_{ab}(Z)]$ such that

$$\Delta(Z)=\sum_{a\leq b}\tau'\left(\widetilde{B}_{ab}(Z)\right)\cdot X_{ab}(Z)+\underbrace{\sum_{a\leq b}\frac{X_{ab}(Z)^2}{2n_{ab}(Z)}\cdot\frac{1}{\xi_{ab}(Z)(1-\xi_{ab}(Z))}}_{Err(Z)}.$$

Similarly, since $\widetilde{B}_{ab}(Z^*) = B_{ab}$, we have $\xi_{ab}(Z^*) \in [\widehat{B}_{ab}(Z^*), B_{ab}]$ such that

$$\Delta(Z^*) = \sum_{a \le b} \tau'(B_{ab}) \cdot X_{ab}(Z^*) + Err(Z^*).$$

We write $\widetilde{B}(Z) = \widetilde{B}$ for simplicity. Then, we have

$$\begin{aligned}
&|\Delta(Z) - \Delta(Z^*)| \\
&= \left| \sum_{a \le b} \tau'(\widetilde{B}_{ab}) X_{ab}(Z) + \sum_{a \le b} \tau'(B_{ab}) X_{ab}(Z^*) + Err(Z) - Err(Z^*) \right| \\
&= \underbrace{\sum_{a \le b} \left| \tau'(\widetilde{B}_{ab}) - \tau'(B_{ab}) \right| |X_{ab}(Z)|}_{(A)} + \underbrace{\left| \sum_{a \le b} \tau'(B_{ab}) \left(X_{ab}(Z) - X_{ab}(Z^*) \right) \right|}_{(B)} + \underbrace{|Err(Z) - Err(Z^*)|}_{(C)},
\end{aligned}$$

and we proceed to bound each term.

Under the event $\mathscr{E}_1(\gamma)$, by Lemma A.1.21, we have that for any $Z \in S_\alpha$,

$$(A) \le \sum_{a \le b} \left| \log \frac{\widetilde{B}_{ab}(1 - B_{ab})}{B_{ab}(1 - \widetilde{B}_{ab})} \right| \cdot \|X(Z)\|_\infty \lesssim \frac{\left\| \widetilde{B} - B \right\|_\infty}{p} \cdot \|X(Z)\|_\infty \lesssim \gamma mnI.$$

Under the event $\mathscr{E}_4(\gamma, \theta)$, we have that for any $Z \in S_\alpha$ with $m > \gamma n$,

$$\begin{aligned}
(B) &= \left| \tau'(p) \sum_a (X_{aa}(Z) - X_{aa}(Z^*)) + \tau'(q) \sum_{a<b} (X_{ab}(Z) - X_{ab}(Z^*)) \right| \\
&= \left| (\tau'(p) - \tau'(q)) \cdot \sum_a (X_{aa}(Z) - X_{aa}(Z^*)) \right| \\
&\le \log \frac{p(1-q)}{q(1-p)} \cdot K \|X(Z) - X(Z^*)\|_\infty \\
&\lesssim \theta mnI,
\end{aligned}$$

where the second equality holds since $\sum_{a \le b} O_{ab}(Z) = \sum_{a \le b} O_{ab}(Z^*)$, and thus $\sum_{a \le b} (X_{ab}(Z) -$

$X_{ab}(Z^*)) = 0$ always holds. Under the event $\mathscr{E}_1(\gamma)$, we have that for any $Z \in S_\alpha$,

$$\begin{aligned}(C) &\le 2 \max_{Z \in S_\alpha} |Err(Z)| \le \max_{Z \in S_\alpha} \sum_{a \le b} \left| \frac{X_{ab}(Z)^2}{n_{ab}(Z)} \cdot \frac{1}{\xi_{ab}(Z)(1-\xi_{ab}(Z))} \right| \\ &\lesssim K^2 \frac{\|X(Z)\|_\infty}{n^2} \cdot \frac{1}{p} \lesssim \gamma^2 n^2 I.\end{aligned}$$

Since $m > \gamma n$, it follows that under the events $\mathscr{E}_1(\gamma)$ and $\mathscr{E}_4(\gamma, \theta)$, there exists some sequence $\varepsilon \to 0$, such that for any $Z \in S_\alpha$ with $m > \gamma n$,

$$|\Delta(Z) - \Delta(Z^*)| \le \varepsilon m n I.$$

The proof is complete. □

Lemma A.1.27. *Recall that $\mathscr{B}(Z)$ is defined in* (1)*. We have*

$$|S_\alpha \cap \mathscr{B}(Z)| \ge \min\left\{ \frac{n}{\beta K} - \frac{n}{\alpha K} - 1, \frac{\alpha n - \beta n}{K} - 1, m \right\} \ge \min\{c_{\alpha,\beta} n, m\},$$

for some constant $c_{\alpha,\beta}$.

Proof. We split the proof of Lemma A.1.27 into three cases.

Case 1. Suppose there are totally k' groups with size $\lceil n/K\alpha \rceil$ (reaching the small size boundary), denoted as the set $\mathscr{K}'$, and $|\mathscr{K}'| = k'$. Then,

$$\begin{aligned}k' \lceil \frac{n}{K\alpha} \rceil = \sum_{a \in \mathscr{K}'} n'_a &= \sum_{a \in \mathscr{K}'} R_{aa} + \sum_{a,b \in \mathscr{K}', a \ne b} R_{ab} + \sum_{a \in \mathscr{K}', b \notin \mathscr{K}'} R_{ab} \\ &= \sum_{a \in \mathscr{K}'} \left(n_a - \sum_{b \in \mathscr{K}' \setminus \{a\}} R_{ba} - \sum_{b \notin \mathscr{K}'} R_{ba} \right) + \sum_{a,b \in \mathscr{K}', a \ne b} R_{ab} + \sum_{a \in \mathscr{K}', b \notin \mathscr{K}'} R_{ab} \\ &= \sum_{a \in \mathscr{K}'} n_a - \sum_{a \in \mathscr{K}'} \sum_{b \notin \mathscr{K}'} R_{ba} + \sum_{a \in K', b \notin \mathscr{K}'} R_{ab} \\ &\ge k' \frac{n}{K\beta} - \sum_{a \notin \mathscr{K}', b \in \mathscr{K}'} R_{ab},\end{aligned}$$

and thus $|\mathscr{B}(Z)\cap S_\alpha| \geq \sum_{a\notin\mathscr{K}',b\in\mathscr{K}'} R_{ab} \geq k'\frac{n}{K\beta} - \lceil\frac{n}{K\alpha}\rceil \geq \frac{n}{K\beta} - \frac{n}{K\alpha} - 1$.

