Optimal Exact Recovery in Semi-Supervised Learning: A Study of Spectral Methods and Graph Convolutional Networks

Haixiao Wang [1] **Zhichao Wang** [1]

Abstract

We delve into the challenge of semi-supervised node classification on the Contextual Stochastic Block Model (CSBM) dataset. Here, nodes from the two-cluster Stochastic Block Model (SBM) are coupled with feature vectors, which are derived from a Gaussian Mixture Model (GMM) that corresponds to their respective node labels. With only a subset of the CSBM node labels accessible for training, our primary objective becomes the accurate classification of the remaining nodes. Venturing into the transductive learning landscape, we, for the first time, pinpoint the information-theoretical threshold for the exact recovery of all test nodes in CSBM. Concurrently, we design an optimal spectral estimator inspired by Principal Component Analysis (PCA) with the training labels and essential data from both the adjacency matrix and feature vectors. We also evaluate the efficacy of graph ridge regression and Graph Convolutional Networks (GCN) on this synthetic dataset. Our findings underscore that graph ridge regression and GCN possess the ability to achieve the information threshold of exact recovery in a manner akin to the optimal estimator when using the optimal weighted self-loops. This highlights the potential role of feature learning in augmenting the proficiency of GCN, especially in the realm of semi-supervised learning.

1. Introduction

Graph Neural Networks (GNNs) have emerged as a powerful method for tackling various problems in the domain of graph-structured data, such as social networks, biological networks, and knowledge graphs. The versatility of GNNs allows for applications ranging from node classification to link prediction and graph classification. To explore the mechanism and functionality behind GNNs, it is natural to assume certain data generation models, such that the fundamental limits of certain tasks appear mathematically. In particular, we focus on the synthetic data sampled from *Contextual Stochastic Block Model* (CSBM), first introduced by (Deshpande et al., 2018).

In the literature, the existing work has been focused on the generalization property of GNN (Esser et al., 2021), (Bruna & Li, 2017), (Baranwal et al., 2021), the role of nonlinearity (Lampert & Scholtes, 2023) and oversmoothing phenomenon (Wu et al., 2022). We focus on the fundamental limits of CSBM and the following questions: (1) When do we expect to classify all nodes correctly? (2) What is the best possible accuracy we can achieve when given the parameters of the data generation model? (3) Can we design an efficient estimator to accomplish the previous tasks and how well GNN performs under this evaluation metric? In addressing this challenge, we venture into the *transductive learning* landscape. For the first time, we identify the *information-theoretical* limits required for classifying all nodes correctly for CSBM, specifically in transductive learning setting. This discovery is pivotal as it sets a benchmark for evaluating the performance of various models and algorithms on this type of data for the node classification problem.

Related work

Graph machine learning on CSBM CSBM has become a popular model for theoretical analysis on the performance of GCNs(Kipf & Welling, 2017). For instance, (Wu et al., 2022) studied over-smoothing on CSBM; the generalization performance of GCNs and graph attentions has been considered in (Baranwal et al., 2021; 2023a; Fountoulakis et al., 2023; Chen et al., 2019); the expressivity of GCNs on CSBM (Oono & Suzuki, 2019); bayesian-inference on nonlinear GCNs (Wei et al., 2022; Baranwal et al., 2023b). Although (Huang et al., 2023) analyzed the feature learning on modified CSBM (SNM therein) and (Lu, 2022) showed the learning performance on SBM, currently, there is no complete analysis of the training dynamics for GCNs on

[1]Department of Mathematics, University of California San Diego, San Diego, California, USA. Correspondence to: Haixiao Wang <h9wang@ucsd.edu>, Zhichao Wang <zhw036@ucsd.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

CSBM.

Generalization theory of GCNs Many works have studied the generalization of GCNs, e.g. Bruna & Li (2017) explored the community detection for SBM with GCNs; (Tang & Liu, 2023) (Shi et al., 2024; Duranthon & Zdeborová, 2023) conjectures and algorithms for the constant degree regime. The roles of self-loops and nonlinearity in CCNs have been analyzed in (Lampert & Scholtes, 2023; Kipf & Welling, 2017). Our results in GCNs also provide a way to choose the optimal self-loop weight in GCN to achieve optimal performance.

Semi-supervised linear regression Semi-supervised linear regression has been studied in a line of research work, e.g. (Azriel et al., 2022; Ryan & Culp, 2015; Chakrabortty & Cai, 2018; Tony Cai & Guo, 2020; Belkin et al., 2004). (Lelarge & Miolane, 2019) proves the asymptotic Bayes risks for GMM in semisupervised learning, whereas we extend this to CSBM under perfect recovery setting. (Nguyen & Couillet, 2023) explored asymptotic Bayes risk of semi-supervised multitask learning on Gaussian mixture.

Main contributions

Our contribution lies in the following three perspectives.

- Mathematically, we derive the necessary and sufficient conditions for classifying all nodes correctly under the CSBM. For scenarios where perfect classification is impossible, we characterize the asymptotic misclassification ratio and design efficient estimators to achieve the lowest possible error rate.
- We discover the optimal spectral method for the exact recovery of all unknown labels in this semi-supervised learning which achieves the information theoretical (IT) threshold for SNR.
- At the same time, we evaluate the efficacy of graph ridge regression and GCN on the CSBM dataset to achieve exact recovery. We find the *optimal weight* for self-loop in the graph to attain IT bound. This provides new insight into modifying the architectures of GCNs for optimal performance.

2. Preliminaries

2.1. Node classification

Let $\mathcal{V}$ and $\mathcal{E}$ denote the set of vertices and edges of graph $\mathcal{G}$ respectively, with $|\mathcal{V}| = N \in \mathbb{N}_+$. Assume that $\mathcal{V}$ is composed of two disjoint sets $\mathcal{V}_+, \mathcal{V}_-$, i.e., $\mathcal{V} = \mathcal{V}_+ \cup \mathcal{V}_-$ and $\mathcal{V}_+ \cap \mathcal{V}_- = \emptyset$. Let $\boldsymbol{y} := [y_1, \dots, y_N]^\top \in \{\pm 1\}^N$ denote the label vector encoding the community memberships, i.e., $\mathcal{V}_+ = \{i \in [N] : y_i > 0\}$ and $\mathcal{V}_- = \{i \in [N] : y_i < 0\}$. Assume the access to $\mathcal{G}$ in practice. The goal is to recover the underlying $\boldsymbol{y}$ using the observations. Let $\widehat{\boldsymbol{y}}$ denote some estimator of $\boldsymbol{y}$. To measure the performance of the above estimator, define the *mismatch* ratio between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$ by

$$\psi_N(\boldsymbol{y}, \widehat{\boldsymbol{y}}) := \frac{1}{N} \min_{s \in \{\pm 1\}} |\{i \in [N] : \boldsymbol{y}_i \neq s \cdot \widehat{\boldsymbol{y}}_i\}|.$$

For the symmetric case, $|\mathcal{V}_+| = |\mathcal{V}_-| = N/2$, the *random guess* estimation, i.e., determining the node label by flipping a fair coin, would achieve 50% accuracy on average. An estimator is meaningful only if it outperforms the random guess, i.e., $\psi_N(\boldsymbol{y}, \widehat{\boldsymbol{y}}) \leq 0.5$. If so, $\widehat{\boldsymbol{y}}$ is said to accomplish *weak* recovery. See (Abbe, 2018) for a detailed introduction.

In this paper, we aim to address another interesting scenario when all the nodes can be perfectly classified, i.e., $\psi_N = 0$, which leads to the concept of *exact* recovery.

Definition 2.1 (Exact recovery)**.** *The $\widehat{\boldsymbol{y}}$ is said to achieve the exact recovery (strong consistency) if*

$$\lim_{N \to \infty} \mathbb{P}(\psi_N(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = 0) = \lim_{N \to \infty} \mathbb{P}(\widehat{\boldsymbol{y}} = \pm\, \boldsymbol{y}) = 1.$$

2.2. Contextual Stochastic Block Model

It is natural to embrace certain data generation models to study the mathematical limits of algorithm performance. The following model is in particular of our interests.

Definition 2.2 (Binary Stochastic Block Model, SBM)**.** *Assume $\mathbf{1}^\top \boldsymbol{y} = 0$, i.e., $|\mathcal{V}_+| = |\mathcal{V}_-| = N/2$. Given $0 < \alpha, \beta < 1$, for any pair of node i and j, the edge $\{i, j\} \in \mathcal{E}$ is sampled independently with probability α if $y_i = y_j$, i.e., $\mathbb{P}(A_{ij} = 1) = \alpha$, otherwise $\mathbb{P}(A_{ij} = 1) = \beta$. Furthermore, if $\boldsymbol{A} \in \{0, 1\}^{N \times N}$ is symmetric and $A_{ii} = 0$ for each $i \in [N]$, we then write $\boldsymbol{A} \sim \mathrm{SBM}(\boldsymbol{y}, \alpha, \beta)$.*

For each node $i \in \mathcal{V}$, there is a feature vector $\boldsymbol{x}_i$ attached to it. We are interested in the scenario where $\boldsymbol{x}_i$ is sampled from the following *Gaussian Mixture Model*.