Case 2. If there is at least one group with size $\lfloor \alpha n/K \rfloor$ (reaching the large size boundary), denoted as group a. It directly follows that

$$m_a = n'_a - R_{aa} \geq n'_a - n_a \geq \lfloor\frac{\alpha n}{K}\rfloor - \frac{\beta n}{K} \geq \frac{\alpha n}{K} - \frac{\beta n}{K} - 1.$$

Case 3. If $\mathscr{B}(Z)\subset S_\alpha$, then $|S_\alpha\cap\mathscr{B}(Z)| = |\mathscr{B}(Z)| = m$.

□

Lemma A.1.28. *Suppose the current label assignment is Z, and Z' corrects the kth sample from a misclassified group b to its true group b'. Then, for any $a,a' \in [K]\setminus\{b,b'\}$, we have*

$$\begin{aligned}
\Delta O_{ab} &= \sum_{i,j} A_{ij}\left(\mathbb{I}\left\{Z_i = a, Z_j = b\right\} - \mathbb{I}\left\{Z'_i = a, Z'_j = b\right\}\right) = \sum_i A_{ik}\mathbb{I}\left\{Z_i = a\right\},\\
\Delta O_{ab'} &= \sum_{i,j} A_{ij}\left(\mathbb{I}\left\{Z_i = a, Z_j = b'\right\} - \mathbb{I}\left\{Z'_i = a, Z'_j = b'\right\}\right) = -\sum_i A_{ik}\mathbb{I}\left\{Z_i = a\right\},\\
\Delta O_{bb} &= \sum_{i<j} A_{ij}\left(\mathbb{I}\left\{Z_i = Z_j = b\right\} - \mathbb{I}\left\{Z'_i = Z'_j = b\right\}\right) = \sum_i A_{ik}\mathbb{I}\left\{Z'_i = b\right\},\\
\Delta O_{b'b'} &= \sum_{i<j} A_{ij}\left(\mathbb{I}\left\{Z_i = Z_j = b'\right\} - \mathbb{I}\left\{Z'_i = Z'_j = b'\right\}\right) = -\sum_i A_{ik}\mathbb{I}\left\{Z_i = b'\right\},\\
\Delta O_{bb'} &= \sum_{i,j} A_{ij}\left(\mathbb{I}\left\{Z_i = b, Z_j = b'\right\} - \mathbb{I}\left\{Z'_i = b, Z'_j = b'\right\}\right) = -\Delta O_{b'b'} - \Delta O_{bb},\\
\Delta O_{aa'} &= 0.
\end{aligned}$$

Lemma A.1.29. *Recall that $\widetilde{B}_{aa'} = \mathbb{E}\left[O_{aa'}(Z)\right]/n_{aa'}(Z)$ for any $a,a' \in [K]$. When $K = 2$, if $(n-2m)(n-2\beta m)/n \to \infty$, we have that*

$$\widetilde{B}_{11} - \widetilde{B}_{12} = \frac{\det(R)(R_{11}-R_{12})(p-q)}{n_1(n_1-1)n_2}(1-o(1)),$$

and by symmetry,

$$\widetilde{B}_{22} - \widetilde{B}_{21} = \frac{\det(R)(R_{22} - R_{21})(p-q)}{n_2(n_2-1)n_1}(1-o(1)).$$

Proof. We use b to denote one group label, and use b' to denote the other. By Lemma A.1.21,

$$\begin{aligned}\widetilde{B}_{b'b'} - \widetilde{B}_{b'b} &= \left(p - \frac{\sum_{k\neq l} R_{b'k}R_{b'l}}{n'_{b'}(n'_{b'}-1)}(p-q)\right) - \left(q + \frac{\sum_k R_{bk}R_{b'k}}{n'_b \cdot n'_{b'}}(p-q)\right) \\ &= \frac{\det(R)(R_{b'b'} - R_{b'b})(p-q)}{n'_{b'}(n'_{b'}-1)n'_b} - \frac{(R_{bb}R_{b'b'} + R_{bb'}R_{b'b})(p-q)}{n'_{b'}(n'_{b'}-1)n'_b},\end{aligned}$$

where $R_{b'b'} - R_{b'b} = n'_{b'} - m \geq n/2\beta - m$. By Lemma A.1.31, $\det(R) \gtrsim n(n-2m)$. Since $R_{bb}R_{b'b'} + R_{bb'}R_{b'b} \lesssim n^2$, under the condition that $(n-2m)(n-2\beta m)/n \to \infty$, the second term is negligible compared to the first term. □

Lemma A.1.30. *For any $Z \in S_\alpha$, suppose Z' corrects one sample from a misclassified group b to its true group b'. Recall $P(Z,Z')$ is defined in* (36)*. Then, we have*

$$\begin{aligned}&P(Z,Z') \\ &= -\sum_{a,l=1}^{K} \widetilde{R}_{al}\left(B_{lb'}\log\frac{\widetilde{B}_{ab'}}{\widetilde{B}_{ab}} + (1-B_{lb'})\log\frac{1-\widetilde{B}_{ab'}}{1-\widetilde{B}_{ab}}\right) \\ &= -\frac{p-q}{n_{b'}(Z)}\sum_{a\in[K]}\log\frac{\widetilde{B}_{ab'}(1-\widetilde{B}_{ab})}{\widetilde{B}_{ab}(1-\widetilde{B}_{ab'})}\left(\sum_{k\neq b'} R_{b'k}(R_{ab'} - R_{ak})\right) - \sum_{a\in[K]}(n_a(Z) - \delta_{ab'})D\left(\widetilde{B}_{ab'}\middle\|\widetilde{B}_{ab}\right),\end{aligned}$$

where $\widetilde{R}_{b'b'} = R_{b'b'} - 1$, otherwise $\widetilde{R}_{ab} = R_{ab}$. Here, $\delta_{ab'} = 1$ when $a = b'$, otherwise $\delta_{ab'} = 0$.

Proof. We write $P(Z,Z') = \sum_{a,a'} P_{a,a'}(Z,Z')$. Since Z' updates one sample from a misclassified group b to its true group b', by Lemma A.1.28, we have for $a \in [K]\setminus\{b,b'\}$,

$$P_{a,b}(Z,Z') + P_{a,b'}(Z,Z') = -\left(\mathbb{E}[\Delta O_{ab}]\log\frac{\widetilde{B}_{ab'}}{\widetilde{B}_{ab}} + (\Delta n_{ab} - \mathbb{E}[\Delta O_{ab}])\log\frac{1-\widetilde{B}_{ab'}}{1-\widetilde{B}_{ab}}\right),$$

where $\Delta n_{ab} = n_a(Z) = n_a(Z')$, and $\mathbb{E}[\Delta O_{ab}] = \sum_{l\in[K]} R_{al}B_{lb'}$. Then, it follows that

$$P_{a,b}(Z,Z') + P_{a,b'}(Z,Z') = -\left(\log\frac{\widetilde{B}_{ab'}}{\widetilde{B}_{ab}}\sum_{l\in[K]} R_{al}B_{lb'} + \log\frac{1-\widetilde{B}_{ab'}}{1-\widetilde{B}_{ab}}\sum_{l\in[K]} R_{al}(1-B_{lb'})\right).$$

Furthermore, we have

$$\begin{aligned}P_{b,b}(Z,Z') + P_{b',b'}(Z,Z') + P_{b,b'}(Z,Z') = &-\left(\mathbb{E}[\Delta O_{bb}]\log\frac{\widetilde{B}_{bb'}}{\widetilde{B}_{bb}} + (\Delta n_{bb} - \mathbb{E}[\Delta O_{bb}])\log\frac{1-\widetilde{B}_{bb'}}{1-\widetilde{B}_{bb}}\right)\\ &-\left(\mathbb{E}[\Delta O_{b'b'}]\log\frac{\widetilde{B}_{bb'}}{\widetilde{B}_{b'b'}} + (\Delta n_{b'b'} - \mathbb{E}[\Delta O_{b'b'}])\log\frac{1-\widetilde{B}_{bb'}}{1-\widetilde{B}_{b'b'}}\right),\end{aligned}$$

where $\Delta n_{bb} = n_b(Z') = n_b(Z) - 1$, $\Delta n_{b'b'} = -n_{b'}(Z)$, $\mathbb{E}[\Delta O_{bb}] = \sum_l R_{bl}B_{lb'} - B_{b'b'}$, and $\mathbb{E}[\Delta O_{b'b'}] = \sum_l R_{b'l}B_{lb'}$. It follows that

$$\begin{aligned}&P_{b,b}(Z,Z') + P_{b',b'}(Z,Z') + P_{b,b'}(Z,Z')\\ =&-\sum_{a\in\{b,b'\}}\left(\log\frac{\widetilde{B}_{ab'}}{\widetilde{B}_{ab}}\sum_l R_{al}B_{lb'} + \log\frac{1-\widetilde{B}_{ab'}}{1-\widetilde{B}_{ab}}\sum_l R_{al}(1-B_{lb'})\right) &(56)\\ &+B_{b'b'}\log\frac{\widetilde{B}_{b'b'}}{\widetilde{B}_{bb}} + (1-B_{b'b'})\log\frac{1-\widetilde{B}_{b'b'}}{1-\widetilde{B}_{bb}}. &(57)\end{aligned}$$

Thus we have

$$P(Z,Z') = -\sum_{a,l=1}^{K}\widetilde{R}_{al}\left(B_{lb'}\log\frac{\widetilde{B}_{ab'}}{\widetilde{B}_{ab}} + (1-B_{lb'})\log\frac{1-\widetilde{B}_{ab'}}{1-\widetilde{B}_{ab}}\right),$$

where $\widetilde{R}_{b'b'} = R_{b'b'} - 1$, otherwise $\widetilde{R}_{aa'} = R_{aa'}$. It leads to the first equality of Lemma A.1.30.

By some calculations, it follows that

$$-P(Z,Z') = \underbrace{\sum_{a,l\in[K]}\widetilde{R}_{al}\left((B_{lb'} - \widetilde{B}_{ab'})\log\frac{\widetilde{B}_{ab'}}{\widetilde{B}_{ab}} + (\widetilde{B}_{ab'} - B_{lb'})\log\frac{1-\widetilde{B}_{ab'}}{1-\widetilde{B}_{ab}}\right)}_{H_1} + \sum_{a,l\in[K]}\widetilde{R}_{al}D\left(\widetilde{B}_{ab'}\middle\|\widetilde{B}_{ab}\right),$$

where H_1 can be written as

$$H_1 = \sum_{a\in[K]} \log \frac{\widetilde{B}_{ab'}(1-\widetilde{B}_{ab})}{\widetilde{B}_{ab}(1-\widetilde{B}_{ab'})} \sum_{l\in[K]} \widetilde{R}_{al}(B_{lb'} - \widetilde{B}_{ab'}).$$

By Lemma A.1.21, we have for $a \neq b'$,

$$\begin{aligned}
\sum_{l\in[K]} R_{al}(B_{lb'} - \widetilde{B}_{ab'}) &= \sum_{l\in[K]} R_{al}(B_{lb'} - q - \frac{\sum_{k\in[K]} R_{ak}R_{b'k}}{n'_a n'_{b'}}(p-q)) \\
&= R_{ab'}(p-q) - \sum_{l\in[K]} R_{al} \frac{\sum_{k\in[K]} R_{ak}R_{b'k}}{n'_a n'_{b'}}(p-q) \\
&= \frac{p-q}{n'_{b'}} \left(R_{ab'} \sum_{k\in[K]} R_{b'k} - \sum_{k\in[K]} R_{ak}R_{b'k} \right) \\
&= \frac{p-q}{n'_{b'}} \left(\sum_{k\neq b'} R_{b'k}(R_{ab'} - R_{ak}) \right),
\end{aligned}$$

and for $a = b'$, we have

$$\begin{aligned}
\sum_{l\in[K]} \widetilde{R}_{b'l}(B_{lb'} - \widetilde{B}_{b'b'}) &= \sum_{l\in[K]} \widetilde{R}_{b'l} \left(B_{lb'} - q - \frac{\sum_{k\in[K]} R^2_{b'k} - n'_{b'}}{n'_{b'}(n'_{b'}-1)} \right) \\
&= \widetilde{R}_{b'b'}(p-q) - \sum_{l\in[K]} \widetilde{R}_{b'l} \frac{\sum_{k\in[K]} R^2_{b'k} - n'_{b'}}{n'_{b'}(n'_{b'}-1)} \\
&= \frac{p-q}{n'_{b'}} \left((R_{b'b'}-1)n'_{b'} - \sum_{k\in[K]} R^2_{b'k} + n'_{b'} \right) \\
&= \frac{p-q}{n'_{b'}} \left(\sum_{k\neq b'} R_{b'k}(R_{b'b'} - R_{b'k}) \right).
\end{aligned}$$

Thus, the result follows by plugging the results into H_1. □

Lemma A.1.31. *For $K = 2$, when $m \leq n/K\beta$, we have* $\det(R) \geq n(n-2m)/4\alpha$.