Definition 2.3 (Gaussian Mixture Model, GMM)**.** *Given $N, d \in \mathbb{N}_+$, label vector $\boldsymbol{y} \in \{\pm 1\}^N$ and some fixed $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ with $\|\boldsymbol{\mu}\|_2 = 1$, we write $\{\boldsymbol{x}_i\}_{i=1}^N \sim \mathrm{GMM}(\boldsymbol{\mu}, \boldsymbol{y}, \theta)$ if $\boldsymbol{x}_i = \theta y_i \boldsymbol{\mu} + \boldsymbol{z}_i \in \mathbb{R}^d$ for each $i \in [N]$, where $\theta > 0$ denote the signal strength, and $\{\boldsymbol{z}_i\}_{i=1}^N \subset \mathbb{R}^d$ are i.i.d. random column vectors sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Then by denoting $\boldsymbol{Z} := [\boldsymbol{z}_1, \dots, \boldsymbol{z}_N]^\top$, we re-write $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ as*

$$\boldsymbol{X} := [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N]^\top = \theta \boldsymbol{y} \boldsymbol{\mu}^\top + \boldsymbol{Z}. \tag{2.1}$$

In particular, this paper focuses on the scenario where $\mathcal{G}$ and $\boldsymbol{X}$ are generated in the following manner.

Definition 2.4 (Contextual Stochastic Block Model, CSBM)**.** *Suppose that $N, d \in \mathbb{N}_+$, $0 \leq \alpha, \beta \leq 1$ and $\theta > 0$. We write $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ if*

(a) the label vector $\boldsymbol{y}$ is uniformly sampled from the set $\{\pm 1\}^N$, satisfying $\mathbf{1}^\top \boldsymbol{y} = 0$;
(b) independently, $\boldsymbol{\mu}$ is sampled from uniform distribution over the $\mathbb{S}^{d-1} := \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 = 1\}$;
(c) given $\boldsymbol{y}$, independently, we sample $\boldsymbol{A} \sim \mathrm{SBM}(\boldsymbol{y}, \alpha, \beta)$ and $\boldsymbol{X} \sim \mathrm{GMM}(\boldsymbol{\mu}, \boldsymbol{y}, \theta)$.

CSBM was first introduced in (Deshpande et al., 2018), where a tight analysis for the inference of latent community structure was provided. The *information-theoretic* thresholds on *exact* recovery (Abbe et al., 2022) and weak recovery (Lu & Sen, 2023) were established under the *unsupervised* learning regime, i.e., none of the node labels is revealed. However, the modern learning methods (Kipf & Welling, 2017) performed on the popular datasets (Yang et al., 2016; Bojchevski & Günnemann, 2018; Shchur et al., 2018) rely on the model training procedure, i.e., a fraction of node labels are revealed, which is the regime we will focus on.

2.3. Semi-supervised learning on graph

Assume that $n \in [0, N)$ node labels are revealed, denoted by $y_1, \ldots, y_n$ without loss of generality. Let $\mathbb{L} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ denote the training samples and $\mathbb{U} = \{\boldsymbol{x}_j\}_{j=n+1}^N$ denote the set of feature vectors corresponding to the unrevealed nodes. Each vertex $v \in \mathcal{V}$ is assigned to either $\mathcal{V}_{\mathbb{L}}$ or $\mathcal{V}_{\mathbb{U}}$ depending on the disclosure of its label, where $n := |\mathcal{V}_{\mathbb{L}}|$, $m := |\mathcal{V}_{\mathbb{U}}|$ with $N = n + m$. Let $\tau := n/N$ denote the *training ratio*. For simplicity, let $\boldsymbol{y}_{\mathbb{L}} \in \{\pm 1\}^n$ and $\boldsymbol{y}_{\mathbb{U}} \in \{\pm 1\}^m$ denote the *revealed* and *unrevealed* label vectors. We further denote $\mathcal{V}_{\mathbb{L},+} = \{i \in \mathcal{V}_{\mathbb{L}} : y_i > 0\}$, $\mathcal{V}_{\mathbb{L},-} = \{i \in \mathcal{V}_{\mathbb{L}} : y_i < 0\}$, $\mathcal{V}_{\mathbb{U},+} = \{i \in \mathcal{V}_{\mathbb{U}} : y_i > 0\}$ and $\mathcal{V}_{\mathbb{U},-} = \{i \in \mathcal{V}_{\mathbb{U}} : y_i < 0\}$. For instance in Figure 1, we aim to recover the labels of $\mathcal{V}_{\mathbb{U},+}$ and $\mathcal{V}_{\mathbb{U},-}$ based on known labels in $\mathcal{V}_{\mathbb{L},+}$ and $\mathcal{V}_{\mathbb{L},-}$. Under the *semi-supervised* regime, the graph $\mathcal{G}$ and feature matrix $\boldsymbol{X}$ are generated in the following manner.

Definition 2.5 (Semisupervised CSBM)**.** *Suppose that $\boldsymbol{y}_{\mathbb{L}}$, $\boldsymbol{y}_{\mathbb{U}}$ are uniformly sampled from $\{\pm 1\}^n$, $\{\pm 1\}^m$ respectively, satisfying $\mathbf{1}_n^\top \boldsymbol{y}_{\mathbb{L}} = \mathbf{1}_m^\top \boldsymbol{y}_{\mathbb{U}} = 0$. After concatenating $\boldsymbol{y} = [\boldsymbol{y}_{\mathbb{L}}^\top, \boldsymbol{y}_{\mathbb{U}}^\top]^\top$, we have $(\boldsymbol{A}, \{\boldsymbol{x}_i\}_{i=1}^N)$ sampled from $\mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ in model 2.4.*

Remark 2.6. *It reduces to the unsupervised regime if $n = 0$.*

Let $\mathcal{X} = \mathrm{span}(\{\boldsymbol{x}_i\}_{i=1}^N)$ denote the *feature space* and $\mathcal{Y} = \{\pm 1\}^N$ denote *label* space. In practice, the access to the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the feature vectors $\{\boldsymbol{x}_i\}_{i=1}^N$ and the revealed labels $\boldsymbol{y}_{\mathbb{L}}$ are guaranteed. At this stage, finding a predictor $h : \mathcal{X} \times \mathcal{G} \times \mathcal{Y}_{\mathbb{L}} \mapsto \mathcal{Y}_{\mathbb{U}}$ is our primary interest. Let $\widehat{\boldsymbol{y}}_{\mathbb{U}}$ denote some estimator of $\boldsymbol{y}_{\mathbb{U}}$. The *mismatch* ratio under the semi-supervised regime can be re-written as

$$\psi_m(\boldsymbol{y}_{\mathbb{U}}, \widehat{\boldsymbol{y}}_{\mathbb{U}}) = \frac{1}{m} \min_{s \in \{\pm 1\}} |\{i \in [m] : (\boldsymbol{y}_{\mathbb{U}})_i \neq s(\widehat{\boldsymbol{y}}_{\mathbb{U}})_i\}|.$$

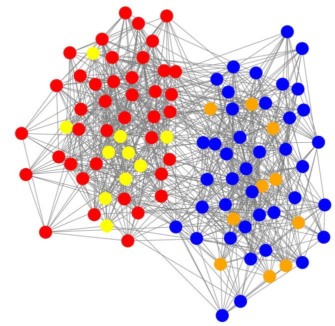


Figure 1. An example of SBM under semi-supervised learning. Red: $\mathcal{V}_{\mathbb{L},+}$; blue: $\mathcal{V}_{\mathbb{L},-}$; yellow: $\mathcal{V}_{\mathbb{U},+}$; and orange $\mathcal{V}_{\mathbb{U},-}$.

2.4. Graph-based transductive learning

To efficiently represent the training and test data, we define the following two *sketching* matrices.

Definition 2.7. *Define $\boldsymbol{S}_{\mathbb{L}} \in \{0,1\}^{n \times N}$, $\boldsymbol{S}_{\mathbb{U}} \in \{0,1\}^{m \times N}$*

$$(\boldsymbol{S}_{\mathbb{L}})_{ij} := \mathbb{1}\{i = j\} \cap \mathbb{1}\{i \in \mathcal{V}_{\mathbb{L}}\},$$
$$(\boldsymbol{S}_{\mathbb{U}})_{ij} := \mathbb{1}\{i = j\} \cap \mathbb{1}\{i \in \mathcal{V}_{\mathbb{U}}\}.$$

Immediately, $\boldsymbol{y}_{\mathbb{L}} = \boldsymbol{S}_{\mathbb{L}}\boldsymbol{y}$, $\boldsymbol{y}_{\mathbb{U}} = \boldsymbol{S}_{\mathbb{U}}\boldsymbol{y}$. Define $\boldsymbol{X}_{\mathbb{L}} := \boldsymbol{S}_{\mathbb{L}}\boldsymbol{X}$, $\boldsymbol{X}_{\mathbb{U}} := \boldsymbol{S}_{\mathbb{U}}\boldsymbol{X}$, then $\boldsymbol{X} = [\boldsymbol{X}_{\mathbb{L}}^\top, \boldsymbol{X}_{\mathbb{U}}^\top]^\top$. The adjacency matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ adapts the following block form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{\mathbb{L}} & \boldsymbol{A}_{\mathbb{L}\mathbb{U}} \\ \boldsymbol{A}_{\mathbb{U}\mathbb{L}} & \boldsymbol{A}_{\mathbb{U}} \end{bmatrix} := \begin{bmatrix} \boldsymbol{S}_{\mathbb{L}}\boldsymbol{A}\boldsymbol{S}_{\mathbb{L}}^\top & \boldsymbol{S}_{\mathbb{L}}\boldsymbol{A}\boldsymbol{S}_{\mathbb{U}}^\top \\ \boldsymbol{S}_{\mathbb{U}}\boldsymbol{A}\boldsymbol{S}_{\mathbb{L}}^\top & \boldsymbol{S}_{\mathbb{U}}\boldsymbol{A}\boldsymbol{S}_{\mathbb{U}}^\top \end{bmatrix}. \quad (2.2)$$

In *inductive* learning, algorithms are unaware of the nodes for testing during the learning stage, i.e., only $\boldsymbol{A}_{\mathbb{L}}$ and $\boldsymbol{X}$ are used for training. The test graph $\boldsymbol{A}_{\mathbb{U}}$ is disjoint from $\boldsymbol{A}_{\mathbb{L}}$ and entirely unseen by the algorithm during the training procedure, since $\boldsymbol{A}_{\mathbb{L}\mathbb{U}}$ is not used either. Notably, this kind of information wastage will reduce the estimator's accuracy.

In contrast, the entire graph $\boldsymbol{A}$ is used for algorithm training under *transductive* learning. The estimator benefits from the message-passing mechanism among seen and unseen nodes.

3. Main results

To state our main results, we start with several basic assumptions. Recall that $\tau := n/N$ denotes the *training ratio*, where $\tau \in (0, 1)$ is some fixed constant.

Assumption 3.1 (Asymptotics)**.** *Let q_m be some function of m and $q_m \to \infty$ as $m \to \infty$. For $\mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ in model 2.4, we assume $\alpha = a \cdot q_m/m$ and $\beta = b \cdot q_m/m$, for some constants $a \neq b \in \mathbb{R}^+$, and*

$$c_\tau := \theta^4[q_m(\theta^2 + (1-\tau)d/m)]^{-1}, \quad (3.1)$$

is a fixed positive constant as $m \to \infty$. Furthermore, we fix $n/N = \tau \in (0, 1)$ as $N, n, m \to \infty$.

For instance, $\tau = 0.2$, $\alpha = 0.3$ and $\beta = 0.05$ in Figure 1. For $a, b \in \mathbb{R}^+$, denote $a_\tau = (1-\tau)^{-1}a$, $b_\tau = (1-\tau)^{-1}b$. Define the following *rate function* by

$$I(a_\tau, b_\tau, c_\tau) \coloneqq [(\sqrt{a_\tau} - \sqrt{b_\tau})^2 + c_\tau]/2, \tag{3.2}$$

which will be applied to our large deviation analysis.

3.1. Information-theoretic limits

Note that $\boldsymbol{y} = [\boldsymbol{y}_{\mathbb{L}}^\top, \boldsymbol{y}_{\mathbb{U}}^\top]^\top$, $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$, we first present the necessary condition for any estimator $\widehat{\boldsymbol{y}}_{\mathbb{U}}$ to reconstruct $\boldsymbol{y}_{\mathbb{U}}$ exactly.

Theorem 3.2 (Impossibility)**.** *Under Assumption 3.1 with $q_m = \log(m)$, as $m \to \infty$, every algorithm will misclassify at least 2 vertices with probability tending to 1 if $I(a_\tau, b_\tau, c_\tau) < 1$.*

We explain the proof sketch above. For the node classification problem, the best estimator is the Maximum Likelihood Estimator (MLE). If MLE fails exact recovery, then no other algorithm could achieve exact recovery. When $I(a_\tau, b_\tau, c_\tau) < 1$, we can prove that with high probability, MLE will not return the true label vector $\boldsymbol{y}_{\mathbb{U}}$, but some other configuration $\widetilde{\boldsymbol{y}}_{\mathbb{U}} \neq \boldsymbol{y}_{\mathbb{U}}$ instead, which leads to the failure of exact recovery. Similar idea showed up in (Abbe et al., 2015; Kim et al., 2018; Wang, 2023) before. On the other hand, the following result concerns the fundamental limits of any algorithm.

Theorem 3.3. *Under Assumption 3.1 with $q_m \gg 1$, any sequence of estimators $\widehat{\boldsymbol{y}}_{\mathbb{U}}$ satisfies*

$$\liminf_{m\to\infty} q_m^{-1} \log \mathbb{E}\psi_m(\boldsymbol{y}_{\mathbb{U}}, \widehat{\boldsymbol{y}}_{\mathbb{U}}) \geq -I(a_\tau, b_\tau, c_\tau).$$

Informally, the result of Theorem 3.3 can be interpreted as $\mathbb{E}\psi_m(\boldsymbol{y}_{\mathbb{U}}, \widehat{\boldsymbol{y}}_{\mathbb{U}}) \geq e^{-I(a_\tau, b_\tau, c_\tau)q_m}$, which gives the lower bound on the expected mismatch ratio for any estimator $\widehat{\boldsymbol{y}}_{\mathbb{U}}$. This rate function $I(a_\tau, b_\tau, c_\tau)$ in (3.2) is derived from the analysis of the *large deviation principle* (LDP) for $\boldsymbol{A}$ and $\boldsymbol{X}$, with details deferred to Lemma A.1.

3.2. Optimal spectral estimator

3.2.1. The construction of spectral estimators

Define the *hollowed Gram* matrix $\boldsymbol{G} = \mathcal{H}(\boldsymbol{X}\boldsymbol{X}^\top) \in \mathbb{R}^{N\times N}$ by $G_{ij} = \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \mathbb{1}_{\{i\neq j\}}$. Similarly, $\boldsymbol{G}$ adapts the block form as in (2.2). Let $\lambda_i(\boldsymbol{A})$, $\lambda_i(\boldsymbol{A}_{\mathbb{U}})$ (resp. $\lambda_i(\boldsymbol{G})$, $\lambda_i(\boldsymbol{G}_{\mathbb{U}})$) denote the i-th largest eigenvalue of $\boldsymbol{A}$, $\boldsymbol{A}_{\mathbb{U}}$ (resp. $\boldsymbol{G}$, $\boldsymbol{G}_{\mathbb{U}}$), and $\boldsymbol{u}_i(\boldsymbol{A}_{\mathbb{U}})$, $\boldsymbol{u}_i(\boldsymbol{G}_{\mathbb{U}})$ are the corresponding unit eigenvectors. Define the index $\ell^\star = 2\cdot\mathbb{1}\{a > b\} + m\cdot\mathbb{1}\{a < b\}$ and the ratio

$$\widehat{\kappa}_{\ell^\star} = \log\Big(\frac{\lambda_1(\boldsymbol{A}_{\mathbb{U}}) + \lambda_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})}{\lambda_1(\boldsymbol{A}_{\mathbb{U}}) - \lambda_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})}\Big).$$

The index $\ell^\star$ is used to differentiate the homophilic $(a > b)$ and heterophilic $(a < b)$ graphs. We then define

$$\widehat{\boldsymbol{y}}_{\mathrm{SBM}} \coloneqq \widehat{\kappa}_{\ell^\star}\big(\frac{1}{\sqrt{m}}\boldsymbol{A}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}} + \lambda_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})\boldsymbol{u}_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})\big)$$

$$\widehat{\boldsymbol{y}}_{\mathrm{GMM}} \coloneqq \frac{2\lambda_1(\boldsymbol{G})}{N\lambda_1(\boldsymbol{G}) + dN}\Big(\frac{\boldsymbol{G}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}}}{\sqrt{m}} + \lambda_1(\boldsymbol{G}_{\mathbb{U}})\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}})\Big).$$

It is natural to discard the graph estimator when $a = b$ reflected by $\widehat{\kappa}_{\ell^\star} = 0$, since no algorithm could outperform random guess on the Erdős-Rényi graph. Consequently, the ideal estimator, inspired by *semi-supervised principal component analysis*, is given by $\mathrm{sign}(\widehat{\boldsymbol{y}}_{\mathrm{PCA}})$ with

$$\widehat{\boldsymbol{y}}_{\mathrm{PCA}} \coloneqq \widehat{\boldsymbol{y}}_{\mathrm{SBM}} + \widehat{\boldsymbol{y}}_{\mathrm{GMM}}, \tag{3.4}$$

where the estimated partitions are collected by $\widehat{\mathcal{V}}_{\mathbb{U},+} \coloneqq \{i \in \mathcal{V}_{\mathbb{U}} : (\widehat{\boldsymbol{y}}_{\mathrm{PCA}})_i > 0\}$ and $\widehat{\mathcal{V}}_{\mathbb{U},-} \coloneqq \{i \in \mathcal{V}_{\mathbb{U}} : (\widehat{\boldsymbol{y}}_{\mathrm{PCA}})_i < 0\}$.

3.2.2. The regime $q_m \gtrsim \log(m)$

Theorem 3.2 and Theorem 3.4 (a) establish the sharp threshold for exact recovery, i.e., $I(a_\tau, b_\tau, c_\tau) = 1$, verified by the numerical simulations in Figures 2 and 3.

Theorem 3.4. *Let Assumption 3.1 hold and $q_m \gtrsim \log(m)$.*

(a) (Exact). When $I_\tau = I(a_\tau, b_\tau, c_\tau) > 1$, $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ achieves exact recovery with probability at least $1 - m^{1-I_\tau}$.