Proof. Suppose $n_1 = n_1(Z^*)$, $n_2 = n_2(Z^*)$, and $m = d(Z, Z^*)$. Then, we can write R as

$$R = \begin{pmatrix} n_1 - \left(\frac{m}{2} - x\right) & \frac{m}{2} + x \\ \frac{m}{2} - x & n_2 - \left(\frac{m}{2} + x\right) \end{pmatrix},$$

for some $x \in [-m/2, m/2]$. Therefore,

$$\det(R) = n_1 n_2 - \frac{m}{2} n - x(n_1 - n_2) = (n - n_2)n_2 - x(n - 2n_2) - mn/2.$$

Without loss of generality, we assume $\frac{\beta n}{2} \geq n_1 \geq \frac{n}{2} \geq n_2 \geq \frac{n}{2\beta}$. Then, we have the following conditions that

$$n_1 + 2x = n - n_2 + 2x \leq \frac{n\alpha}{2}, \quad n_2 - 2x \geq \frac{n}{2\alpha}.$$

Combine all these conditions, and it directly follows that

$$\begin{aligned} \det(R) &\geq \frac{n}{2}\left(1 - \frac{1}{\alpha}\right) n_2 + \frac{n^2}{4\alpha} - \frac{mn}{2} \\ &\geq \frac{n}{2}\left(\frac{n}{2\beta} + \frac{n}{2\alpha}\left(1 - \frac{1}{\beta}\right) - m\right) \\ &= \frac{n}{2}\left(\frac{n}{2\beta}\left(1 - \frac{1}{\alpha}\right) + \frac{n}{2\alpha} - m\right) \\ &\geq \frac{n(n-2m)}{4\alpha}, \end{aligned}$$

where the last inequality holds since $m \leq \frac{n}{2\beta}$. □

Lemma A.1.32. *When $K = 2$, for any $Z \in S_\alpha$ with $d(Z, Z^*) = m$, denote $\varepsilon_m = 1 - 2m/n$ for simplicity. If $(n-2m)(n-2\beta m)/n \to \infty$ and $(1-2m/n)^3 nI \to \infty$, then we have*

$$\min_{Z' \in B(Z) \cap S_\alpha} -P(Z, Z') \geq \frac{\varepsilon_m^3 nI}{2\alpha^2} \max\{1 - \beta + 1/\alpha, \varepsilon_m\}(1 - o(1)).$$

Proof. Suppose $Z' \in \mathscr{B}(Z) \cap S_\alpha$ is corrects one sample from a misclassified group b to its true

group b'. Decompose $-P(Z,Z') = (A)+(B)$ by Lemma A.1.30, where

$$\begin{aligned}(A) &= \frac{p-q}{n_{b'}(Z)} \sum_{a\in[K]} \log \frac{\widetilde{B}_{ab'}(1-\widetilde{B}_{ab})}{\widetilde{B}_{ab}(1-\widetilde{B}_{ab'})} \left(\sum_{k\neq b'} R_{b'k}(R_{ab'}-R_{ak}) \right) \\ &= \frac{R_{b'b}(p-q)}{n'_{b'}} \sum_{a\in\{b,b'\}} \left[\log \left(\frac{\widetilde{B}_{ab'}(1-\widetilde{B}_{ab})}{\widetilde{B}_{ab}(1-\widetilde{B}_{ab'})} \right) (R_{ab'}-R_{ab}) \right]\end{aligned}$$

and

$$(B) = \sum_{a\in[K]} (n_a(Z)-\delta_{ab'}) D\left(\widetilde{B}_{ab'} \middle\| \widetilde{B}_{ab}\right)$$

By Lemma A.1.29, when $a = b'$, under the condition that $(n-2m)(n-2\beta m)/n \to \infty$, since $R_{b'b'} - R_{b'b} = n_{b'} - m > 0$, we have

$$\begin{aligned}\log \left(\frac{\widetilde{B}_{b'b'}(1-\widetilde{B}_{b'b})}{\widetilde{B}_{b'b}(1-\widetilde{B}_{b'b'})} \right) (R_{b'b'}-R_{b'b}) &\geq \frac{\widetilde{B}_{b'b'}-\widetilde{B}_{b'b}}{\widetilde{B}_{b'b'}(1-\widetilde{B}_{b'b'})} \cdot (R_{b'b'}-R_{b'b}) \\ &\geq (1-o(1)) \frac{\det(R)(R_{b'b'}-R_{b'b})^2(p-q)}{n'_{b'}(n'_{b'}-1)n'_b} \cdot \frac{1}{\widetilde{B}_{b'b'}(1-\widetilde{B}_{b'b'})} \\ &\geq (1-o(1)) \frac{\det(R)(R_{b'b'}-R_{b'b})^2(p-q)}{n'_{b'}(n'_{b'}-1)n'_b} \cdot \frac{1}{p},\end{aligned}$$

and a similar argument also applies to the case with $a = b$. Hence, it follows that

$$\begin{aligned}(A) &\geq \frac{R_{b'b}(p-q)^2\det(R)}{n'_b} \left(\frac{(R_{b'b'}-R_{b'b})^2}{n'^2_{b'}n'_b} + \frac{(R_{bb}-R_{bb'})^2}{n'^2_b n'_{b'}} \right) (1-o(1)) \\ &\geq \frac{R_{b'b}(p-q)^2\det(R)}{n'_b} \frac{(n-2m)^2}{n'_{b'}n'_b n} (1-o(1)) \\ &\geq \frac{8R_{b'b}(p-q)^2\det(R)(n-2m)^2}{\alpha n^4} (1-o(1)),\end{aligned}$$

where the second inequality holds by Cauchy-Schwarz inequality,

$$\left(\frac{(R_{b'b'}-R_{b'b})^2}{n'_{b'}} + \frac{(R_{bb}-R_{bb'})^2}{n'_b} \right) \cdot (n'_{b'}+n'_b) \geq (R_{b'b'}-R_{b'b}+R_{bb}-R_{bb'})^2 = (n-2m)^2.$$