(b) (Optimal). When $I_\tau = I(a_\tau, b_\tau, c_\tau) \leq 1$, it follows

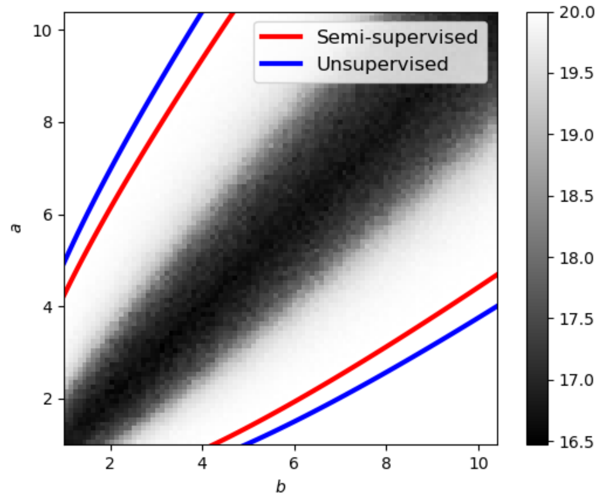
$$\limsup_{m\to\infty} q_m^{-1} \log \mathbb{E}\psi_m(\boldsymbol{y}_{\mathbb{U}}, \widehat{\boldsymbol{y}}_{\mathrm{PCA}}) \leq -I_\tau.$$

Informally, the second part of Theorem 3.4 can be understood as $\mathbb{E}\psi_m(\boldsymbol{y}_{\mathbb{U}}, \widehat{\boldsymbol{y}}_{\mathbb{U}}) \leq e^{-I_\tau\cdot q_m}$, which establishes an upper bound of the expected mismatch ratio. It matches the lower bound in Theorem 3.3. In that sense, even though exact recovery is impossible when $I(a_\tau, b_\tau, c_\tau) \leq 1$ by Theorem 3.2, the estimator $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ in (3.4) arrives the lowest possible error rate when $q_m \gtrsim \log(m)$.

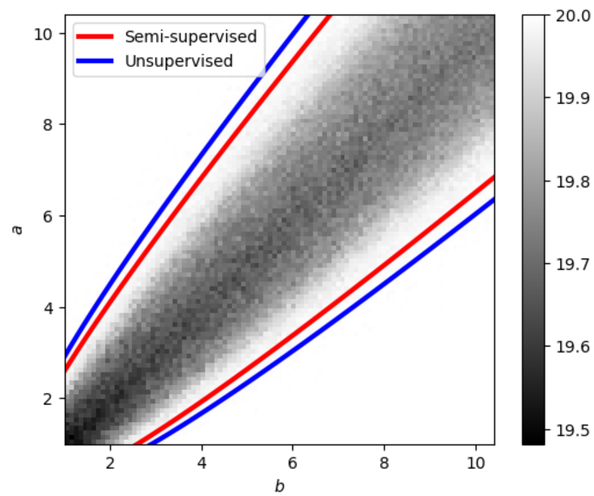
3.2.3. The regime $1 \ll q_m \ll \log(m)$

When the graph becomes even sparser, where the expected degree of each vertex goes to infinity slower than $\log(m)$, the previous estimator $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ in (3.4) is no longer valid. There are two main issues. First, $\widehat{\kappa}_{\ell^\star}$ was designed for the estimation of $\log(a_\tau/b_\tau)$, but it does not converge to $\log(a_\tau/b_\tau)$ anymore when $1 \ll q_m \ll \log(m)$, since $\lambda_1(\boldsymbol{A}_{\mathbb{U}})$ and $\lambda_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})$ no longer concentrate around $\frac{\alpha+\beta}{2}$ and $\frac{\alpha-\beta}{2}$ (Feige & Ofek, 2005). To get rid of that, we refer to the quadratic forms $\mathbf{1}^\top\boldsymbol{A}_{\mathbb{L}}\mathbf{1}$ and $\boldsymbol{y}_{\mathbb{L}}^\top\boldsymbol{A}_{\mathbb{L}}\boldsymbol{y}_{\mathbb{L}}$, which still present good concentration properties. Formally, we use the following $\widetilde{\kappa}_{\ell^\star}$ instead

$$\widetilde{\kappa}_{\ell^\star} \coloneqq \log\Big(\frac{\mathbf{1}^\top\boldsymbol{A}_{\mathbb{L}}\mathbf{1} + \boldsymbol{y}_{\mathbb{L}}^\top\boldsymbol{A}_{\mathbb{L}}\boldsymbol{y}_{\mathbb{L}}}{\mathbf{1}^\top\boldsymbol{A}_{\mathbb{L}}\mathbf{1} - \boldsymbol{y}_{\mathbb{L}}^\top\boldsymbol{A}_{\mathbb{L}}\boldsymbol{y}_{\mathbb{L}}}\Big).$$



(a) $c_\tau = 0.5$.



(b) $c_\tau = 1.5$

Figure 2. Performance of $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ in (3.4): fix $N = 800$, $\tau = 0.25$ and vary a (y-axis) and b (x-axis) from 1 to 10.5. For each parameter configuration (a_τ, b_τ, c_τ), we compute the frequency of exact recovery over 20 independent runs. Light color represents a high chance of success. Phase transitions occurs at the red curve $I(a_\tau, b_\tau, c_\tau) = 1$, as proved by Theorems 3.2 and 3.4.

The second issue is that, the entrywise eigenvector analysis of $\boldsymbol{u}_2(\boldsymbol{A}_{\mathbb{U}})$ breaks down due to the lack of concentration. To overcome that, we let $\widehat{\boldsymbol{y}}_{\mathrm{G}} = \mathrm{sign}(\widehat{\boldsymbol{y}}_{\mathrm{GMM}})$. Note that $\boldsymbol{A}_{\mathbb{U}}\widehat{\boldsymbol{y}}_{\mathrm{G}}$ is closed to $\sqrt{m}\lambda_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})\boldsymbol{u}_{\ell^\star}(\boldsymbol{A}_{\mathbb{U}})$, then the new graph estimator is defined through

$$\widetilde{\boldsymbol{y}}_{\mathrm{SBM}} \coloneqq \widetilde{\kappa}_{\ell^\star}\big(\boldsymbol{A}_{\mathbb{U}\mathbb{L}}\boldsymbol{y}_{\mathbb{L}} + \boldsymbol{A}_{\mathbb{U}}\widehat{\boldsymbol{y}}_{\mathrm{G}}\big)/\sqrt{m} \qquad (3.5)$$

Combining the above reasoning together, the estimator for under the general case is given by $\mathrm{sign}(\widetilde{\boldsymbol{y}}_{\mathrm{PCA}})$, where

$$\widetilde{\boldsymbol{y}}_{\mathrm{PCA}} = \widetilde{\boldsymbol{y}}_{\mathrm{SBM}} + \widehat{\boldsymbol{y}}_{\mathrm{GMM}}.$$

The following result shows that $\widetilde{\boldsymbol{y}}_{\mathrm{PCA}}$ achieves the lowest possible expected error rate when $1 \ll q_m \ll \log(m)$.

Theorem 3.5. *Let Assumption 3.1 hold, then it follows*

$$\limsup_{m\to\infty} q_m^{-1} \log \mathbb{E}\psi_m\big(\boldsymbol{y}_{\mathbb{U}}, \mathrm{sign}(\widetilde{\boldsymbol{y}}_{\mathrm{PCA}})\big) \leq -I(a_\tau, b_\tau, c_\tau).$$

3.2.4. Comparation with unsupervised regime

When only the sub-graph $\mathcal{G}_{\mathbb{U}} = (\mathcal{V}_{\mathbb{U}}, \mathcal{E}_{\mathbb{U}})$ is observed, it becomes an unsupervised learning task on $\mathcal{G}_{\mathbb{U}}$, where the data is equivalently sampled from $(\boldsymbol{A}_{\mathbb{U}}, \{\boldsymbol{x}_i\}_{i=1}^m) \sim \mathrm{CSBM}(\boldsymbol{y}_{\mathbb{U}}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ with $\alpha = aq_m/m$ and $\beta = bq_m/m$. The rate function can be obtained by simply taking $\tau = 0$ with $a_0 = a$, $b_0 = b$, and $c_0 = q_m^{-1}(\theta^2 + d/m)^{-1}\theta^4$, aligning with the result in (Abbe et al., 2022). The difference between the two boundaries $I(a_\tau, b_\tau, c_\tau) = 1$ (red) and $I(a_0, b_0, c_0) = 1$ (blue) is presented in Figure 2. A crucial observation is that, the extra information from $\boldsymbol{X}_{\mathbb{U}}$, $\boldsymbol{A}_{\mathbb{U}}$ and $\boldsymbol{A}_{\mathbb{U}\mathbb{L}}$ shrinks the boundary for exact recovery, making the task easier compared with the unsupervised regime.

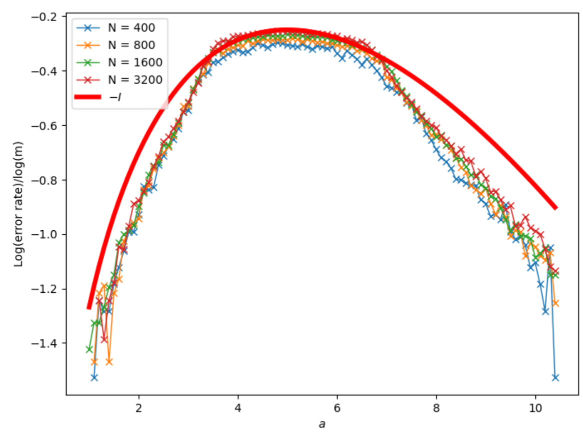


Figure 3. The y-axis is $q_m^{-1}\log(\mathbb{E}\psi_m)$, the average mismatch ratio on the logarithmic scale. The x-axis is a, varying from 0 to 10.5. Fix $b = 5$, $\tau = 0.25$, $c_\tau = 0.5$. The red curve is $-I(a_\tau, b_\tau, c_\tau)$, the lower bound predicted by Theorem 3.3. The experiments over different N shows that $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ achieves the information-theoretical limits, as proved in Theorems 3.4 and 3.5.

3.3. Performance of ridge regression on linear GCN

For $\mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$, in this section, we focus on analyzing how these parameters a, b, c_τ and τ defined in Assumption 3.1 affect the learning performances of the *lin-*

ear graph convolutional networks. We consider a graph convolutional kernel $h(\boldsymbol{X}) \in \mathbb{R}^{N\times d}$ which is a function of data matrix $\boldsymbol{X}$ and adjacency matrix $\boldsymbol{A}$ sampled from $\mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$. We add self-loops and define the new adjacency matrix $\boldsymbol{A}_\rho := \boldsymbol{A} + \rho \boldsymbol{I}_N$, where $\rho > 0$ represents the intensity of self connections in the graph. Let $\boldsymbol{D}_\rho$ be the diagonal matrix whose diagonals are the average degree for $\boldsymbol{A}_\rho$, i.e., $[\boldsymbol{D}_\rho]_{ii} = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^N (\boldsymbol{A}_\rho)_{ij}$ for each $i \in [N]$. For the linear graph convolutional layer, we will consider the following normalization:

$$h(\boldsymbol{X}) = \frac{1}{\sqrt{N q_m}} \boldsymbol{D}_\rho^{-1} \boldsymbol{A}_\rho \boldsymbol{X}, \tag{3.6}$$

Denote $\boldsymbol{D} := \boldsymbol{D}_0$, indicating no self-loop added, and D_0 as the first diagonal of $\boldsymbol{D}$. We study the linear ridge regression on $h(\boldsymbol{X})$. Compared with the general GCN defined in (Kipf & Welling, 2017), here we simplify the graph convolutional layer by replacing the degree matrix of $\boldsymbol{A}$ by the average degree among all vertices. In this case, we can directly employ $\widetilde{h}(\boldsymbol{X}) = \frac{1}{\widetilde{d}\cdot\sqrt{N q_m}} \boldsymbol{A}_\rho \boldsymbol{X}$ to approximate the $h(\boldsymbol{X})$, where $\widetilde{d}$ is the expected average degree defined by (C.1). Notice that for sparse graph $\boldsymbol{A}$ under Assumption 3.1, the degree concentration for each degree is *not* guaranteed, which is a different situation from (Baranwal et al., 2021; 2023a).

We now consider transductive learning on CSBM following the idea from (Baranwal et al., 2021; Shi et al., 2024). Recall that the vertex set $\mathcal{V}$ is split into two disjoint sets $\mathcal{V}_{\mathbb{L}}$ and $\mathcal{V}_{\mathbb{U}}$, where $n = |\mathcal{V}_{\mathbb{L}}|$, $m = |\mathcal{V}_{\mathbb{U}}|$ and $N = n + m$. The training ratio $\tau = \frac{n}{N}$ as $N \to \infty$. From Definition 2.7, we know that $\boldsymbol{S}_{\mathbb{L}} \in [0,1]^{n\times N}$, $\boldsymbol{S}_{\mathbb{U}} \in [0,1]^{m\times N}$, $\boldsymbol{S}_{\mathbb{L}}\boldsymbol{X} = \boldsymbol{X}_{\mathbb{L}} \in \mathbb{R}^{n\times d}$, and $\boldsymbol{S}_{\mathbb{U}}\boldsymbol{X} = \boldsymbol{X}_{\mathbb{U}} \in \mathbb{R}^{m\times d}$. Then, the empirical loss of linear ridge regression (LLR) on $h(\boldsymbol{X})$ can be written as

$$L(\boldsymbol{\beta}) = \frac{1}{n}\|\boldsymbol{S}_{\mathbb{L}}(h(\boldsymbol{X})\boldsymbol{\beta} - \boldsymbol{y})\|_2^2 + \frac{\lambda}{n}\|\boldsymbol{\beta}\|_2^2,$$

for any $\lambda > 0$, where the solution to this problem is

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^d} L(\boldsymbol{\beta}) \\ &= (h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{y}, \end{aligned}\tag{3.7}$$

where $\boldsymbol{P}_{\mathbb{L}} = \boldsymbol{S}_{\mathbb{L}}^\top \boldsymbol{S}_{\mathbb{L}} \in \mathbb{R}^{N\times N}$ is a diagonal matrix. Similarly, define $\boldsymbol{P}_{\mathbb{U}} = \boldsymbol{S}_{\mathbb{U}}^\top \boldsymbol{S}_{\mathbb{U}} \in \mathbb{R}^{N\times N}$. Then the estimator of this linear ridge regression for $\lambda > 0$ is

$$\widehat{\boldsymbol{y}}_{\mathrm{LRR}} = \boldsymbol{S}_{\mathbb{U}} h(\boldsymbol{X})(h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{y}. \tag{3.8}$$

In the following, we aim to analyze the misclassification rate $\psi_m(\boldsymbol{y}_{\mathbb{U}}, \widehat{\boldsymbol{y}}_{\mathrm{LRR}})$, the associated test and training risks in mean square error (MSE) defined by

$$\mathcal{R}(\lambda) := \frac{1}{m}\|\boldsymbol{S}_{\mathbb{U}}(h(\boldsymbol{X})\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\|_2^2 \tag{3.9}$$

$$\mathcal{E}(\lambda) := \frac{1}{n}\|\boldsymbol{S}_{\mathbb{L}}(h(\boldsymbol{X})\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\|_2^2. \tag{3.10}$$

Notice that Shi et al. (2024) also studied the asymptotic test and training risks for CSBM on linear GCNs but in a sparser graph $\boldsymbol{A}$ with constant average degree. They utilized statistical physics methods with some Gaussian equivalent conjecture to compute the asymptotic risks. Below, we provide detailed statements for the exact recovery thresholds of $\widehat{\boldsymbol{y}}_{\mathrm{LRR}}$.

Theorem 3.6 (Exact recovery for graph convolution linear ridge regression)**.** *Consider the ridge regression on the linear graph convolution $h(\boldsymbol{X})$ defined in (3.6) with estimator $\widehat{\boldsymbol{y}}_{\mathrm{LLR}}$ in (3.8). Assume that $\rho \lesssim q_m$, $\theta^2 = (1 + o(1))c_\tau q_m$ and $q_m \lesssim d \lesssim \sqrt{N q_m}$.Then, under Assumption 3.1, we can conclude that*

(a) When $\rho = 0$, then $\mathbb{P}(\psi_m(\boldsymbol{y}_{\mathbb{U}}, \mathrm{sign}(\widehat{\boldsymbol{y}}_{\mathrm{LRR}})) = 0) \to 1$ as long as $I(a_\tau, b_\tau, 0) > 1$.

(b) When

$$\rho = \frac{2c_\tau}{\log(a_\tau/b_\tau)} q_m, \tag{3.11}$$

then $\mathbb{P}(\psi_m(\boldsymbol{y}_{\mathbb{U}}, \mathrm{sign}(\widehat{\boldsymbol{y}}_{\mathrm{LRR}})) = 0) \to 1$ as long as $I(a_\tau, b_\tau, c_\tau) > 1$.

(c) When $\rho = s q_m$ for some constant $s \in \mathbb{R}$, then

$$\mathbb{P}(\psi_m(\boldsymbol{y}_{\mathbb{U}}, \mathrm{sign}(\widehat{\boldsymbol{y}}_{\mathrm{LRR}})) = 0) \to 1$$

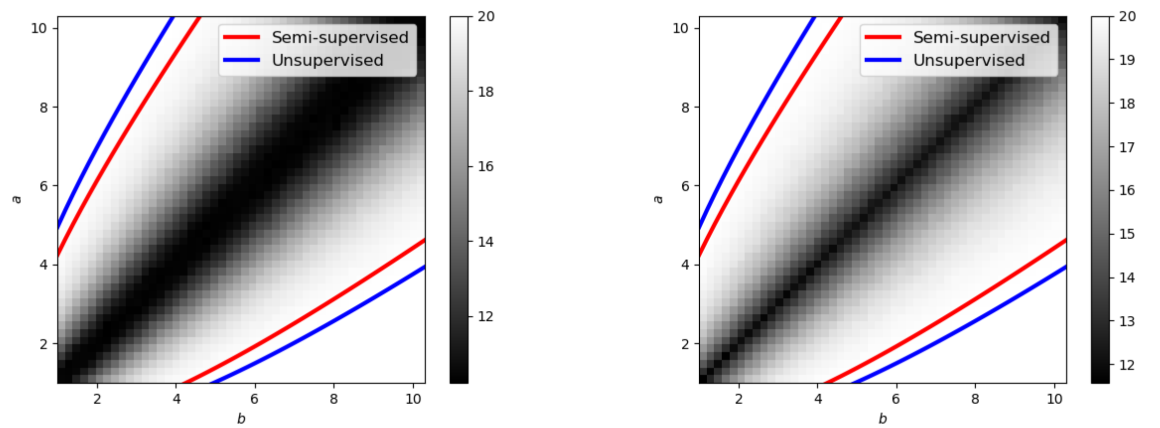
when $J(a_\tau, b_\tau, c_\tau, \zeta, s) > 1$, as $m \to \infty$, where $\zeta := \frac{\kappa\tau}{\kappa^2\tau + \lambda}$ and $\kappa := \sqrt{c_\tau}\cdot\frac{a_\tau - b_\tau + 2s}{a_\tau + b_\tau + 2s}$ for $\lambda > 0$. Here rate function $J(a_\tau, b_\tau, c_\tau, \zeta, s)$ is defined in Lemma C.2. Additionally, we know that the exact recovery region $\{(a_\tau, b_\tau, c_\tau) : J(a_\tau, b_\tau, c_\tau, \zeta, s) > 1\}$ is a subset of the optimal region $\{(a_\tau, b_\tau, c_\tau) : I(a_\tau, b_\tau, c_\tau) > 1\}$.