Third inequalities holds since $n'_{b'}n_{b'} \leq (n'_{b'}+n'_b)^2/4$ and $n'_b \leq \alpha n/2$. We proceed to lower bound

the term (B). By Lemma A.1.24, we have $D(x\|y) \geq (x-y)^2/2p$, and

$$
\begin{aligned}
& n'_{b'} D\left(\widetilde{B}_{b'b'} \middle\| \widetilde{B}_{b'b}\right) + n'_b D\left(\widetilde{B}_{bb'} \middle\| \widetilde{B}_{bb}\right) \\
\geq & \det(R)^2 \frac{(p-q)^2}{2p} \left(\frac{(R_{b'b'} - R_{b'b})^2}{n'^3_{b'} n'^2_b} + \frac{(R_{bb} - R_{bb'})^2}{n'^3_b n'^2_{b'}} \right) (1 - o(1)) \\
\geq & \frac{(p-q)^2}{2p} \det(R)^2 \frac{(n-2m)^2}{n'^2_b n'^2_{b'} n} (1 - o(1)) \\
\geq & \frac{8(p-q)^2 \det(R)^2 (n-2m)^2}{n^5 p} (1 - o(1)).
\end{aligned}
$$

Upper bound Kullback-Leibler divergence by χ^2-divergence, and we have that $D\left(\widetilde{B}_{bb'} \middle\| \widetilde{B}_{bb}\right) \leq CI$ for some constant C. Under the condition that $(1-2m/n)^3 nI \to \infty$, we have

$$
(B) \geq \frac{8(p-q)^2 \det(R)^2 (n-2m)^2}{n^5 p} (1 - o(1)).
$$

It directly follows that

$$
-P(Z, Z') = (A) + (B) \geq \frac{8(p-q)^2 (n-2m)^2 \det(R)}{n^4 p} \left(\frac{R_{b'b}}{\alpha} + \frac{\det(R)}{n} \right) (1 - o(1)). \tag{58}
$$

Note that

$$
\frac{n}{2\alpha} \leq R_{b'b} + R_{b'b'} \leq R_{b'b} + \frac{\beta n}{2} - (m - R_{b'b}),
$$

and thus

$$
R_{b'b} \geq \max\left\{ \frac{m}{2} - \frac{\beta n}{4} + \frac{n}{4\alpha},\ 0 \right\}. \tag{59}
$$

By Lemma A.1.31, we have

$$
\begin{aligned}
-P(Z, Z') &\geq \frac{(p-q)^2 (n-2m)^3}{\alpha^2 n^3 p} \left[\max\left\{ m - \frac{\beta n}{2} + \frac{n}{2\alpha}, 0 \right\} + \frac{n-2m}{2} \right] (1 - o(1)) \\
&= \frac{(p-q)^2 (n-2m)^3}{\alpha^2 n^3 p} \max\left\{ \frac{n}{2}(1 - \beta + 1/\alpha), \frac{n-2m}{2} \right\} (1 - o(1)),
\end{aligned}
$$

where $(p-q)^2/p \geq I$, and thus the result follows. □

Lemma A.1.33. *Under the conditions of Lemma A.1.32, for $Z \in S_\alpha$ with $d(Z,Z') = m$, we have that*

$$-P(Z,Z') \gtrsim \frac{(n-2m)^2 \det(R) I}{n^3}.$$

for any $Z' \in \mathscr{B}(Z) \cap S_\alpha$.

Proof. By (58), (59) and Lemma A.1.31, we have

$$-P(Z,Z') \gtrsim \frac{(n-2m)^2 \det(R) I}{n^3} \max\left\{1-\beta+1/\alpha, 1-2m/n\right\}.$$

Since $m \leq n/2\beta$, then $1-2m/n \to 0$ only if $\beta \to 1$, and thus $1-\beta+1/\alpha$ is a constant. Hence, the result follows. □

Lemma A.1.34. *For $K = 2$, we have*

$$\max_{Z' \in \mathscr{B}(Z) \cap S_\alpha} \max_{a,a' \in \{1,2\}} \left| \mathbb{E}\left[\Delta O_{aa'}\right] - \widetilde{B}_{aa'} \Delta n_{b'b'} \right| \leq C(n-2m)(p-q)$$

for some constant C depending on α.

Proof. Without loss of generality, suppose the current state is Z, and Z' corrects one sample from misclassified group 1 to true group 2. We write $n_1 = n_1(Z^*)$, $n_2 = n_2(Z^*)$ for simplicity. Then, we have

$$R_Z = \begin{pmatrix} n_1 - s & t \\ s & n_2 - t \end{pmatrix}, \quad R_{Z'} = \begin{pmatrix} n_1 - s & t-1 \\ s & n_2 - t + 1 \end{pmatrix},$$

where $t = \sum_{i=1}^n \mathbb{I}\left\{Z_i = 1,\ Z_i^* = 2\right\}$, and $s = \sum_{i=1}^n \mathbb{I}\left\{Z_i = 2,\ Z_i^* = 1\right\}$. Thus, $t+s = m$. By Lemma

A.1.28, we have

$$\begin{aligned}
\Delta O_{11} - \widetilde{B}_{11}\Delta n_{11} &= \frac{-(n_1-s)(n_1-m)}{n_1'}, \\
\Delta O_{22} - \widetilde{B}_{22}\Delta n_{22} &= \frac{-s(n_2-m+1)}{n_2'-1}, \\
\Delta O_{11} - \widetilde{B}_{12}\Delta n_{11} &= \frac{-s(n_1-m)}{n_2'} - \frac{n_1 s + n_2 t - m}{n_1' n_2'}, \\
\Delta O_{22} - \widetilde{B}_{12}\Delta n_{22} &= \frac{-(n_1-s)(n_2-m)}{n_1'},
\end{aligned}$$

and it directly follows that

$$\max_{a,a' \in \{1,2\}} \left| \mathbb{E}\left[\Delta O_{aa'}\right] - \widetilde{B}_{aa'}\Delta n_{aa'} \right| \lesssim (n-2m)(p-q).$$

□

Lemma A.1.35. *For any $Z \in S_\alpha$, write $R_Z(a,b)$ as R_{ab} for simplicity. Then, we have*

$$\sum_{a,b,k,l} R_{ak} R_{bl} D\left(B_{kl} \middle\| \widetilde{B}_{ab}\right) \lesssim mnI.$$

Proof. Since $D(x\|y) \leq \frac{(x-y)^2}{y(1-y)}$ for any $x, y \in (0,1)$, we have

$$\begin{aligned}
\sum_{a,b,k,l} R_{ak} R_{bl} D\left(B_{kl} \middle\| \widetilde{B}_{ab}\right) &\leq \frac{1}{q(1-q)} \sum_{a,b,k,l} R_{ak} R_{bl} (B_{kl} - \widetilde{B}_{ab})^2 \\
&= \frac{1}{q(1-q)} \sum_{a,b,k,l} R_{ak} R_{bl} ((B_{kl} - B_{ab}) + (B_{ab} - \widetilde{B}_{ab}))^2 \\
&\leq \frac{2}{q(1-q)} \sum_{a,b,k,l} R_{ak} R_{bl} \left((B_{kl} - B_{ab})^2 + (B_{ab} - \widetilde{B}_{ab})^2 \right) \\
&= (A) + (B),
\end{aligned}$$

where by Lemma A.1.21,

$$(B) = \frac{1}{q(1-q)} \sum_{a,b,k,l} R_{ak} R_{bl} (\widetilde{B}_{ab} - B_{ab})^2 \leq \frac{2(p-q)^2}{q(1-q)} n^2 \left(\frac{2K\alpha m}{n} \right)^2 \lesssim mnI,$$

and

$$(A) \leq \frac{2(p-q)^2}{q(1-q)} \left(\sum_a \sum_{k \neq l} R_{ak} R_{al} + \sum_{a \neq b} \sum_k R_{ak} R_{bk} \right) \lesssim mnI,$$

since $\sum_a \sum_{k \neq l} R_{ak} R_{al} \leq \sum_a n_a(Z) m_a \leq mn$, and a similar bound holds for $\sum_{a \neq b} \sum_k R_{ak} R_{bk}$. By combining (A) and (B), the proof is complete. □

Lemma A.1.36. *Let $\gamma \to 0$ be any positive sequence. Under the events $\mathscr{E}_1(\bar{\varepsilon})$ and $\mathscr{E}_2$ defined in* (6)*, for any $Z \in S_\alpha$ with $m \leq \gamma n$, we have*

$$\sum_{a<b} n_{ab} \cdot D\left(\frac{O_{ab}}{n_{ab}} \middle\| \frac{O_{ab}(Z)}{n_{ab}(Z)} \right) \leq Cm^2 I$$

for some constant C only depending on K, β, α.

Proof. For any $a, b \in [K]$, we have

$$\begin{aligned} \left| \frac{O_{ab}}{n_{ab}} - \frac{O_{ab}(Z)}{n_{ab}(Z)} \right| &= \left| \frac{X_{ab}(Z^*)}{n_{ab}} - \frac{X_{ab}(Z)}{n_{ab}(Z)} + B_{ab} - \widetilde{B}_{ab} \right| \\ &\leq \underbrace{\left| B_{ab} - \widetilde{B}_{ab} \right|}_{(A)} + \underbrace{\left| \frac{X_{ab}(Z^*)\,(n_{ab}(Z) - n_{ab})}{n_{ab}(Z) \cdot n_{ab}} \right|}_{(B)} + \underbrace{\left| \frac{X_{ab}(Z^*) - X_{ab}(Z)}{n_{ab}(Z)} \right|}_{(C)}. \end{aligned}$$

By Lemma A.1.21, we have

$$(A) \leq \frac{2K\alpha m}{n}(p-q) \asymp \frac{m}{n}(p-q).$$

Under the event $\mathscr{E}_1(\bar{\varepsilon})$, we have

$$(B) \leq \frac{\bar{\varepsilon} n^2 (p-q) \cdot 2mn}{(n/K\alpha)^4} \asymp \bar{\varepsilon} \frac{m}{n}(p-q),$$

where $\bar{\varepsilon}$ is the positive sequence that defines the event $\mathscr{E}_1(\bar{\varepsilon})$. Under the event $\mathscr{E}_2$, we have that

$$(C) \leq \frac{(\alpha+\beta)mn(p-q)/K}{(n/K\alpha)^2} \asymp \frac{m}{n}(p-q).$$

Hence, it follows that

$$\left| \frac{O_{ab}}{n_{ab}} - \frac{O_{ab}(Z)}{n_{ab}(Z)} \right| \lesssim \frac{m}{n}(p-q),$$

for all $Z \in S_\alpha$ with $m \leq \gamma n$. Furthermore, under the event $\mathscr{E}_1(\bar{\varepsilon})$, we have

$$\left| \frac{O_{ab}(Z)}{n_{ab}(Z)} - B_{ab} \right| \leq \left| \frac{X_{ab}(Z)}{n_{ab}(Z)} \right| + \left| \widetilde{B}_{ab} - B_{ab} \right| \lesssim (\bar{\varepsilon} + \gamma)(p-q),$$

and thus $\frac{O_{ab}(Z)}{n_{ab}(Z)} \gtrsim p$. Hence, by $D(x\|y) \leq \frac{(x-y)^2}{y(1-y)}$ for any $x, y \in (0,1)$, we have that,

$$\sum_{a \leq b} n_{ab} \cdot D\left(\frac{O_{ab}}{n_{ab}} \middle\| \frac{O_{ab}(Z)}{n_{ab}(Z)} \right) \lesssim K^2 n^2 \left(\frac{m}{n} \right)^2 \frac{(p-q)^2}{p} \lesssim m^2 I.$$

□

A.2 Proofs in Chapter 4

Lemma A.2.1. *Suppose $Y = X + o_P(1)$, where X, Y are $O_P(1)$. Then, there exists some $\varepsilon_n \to 0$, such that*

$$\left| \frac{\mathbb{E}\mathscr{K}(\frac{Y}{\varepsilon_n})}{\mathbb{E}\mathscr{K}(\frac{X}{\varepsilon_n})} - 1 \right| = o(1),$$

where $\mathscr{K}$ can be indicator kernel or any fixed kernel function with $\mathbb{E}[\nabla \log \mathscr{K}(u)]$ bounded.