Consequently, $\rho = \frac{2c_\tau}{\log(a/b)} q_m$ is the *optimal* weighted self-loop to attain the exact recovery of labels $\boldsymbol{y}_{\mathbb{U}}$ in this semi-supervised learning with linear ridge regression on $h(\boldsymbol{X})$. This is because in this case, the exact recovery for $\widehat{\boldsymbol{y}}_{\mathrm{LRR}}$ matches the information-theoretic lower bound in Theorem 3.3, i.e., below this threshold, no algorithms can perfectly recover all the unknown labels in $\mathcal{V}_{\mathbb{U}}$.

Theorem 3.7 (Asymptotic training and test errors)**.** *Consider $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$. Suppose that $\rho/q_m \to s \in \mathbb{R}$ and $d \lesssim N$. Under the Assumption 3.1, the training and test errors for linear ridge regression on $h(\boldsymbol{X})$ defined by (3.7) are asymptotically satisfying the following results. For any fixed $\lambda > 0$, both training and test errors in MSE loss defined in (3.9) and (3.10) satisfy*

$$\mathcal{E}(\lambda) \text{ and } \mathcal{R}(\lambda) \to \frac{\lambda^2}{(\kappa^2\tau + \lambda)^2},$$

almost surely, as $m, N \to \infty$, where κ is defined in Theorem 3.6 (c).



(a) Without self-loop. (b) With optimal self-loop

Figure 4. Performance of $\widehat{\boldsymbol{y}}_{\mathrm{LRR}}$ in (3.8). Fix $N = 800$, $\tau = 0.25$, $c_\tau = 0.5$. Compute the frequency of exact recovery over 20 independent runs. When $I(a_\tau, b_\tau, c_\tau) > 1$, $\widehat{\boldsymbol{y}}_{\mathrm{LRR}}$ achieves exact recovery, as proved in Theorem 3.6 (a) and (b).

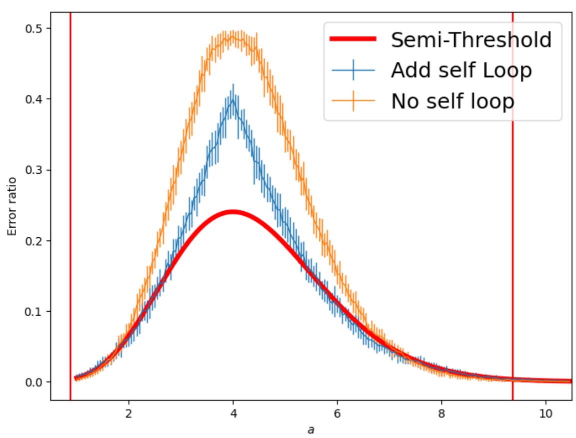


Figure 5. The y-axis is $\mathbb{E}\psi_m$, the average mismatch ratio over 20 independent runs. The x-axis is a, varying from 0 to 10.5. Fix $b = 4$, $\tau = 0.25$, $c_\tau = 0.5$, $N = 400$. The red curve is $m^{-I(a_\tau, b_\tau, c_\tau)}$, the lower bound predicted by Theorem 3.3 with $q_m = \log(m)$. The experiments shows that $\widehat{\boldsymbol{y}}_{\mathrm{LRR}}$ achieves a lower mismatch ratio when adding self-loop in the area $I(a_\tau, b_\tau, c_\tau) < 1$, where the exact recovery is impossible.

3.4. Performance of GCN with gradient-based training

In this section, we study the feature learning of GCN on $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ with n known labels and m unknown labels to be classified. We focus on gradient-based training processes. From Section 3.3, we can indicate that the self-connection (or self-loop) weight ρ plays an important role in exact recovery on test feature vertices. It turns out that we need to find the optimal ρ in (3.11) for self-loop weight to ensure the exact recovery threshold approaches to the IT bound studied in Section 3.1 for graph learning. A similar idea is also mentioned in (Kipf & Welling, 2017), whereas the equal status of self-connections and edges to neighboring nodes may not be a good assumption for a general graph dataset. Hence, we raise a modified training process for GCNs: in general, learning on graphs also requires learning the optimal self-loop weight for the graph, i.e., we should also view parameter ρ in graph $\boldsymbol{A}_\rho = \boldsymbol{A} + \rho \boldsymbol{I}_N$ as a trainable parameter. Although the optimal ρ in Section 3.3 for CSBM on semisupervised learning is due to LDP analysis (see Appendix C.1), we can generally apply a spectral method to achieve oracle ρ in (3.11).

For simplicity, in this section, we denote

$$\boldsymbol{A}_s = \boldsymbol{A} + s q_m \boldsymbol{I}_N, \quad \boldsymbol{D}_s = s q_m \boldsymbol{I}_N + \boldsymbol{D},$$

where $\boldsymbol{D}$ is a diagonal matrix with the average degree for each diagonal. In this section, we view $s \in \mathbb{R}$ as a trainable parameter. Let us consider a general two-layer graph convolutional neural network defined by

$$f(\boldsymbol{X}) := \frac{1}{\sqrt{K}} \sigma(\boldsymbol{D}_s^{-1} \boldsymbol{A}_s \boldsymbol{X} \boldsymbol{W}) \boldsymbol{a} \tag{3.12}$$

where the first-layer weight matrix is $\boldsymbol{W} \in \mathbb{R}^{d \times K}$ and second layer weight matrix is $\boldsymbol{a} \in \mathbb{R}^K$ for some $K \in \mathbb{N}$. Here, $\boldsymbol{W}, s$ are training parameters for this GCN. We aim to train this neural network with training label $\boldsymbol{y}_{\mathbb{L}}$ to predict the labels for vertices in $\mathcal{V}_{\mathbb{U}}$. Notice that when training $\boldsymbol{W}$, we want $\boldsymbol{W}$ to learn (align with) the correct feature $\boldsymbol{\mu}$ in the dataset. As studied in (Baranwal et al., 2021), for CSBM with a large threshold, the data point feature is linearly separable, hence there is no need to introduce a nonlinear convolution layer in (3.12). So we will consider $\sigma(x) = x$ below. In practice, nonlinearity for node classification may not be important in certain graph learning problems (Wu et al., 2019; He et al., 2020).

We train this GCN in two steps. First, we train the $\boldsymbol{W}$ with a large gradient descent step on training labels. By choosing the suitable learning rate, we can allow $\boldsymbol{W}$ to learn the feature $\boldsymbol{\mu}$. Let us define the MSE loss by

$$\mathcal{L}(\boldsymbol{W}, s) = \frac{1}{2n}(f(\boldsymbol{X}) - \boldsymbol{y})^\top \boldsymbol{P}_{\mathbb{L}}(f(\boldsymbol{X}) - \boldsymbol{y}).$$

The analysis for GD with a large learning rate to achieve feature learning is analogous with (Ba et al., 2022; Damian et al., 2022). We extend this analysis to one-layer GCNs. Precisely, we take a one-step GD with a weight decay λ_1 and learning rate η_1:

$$\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} - \eta_1\Big(\nabla_{\boldsymbol{W}^{(0)}}\mathcal{L}(\boldsymbol{W}^{(0)}, s^{(0)}) + \lambda_1 \boldsymbol{W}^{(0)}\Big).$$

Secondly, we find out the optimal s based on (3.11). Here, we only use training labels $\boldsymbol{y}_{\mathbb{L}}$. Let

$$s^{(1)} = \frac{2}{n^2 q_m}(\boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{X}_{\mathbb{L}} \boldsymbol{W}^{(1)} \boldsymbol{a}) \Big/ \log\left(\frac{\mathbf{1}^\top \boldsymbol{A}\mathbf{1} + \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}} \boldsymbol{y}_{\mathbb{L}}}{\mathbf{1}^\top \boldsymbol{A}\mathbf{1} - \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}} \boldsymbol{y}_{\mathbb{L}}}\right).$$

This construction resembles the spectral methods defined in (3.5). Meanwhile, we can also replace this estimator with the gradient-based method to optimize s in MSE loss which is shown in Appendix D. However, to attain the IT bound, the *nonlinearity* of $\sigma(x)$ in (3.12) plays an important role when applying GD to find optimal self-loop weight s. This observation is consistent with results by Wei et al. (2022); Baranwal et al. (2023b), where nonlinearity needed for GCN to obtain certain Bayes optimal in sparse graph learning.

Assumption 3.8. *Consider $N, d, K \to \infty$, $n \asymp N$, $K \asymp N$, $q_m = \log(m)$ and $d = o(q_m^2)$. We assume that at initialization $s^{(0)} = 0$, and $\sqrt{K} \cdot [\boldsymbol{W}^{(0)}]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\sqrt{K} \cdot [\boldsymbol{a}]_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{\pm 1\}$, for all $i \in [d], j \in [K]$.*

With initialization stated in Assumption 3.8 and trained parameters $\boldsymbol{W}^{(1)}$ and $s^{(1)}$, we derive a GCN estimator for unknown labels which matches with Theorems 3.3 and 3.2.