Proof. Denote $Y = X + W$, and $W = o_P(1)$. There exists some positive sequence $\eta_n, \xi_n \to 0$, such

that $\mathbb{P}\{\|W\| \geq \eta_n\} \leq \xi_n$. Denote event $\mathscr{E} = \{\|W\| \leq \eta_n\}$. Then, we have

$$\left|\mathbb{E}\left(\mathscr{K}\left(\frac{Y}{\varepsilon_n}\right) - \mathscr{K}\left(\frac{X}{\varepsilon_n}\right)\right)\right| \leq C \cdot \mathbb{P}\{\mathscr{E}^C\} + \left|\mathbb{E}\left(\mathscr{K}\left(\frac{Y}{\varepsilon_n}\right) - \mathscr{K}\left(\frac{X}{\varepsilon_n}\right)\right)\mathbb{I}\{\mathscr{E}\}\right|,$$

where

$$\begin{aligned}
&\left|\mathbb{E}\left(\mathscr{K}\left(\frac{Y}{\varepsilon_n}\right) - \mathscr{K}\left(\frac{X}{\varepsilon_n}\right)\right)\mathbb{I}\{\mathscr{E}\}\right| \\
=&\left|\mathbb{E}\int_{t\in[0,1]} \nabla\mathscr{K}\left(\frac{X+tW}{\varepsilon_n}\right)^T \frac{W}{\varepsilon_n}\mathbb{I}\{\mathscr{E}\}\,dt\right| \\
\leq&\left|\int_{t\in[0,1]} \mathbb{E}\left\|\nabla\mathscr{K}\left(\frac{X+tW}{\varepsilon_n}\right)\right\| \cdot \left\|\frac{W}{\varepsilon_n}\right\|\mathbb{I}\{\mathscr{E}\}\,dt\right| \\
\leq&\frac{1}{\varepsilon_n}\left|\int_{t\in[0,1]} \sqrt{\mathbb{E}\left\|\nabla\mathscr{K}\left(\frac{X+tW}{\varepsilon_n}\right)\right\|^2} \cdot \sqrt{\mathbb{E}\|W\|^2\mathbb{I}\{\mathscr{E}\}}dt\right| \\
\leq&\frac{1}{\varepsilon_n}\left|\int_{t\in[0,1]} \sqrt{\mathbb{E}\left\|\nabla\mathscr{K}\left(\frac{X+tW}{\varepsilon_n}\right)\right\|^2} \cdot \sqrt{\mathbb{E}\|W\|^2\mathbb{I}\{\mathscr{E}\}}dt\right|.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\mathbb{E}\left\|\nabla\mathscr{K}\left(\frac{X+tW}{\varepsilon_n}\right)\right\|^2 &= \varepsilon_n^d \int_u \|\nabla\mathscr{K}(u)\|^2 f_{X+tW}(\varepsilon_n u)du \\
&= \varepsilon_n^d \int_u \|\nabla\log\mathscr{K}(u)\|^2 \mathscr{K}(u)^2 f_{X+tW}(\varepsilon_n u)du \\
&\leq \varepsilon_n^d \int_u \|\nabla\log\mathscr{K}(u)\|^2 \mathscr{K}(u)^2 f_{X+tW}(\varepsilon_n u)du \\
&\leq C\varepsilon_n^d \mathbb{E}\|\nabla\log\mathscr{K}(u)\|^2.
\end{aligned}$$

Hence, we have

$$\left|\mathbb{E}\left(\mathscr{K}\left(\frac{Y}{\varepsilon_n}\right) - \mathscr{K}\left(\frac{X}{\varepsilon_n}\right)\right)\right| \lesssim \varepsilon_n^{d/2-1}\eta_n + \xi_n,$$

and it is clear that $\mathbb{E}[\mathscr{K}(X/\varepsilon_n)] = (1+o(1))f_X(0)\varepsilon_n^d$. It follows that the ratio is at the order

$$\varepsilon_n^{-1-d/2}\eta_n + \varepsilon_n^{-d}\xi_n.$$

Hence, there exists some ε_n such that the result follows. □

REFERENCES

[1] Emmanuel Abbe. Community detection and the stochastic block model. 2016.

[2] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[3] Zhi-Dong Bai and Xuming He. Asymptotic distributions of the maximal depth estimators for regression and mulitvariate location. In *Advances In Statistics*, pages 241–262. World Scientific, 2008.

[4] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

[5] Meïli Baragatti and Pierre Pudlo. An overview on approximate bayesian computation. In *ESAIM: Proceedings*, volume 44, pages 291–299. EDP Sciences, 2014.

[6] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[7] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[8] Nathanaël Berestycki. Mixing times of markov chains: Techniques and examples. *Alea-Latin American Journal of Probability and Mathematical Statistics*, 2016.

[9] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate bayesian computation with the wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019.

[10] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, pages 1922–1943, 2013.

[11] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[12] Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *Foundations of Computer Science, 1997. Proceedings., 38th Annual Symposium on*, pages 223–231. IEEE, 1997.

[13] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[14] Alain Celisse, Jean-Jacques Daudin, Laurent Pierre, et al. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899, 2012.

[15] Niladri S Chatterji, Jelena Diakonikolas, Michael I Jordan, and Peter L Bartlett. Langevin monte carlo without smoothness. *arXiv preprint arXiv:1905.13285*, 2019.

[16] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[17] Mu-Fa Chen and Feng-Yu Wang. Estimation of spectral gap for elliptic operators. *Transactions of the American Mathematical Society*, 349(3):1239–1267, 1997.

[18] Xiang Cheng and Peter Bartlett. Convergence of langevin mcmc in kl-divergence. *arXiv preprint arXiv:1705.09048*, 2017.

[19] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Mmd-bayes: Robust bayesian estimation via maximum mean discrepancy. *arXiv preprint arXiv:1909.13339*, 2019.

[20] Wei Chu, S Sathiya Keerthi, and Chong Jin Ong. Bayesian support vector regression using a unified loss function. *IEEE transactions on neural networks*, 15(1):29–44, 2004.

[21] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[22] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[23] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[24] Tanumay Datta, N Ashok Kumar, Ananthanarayanan Chockalingam, and B Sundar Rajan. A novel mcmc algorithm for near-optimal detection in large-scale uplink mulituser mimo systems. In *2012 Information Theory and Applications Workshop*, pages 69–77. IEEE, 2012.

[25] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[26] Persi Diaconis and Daniel Stroock. Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability*, pages 36–61, 1991.

[27] Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. 2016.

[28] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. *arXiv preprint arXiv:1612.07471*, 2016.

[29] Alain Durmus, Eric Moulines, and Marcelo Pereyra. Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

[30] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

[31] 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.

[32] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[33] David T Frazier, Gael M Martin, Christian P Robert, and Judith Rousseau. Asymptotic properties of approximate bayesian computation. *Biometrika*, 105(3):593–607, 2018.

[34] Chao Gao, Jiyi Liu, Yuan Yao, and Weizhi Zhu. Robust estimation and generative adversarial nets. *arXiv preprint arXiv:1810.02030*, 2018.

[35] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826, 2002.

[36] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[37] Ulf Grenander. Tutorial in pattern theory. *Report, Division of Applied Mathematics*, 1983.

[38] Ulf Grenander and Michael I Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

[39] Yongtao Guan, Stephen M Krone, et al. Small-world mcmc and convergence to multi-modal distributions: From slow mixing to fast mixing. *The Annals of Applied Probability*, 17(1):284–304, 2007.

[40] Venkatesan Guruswami. Rapidly mixing markov chains: a comparison of techniques (a survey). *arXiv preprint arXiv:1603.01512*, 2016.

[41] Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, 28(2):411–425, 2018.

[42] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[43] Jake M Hofman and Chris H Wiggins. Bayesian approach to network modularity. *Physical review letters*, 100(25):258701, 2008.

[44] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[45] Adrian Hutter, James R Wootton, and Daniel Loss. Efficient markov chain monte carlo algorithm for the surface code. *Physical Review A*, 89(2):022326, 2014.

[46] Marko Järvenpää, Michael U Gutmann, Arijus Pleska, Aki Vehtari, Pekka Marttinen, et al. Efficient acquisition rules for model-based approximate bayesian computation. *Bayesian Analysis*, 14(2):595–622, 2019.

[47] Jack Jewson, Jim Q Smith, and Chris Holmes. Principles of bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

[48] Rafail Z Khas' minskii. Ergodic properties of recurrent diffusion processes and stabilization of the solution to the cauchy problem for parabolic equations. *Theory of Probability & Its Applications*, 5(2):179–196, 1960.

[49] Anthony Lee. On the choice of mcmc kernels for approximate bayesian computation with smc samplers. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2012.

[50] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[51] Wentao Li and Paul Fearnhead. On the asymptotic efficiency of approximate bayesian computation estimators. *Biometrika*, 105(2):285–299, 2018.

[52] Frank McSherry. Spectral partitioning of random graphs. In *focs*, page 529. IEEE, 2001.

[53] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[54] Sean P Meyn, Robert L Tweedie, et al. Computable bounds for geometric convergence rates of markov chains. *The Annals of Applied Probability*, 4(4):981–1011, 1994.

[55] Jesper Møller, Anthony N Pettitt, Robert Reeves, and Kasper K Berthelsen. An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.

[56] Ravi Montenegro, Prasad Tetali, et al. Mathematical aspects of mixing times in markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3):237–354, 2006.

[57] Kimia Nadjahi, Valentin De Bortoli, Alain Durmus, Roland Badeau, and Umut Şimşekli. Approximate bayesian computation with the sliced-wasserstein distance. *arXiv preprint arXiv:1910.12815*, 2019.

[58] Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

[59] Mark EJ Newman and Elizabeth A Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569, 2007.

[60] Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

[61] Giorgio Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

[62] Gareth W Peters, Balakrishnan Kannan, Ben Lasscock, Chris Mellen, Simon Godsill, et al. Bayesian cointegrated vector autoregression models incorporating alpha-stable noise for inter-day price movements via approximate bayesian computation. *Bayesian Analysis*, 6(4):755–792, 2011.

[63] Dennis Prangle. Summary statistics in approximate bayesian computation. *arXiv preprint arXiv:1512.05633*, 2015.

[64] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[65] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

[66] Dana Randall. Rapidly mixing markov chains with applications in computer science and physics. *Computing in Science & Engineering*, 8(2):30–41, 2006.

[67] Gareth O Roberts, Jeffrey S Rosenthal, et al. Geometric ergodicity and hybrid markov chains. *Electron. Comm. Probab*, 2(2):13–25, 1997.

[68] Gareth O Roberts, Jeffrey S Rosenthal, et al. General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71, 2004.

[69] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[70] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[71] Peter J Rossky, JD Doll, and HL Friedman. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

[72] Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics surveys*, 8:45, 2014.

[73] Alistair Sinclair. Improved bounds for mixing rates of markov chains and multicommodity flow. *Combinatorics, probability and Computing*, 1(4):351–370, 1992.

[74] Tom AB Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of classification*, 14(1):75–100, 1997.

[75] Yuttapong Thawornwattana, Daniel Dalquen, Ziheng Yang, et al. Designing simple and efficient markov chain monte carlo proposal kernels. *Bayesian Analysis*, 2018.

[76] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[77] Zhuowen Tu and Song-Chun Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):657–673, 2002.

[78] SL van der Pas, J-B Salomond, Johannes Schmidt-Hieber, et al. Conditions for posterior contraction in the sparse normal means problem. *Electronic journal of statistics*, 10(1):976–1000, 2016.

[79] SL van der Pas, AW van der Vaart, et al. Bayesian community detection. *Bayesian Analysis*, 2017.

[80] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[81] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[82] Dawn B Woodard, Jeffrey S Rosenthal, et al. Convergence rate of markov chain methods for genomic motif discovery. *The Annals of Statistics*, 41(1):91–124, 2013.

[83] Keijan Wu, Naoise Nunan, John W Crawford, Iain M Young, and Karl Ritz. An efficient markov chain model for the simulation of heterogeneous soil structure. *Soil Science Society of America Journal*, 68(2):346–351, 2004.

[84] Pan Xu, Jinghui Chen, and Quanquan Gu. Global convergence of langevin dynamics based algorithms for nonconvex optimization. *arXiv preprint arXiv:1707.06618*, 2017.

[85] Yun Yang, Martin J Wainwright, Michael I Jordan, et al. On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

[86] Ming Yu, Varun Gupta, and Mladen Kolar. An influence-receptivity model for topic based information cascades. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1141–1146. IEEE, 2017.

[87] Ming Yu, Varun Gupta, and Mladen Kolar. Learning influence-receptivity network structure with guarantee. *arXiv preprint arXiv:1806.05730*, 2018.

[88] Ming Yu, Varun Gupta, and Mladen Kolar. Estimation of a low-rank topic-based model for information cascades. *arXiv preprint arXiv:1709.01919*, 2019.

[89] 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.

[90] Bumeng Zhuo and Chao Gao. Mixing time of metropolis-hastings for bayesian community detection. *arXiv preprint arXiv:1811.02612*, 2018.