Theorem 3.9. *Under Assumptions 3.1 and 3.8, suppose that learning rate $\eta_1 = \Theta(K/\sqrt{q_m})$ and weight decay rate $\lambda_1 = \eta_1^{-1}$. Then, estimator $\widehat{\boldsymbol{y}}_{\mathrm{GCN}} = \boldsymbol{S}_{\mathbb{U}} f(\boldsymbol{X})$ with $\boldsymbol{W} = \boldsymbol{W}^{(1)}$ and $s = s^{(1)}$ satisfies that*

$$\mathbb{P}(\psi_m(\boldsymbol{y}_{\mathbb{U}}, \mathrm{sign}(\widehat{\boldsymbol{y}}_{\mathrm{GCN}})) = 0) \to 1$$

when $I(a_\tau, b_\tau, c_\tau) > 1$, as $m \to \infty$. Hence, GCN can attain the IT bound for the exact recovery of CSBM.

Remark 3.10. *(Duranthon & Zdeborová, 2023) proposed the AMP-BP algorithm to solve the community detection problem under CSBM, where the expected degree of each vertex is constant, i.e., $q_m = O(1)$. By contrast, this manuscript focuses on the regime $q_m \gg 1$ as in Assumption 3.1. Theorem 3.9 shows that the GCN achieves exact recovery when $I(a_\tau, b_\tau, c_\tau) > 1$. However, the performance of the GCN is not characterized when $I(a_\tau, b_\tau, c_\tau) < 1$, and it is still unclear whether it would match the lower bound proved in Theorem 3.3, i.e., the optimality of GCN remains open. From simulations in Figure 6, we observe that below the IT bound ($I(a_\tau, b_\tau, c_\tau) < 1$), there is a gap between theoretical optimal error (red curve) and the simulated mismatch ratio by GCN estimators.*

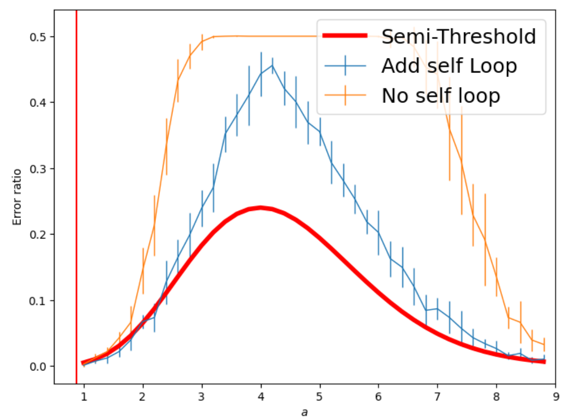


Figure 6. Mismatch ratio difference of $\widehat{\boldsymbol{y}}_{\mathrm{GCN}}$ when with or without self-loop for fixed $b = 4$, $c_\tau = 0.5$, $N = 400$.

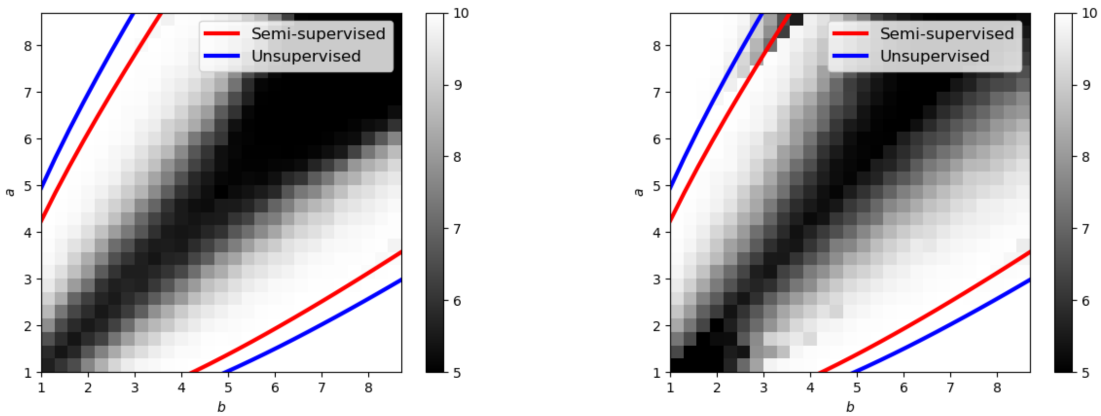
4. Numerical simulations

4.1. Optimal spectral method

The efficacy of the spectral estimator $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ is demonstrated in Figures 2 (a) and 3 for $c_\tau = 0.5$ and Figure 2 (b) for $c_\tau = 1.5$. We fix $N = 800$, $\tau = 0.25$, but vary a (y-axis) and b (x-axis) from 1 to 10.5 in Figure 2, and compute the frequency of exact recovery over 20 independent trials for each parameter configuration (a_τ, b_τ, c_τ). Here, a lighter color represents a higher success chance. The (red) and (blue) curves represent the boundaries for exact recovery under semi-supervised and unsupervised regimes respectively. A larger c_τ implies a stronger signal in node features, which shrinks the boundary for exact recovery and makes the problem easier. In Figure 3, we fix $b = 5$ but vary a (x-axis) from 1 to 10.5. The simulations for the average mismatch ratio are presented on the logarithmic scale over different choices of N. Clearly, $\log \mathbb{E}\psi_m$ will approach the lower bound (red curve) as proved in Theorems 3.3, 3.4 (2).

4.2. Ridge regression on linear GCNs

The efficacy of the ridge estimator $\widehat{\boldsymbol{y}}_{\mathrm{LRR}}$ is presented in Figures 4 and 5. We fix $N = 800$, $\tau = 0.25$ and $c_\tau = 0.5$



(a) Exact recovery counts without self-loop.

(b) Exact recovery counts with self loop ρ.

Figure 7. Performance of $\widehat{\boldsymbol{y}}_{\mathrm{GCN}}$ when $N = 400$, $\tau = 0.25$, $c_\tau = 0.5$.

in Figure 4, but vary a (y-axis) and b (x-axis) from 1 to 10.5, where 20 independent trials are performed on each (a_τ, b_τ, c_τ). The difference between the (a) and (b) lies on the choice of the self-loop density ρ, where we take $\rho = 0$ in (a) but $\rho = 2c_\tau q_m \log(a_\tau/b_\tau)$ in (b) as (3.11). In Figure 5, we fix $b = 4$, $N = 400$ but vary a (x-axis) from 1 to 10.5. When $I(a_\tau, b_\tau, c_\tau) < 1$, the performance difference between the choices of ρ are presented. From simulations, the average mismatch ratio is closer to the predicted lower bound (red curve) when the optimal self-loop is added.

4.3. Gradient-based training on GCN

The efficacy of $\widehat{\boldsymbol{y}}_{\mathrm{GCN}}$ is presented in Figures 6 and 7. Similarly, we fix $N = 400$, $\tau = 0.25$ and $c_\tau = 0.5$, but vary a (y-axis) and b (x-axis) from 1 to 9 in Figure 7. For each (a_τ, b_τ, c_τ), 10 independent trials are performed. We plot the performance when adding self-loops to the graph data, where we take $\rho = 0$ in (a) but $\rho = 2c_\tau q_m \log(a_\tau/b_\tau)$ in (b) as (3.11). In Figure 6, we fix $b = 4$, $c_\tau = 0.5$, $N = 400$ but vary a (x-axis) from 1 to 9. The performance difference between the choices of ρ when $I(a_\tau, b_\tau, c_\tau) < 1$ are presented. From the simulations, the average mismatch ratio is closer to the predicted bound (red curve) when the optimal self-loop is added.

5. Discussion and conclusion

Firstly, as shown in ℓ^* and $\widehat{\kappa}_{\ell^*}$, our results for spectral method, ridge regression, and GCNs all cover Erdős-Rényi graph ($a = b$), homophilic graphs ($a > b$) and heterophilic graphs ($a < b$). When $a = b$, we can only utilize the node feature from GMM for classification, which returns to the semisupervised learning on GMM (Lelarge & Miolane, 2019; Oymak & Cihad Gulcu, 2021; Nguyen & Couillet, 2023). For heterophilic graphs with $a < b$, the optimal self-loop strength ρ defined in (3.11) is negative, which validates the observation in Figure 5 of (Shi et al., 2024).

Our research pioneers the investigation of the exact recovery threshold in semi-supervised learning on CSBM. We present various strategies for achieving exact recovery, including the spectral method, linear ridge regression applied to linear Graph Convolutional Networks (GCNs), and gradient-based training techniques for GCNs. For each method, we establish precise asymptotic lower bounds that depend on the sparsity of the Stochastic Block Model (SBM) and the signal-to-noise ratio (SNR) in Gaussian Mixture Models (GMM), which, in many instances, are optimal when compared to the information-theoretic (IT) bound. Crucially, our findings support the notion that GCNs, when equipped with certain gradient-based training protocols, can flawlessly recover all unlabeled vertices provided the SNR exceeds the IT bound. This finding underscores the effectiveness of GCNs in addressing classification problems within CSBM settings. For future work, we can delve into the exact recovery rates for more intricate and nonlinear graph models, such as XOR-SBM and geometric Gaussian graphs. Additionally, we intend to shed light on the process of feature learning in GCNs and identify the optimal GCN architectures that prevent over-smoothing and meet IT bounds.

Acknowledgements

H.W. and Z.W. acknowledge the support from NSF DMS-2154099. Z.W. is also partially supported by NSF DMS-2055340.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Abbe, E. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018. URL http://jmlr.org/papers/v18/16-480.html.

Abbe, E., Bandeira, A. S., and Hall, G. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

Abbe, E., Fan, J., and Wang, K. An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

Alt, J., Ducatez, R., and Knowles, A. Extremal eigenvalues of critical Erdős–Rényi graphs. *The Annals of Probability*, 49(3):1347–1401, 2021.

Azriel, D., Brown, L. D., Sklar, M., Berk, R., Buja, A., and Zhao, L. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35: 37932–37946, 2022.

Baranwal, A., Fountoulakis, K., and Jagannath, A. Graph convolution for semi-supervised classification: Improved linear separability and out-of-distribution generalization. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 684–693. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/baranwal21a.html.

Baranwal, A., Fountoulakis, K., and Jagannath, A. Effects of graph convolutions in multi-layer networks. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=P-73JPgRs0R.

Baranwal, A., Fountoulakis, K., and Jagannath, A. Optimality of message-passing architectures for sparse graphs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=d1knqWjmNt.

Belkin, M., Matveeva, I., and Niyogi, P. Regularization and semi-supervised learning on large graphs. In *Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings 17*, pp. 624–638. Springer, 2004.

Bojchevski, A. and Günnemann, S. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1ZdKJ-0W.

Bruna, J. and Li, X. Community detection with graph neural networks. *stat*, 1050:27, 2017.

Chakrabortty, A. and Cai, T. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, pp. 1541–1572, 2018.

Chen, Z., Li, L., and Bruna, J. Supervised community detection with line graph neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1g0Z3A9Fm.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Deshpande, Y., Sen, S., Montanari, A., and Mossel, E. Contextual stochastic block models. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/08fc80de8121419136e443a70489c123-Paper.pdf.

Dumitriu, I. and Wang, H. Optimal and exact recovery on the general non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2304.13139*, 2023.

Duranthon, O. and Zdeborová, L. Optimal inference in contextual stochastic block models. *arXiv preprint arXiv:2306.07948*, 2023.

Esser, P., Chennuru Vankadara, L., and Ghoshdastidar, D. Learning theory can (sometimes) explain generalisation in graph neural networks. *Advances in Neural Information Processing Systems*, 34:27043–27056, 2021.

Feige, U. and Ofek, E. O. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27, 2005. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20089.

Fountoulakis, K., Levi, A., Yang, S., Baranwal, A., and Jagannath, A. Graph attention retrospective. *Journal of Machine Learning Research*, 24(246):1–52, 2023.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

Huang, W., Cao, Y., Wang, H., Cao, X., and Suzuki, T. Graph neural networks provably benefit from structural information: A feature learning perspective. *arXiv preprint arXiv:2306.13926*, 2023.

Kim, C., Bandeira, A. S., and Goemans, M. X. Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. *arXiv preprint arXiv:1807.02884*, 2018.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Lampert, M. and Scholtes, I. The self-loop paradox: Investigating the impact of self-loops on graph neural networks. *arXiv preprint arXiv:2312.01721*, 2023.

Lelarge, M. and Miolane, L. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 639–643. IEEE, 2019.

Lu, C. and Sen, S. Contextual stochastic block model: Sharp thresholds and contiguity. *Journal of Machine Learning Research*, 24(54):1–34, 2023.

Lu, W. LEARNING GUARANTEES FOR GRAPH CONVOLUTIONAL NETWORKS ON THE STOCHASTIC BLOCK MODEL. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dpXL6lz4mOQ.

Nguyen, M.-T. and Couillet, R. Asymptotic bayes risk of semi-supervised multitask learning on gaussian mixture. In *International Conference on Artificial Intelligence and Statistics*, pp. 5063–5078. PMLR, 2023.

Oono, K. and Suzuki, T. Graph neural networks exponentially lose expressive power for node classification. *arXiv preprint arXiv:1905.10947*, 2019.

Oymak, S. and Cihad Gulcu, T. A theoretical characterization of semi-supervised learning with self-training for gaussian mixture models. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3601–3609. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/oymak21a.html.

Ryan, K. J. and Culp, M. V. On semi-supervised linear regression in covariate shift problems. *The Journal of Machine Learning Research*, 16(1):3183–3217, 2015.

Shchur, O., Mumme, M., Bojchevski, A., and Günnemann, S. Pitfalls of graph neural network evaluation. *arXiv preprint arXiv:1811.05868*, 2018.

Shi, C., Pan, L., Hu, H., and Dokmanić, I. Homophily modulates double descent generalization in graph convolution networks. *Proceedings of the National Academy of Sciences*, 121(8):e2309504121, 2024.

Tang, H. and Liu, Y. Towards understanding generalization of graph neural networks. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33674–33719. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/tang23f.html.

Tony Cai, T. and Guo, Z. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2):391–419, 2020.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Wang, H. Fundamental limits and strong consistency of binary non-uniform hypergraph stochastic block model. *arXiv preprint arXiv:2306.06845*, 2023.

Wei, R., Yin, H., Jia, J., Benson, A. R., and Li, P. Understanding non-linearity in graph neural networks from the bayesian-inference perspective. *arXiv preprint arXiv:2207.11311*, 2022.

Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., and Weinberger, K. Simplifying graph convolutional networks. In *International conference on machine learning*, pp. 6861–6871. PMLR, 2019.

Wu, X., Chen, Z., Wang, W., and Jadbabaie, A. A non-asymptotic analysis of oversmoothing in graph neural networks. *arXiv preprint arXiv:2212.10701*, 2022.

Yang, Z., Cohen, W., and Salakhudinov, R. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*, pp. 40–48. PMLR, 2016.

A. Information-theoretic limits

In this section, we will provide the proofs for Theorem 3.2 and Theorem 3.3.

A.1. Impossibility for exact recovery

The proof sketch of Theorem 3.2 is presented in this section, with some proofs of Lemmas deferred.

Let $\boldsymbol{y} \in \{\pm 1\}^N$ denote the true label vector with $\boldsymbol{y} = [\boldsymbol{y}_{\mathbb{L}}^\top, \boldsymbol{y}_{\mathbb{U}}^\top]^\top$, where $\boldsymbol{y}_{\mathbb{L}}$ and $\boldsymbol{y}_{\mathbb{U}}$ denote the observed and uncovered label vector respectively. Assume $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ as in model 2.5, and the access to $\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}}$ are provided. Let $\widehat{\boldsymbol{y}}_{\mathbb{U}} \in \{\pm 1\}^m$ denote an estimator of $\boldsymbol{y}_{\mathbb{U}}$ obtained from algorithm. The probability that $\widehat{\boldsymbol{y}}_{\mathbb{U}}$ fails recovering every entry of $\boldsymbol{y}_{\mathbb{U}}$ is

$$\mathbb{P}_{\text{fail}} := \mathbb{P}(\widehat{\boldsymbol{y}}_{\mathbb{U}} \neq \pm \boldsymbol{y}_{\mathbb{U}}) = \sum_{\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}}} [1 - \mathbb{P}(\widehat{\boldsymbol{y}}_{\mathbb{U}} = \pm \boldsymbol{y}_{\mathbb{U}} | \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}})] \cdot \mathbb{P}(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}}),$$

where the *Maximum A Posteriori* (MAP) estimator achieves its minimum. Since the prior distribution of $\boldsymbol{y}$ is uniform sampled in Definition 2.5, the discussion on the ideal estimator can be transferred to *Maximum Likelihood Estimation* (MLE),

$$\widehat{\boldsymbol{y}}_{\text{MLE}} := \underset{\boldsymbol{z} \in \{\pm 1\}^m, \mathbf{1}^\top \boldsymbol{z} = 0}{\arg\max} \; \mathbb{P}(\boldsymbol{A} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) \cdot \mathbb{P}(\boldsymbol{X} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}). \tag{A.1}$$

Furthermore, Lemma A.3 shows the function that MLE is maximizing over $\boldsymbol{z} \in \{\pm 1\}^m$

$$f(\boldsymbol{z}) := \log \mathbb{P}(\boldsymbol{A} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) + \log \mathbb{P}(\boldsymbol{X} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}). \tag{A.2}$$

From the discussion above, MLE is equivalent to the best estimator MAP. No algorithm would be able to assign all labels correctly if MLE fails. In the view of (A.2), the failure of MLE indicates that some configuration $\boldsymbol{\sigma} \in \{\pm 1\}^m$ other than the true $\boldsymbol{y}_{\mathbb{U}}$ achieves its maximum, and MLE prefers $\boldsymbol{\sigma}$ other than $\boldsymbol{y}_{\mathbb{U}}$.

To establish the necessity, we explicitly construct some $\boldsymbol{\sigma} \in \{\pm 1\}^m$ with $\mathbf{1}^\top \boldsymbol{\sigma} = 0$ such that $\boldsymbol{\sigma} \neq \boldsymbol{y}_{\mathbb{U}}$ but $f(\boldsymbol{\sigma}) \geq f(\boldsymbol{y}_{\mathbb{U}})$ when below the threshold, i.e., $I_\tau(a, b, c) < 1$. An example of such $\boldsymbol{\sigma}$ can be constructed as follows. Pick $u \in \mathcal{V}_{\mathbb{U},+}$ and $v \in \mathcal{V}_{\mathbb{U},-}$ where $\mathcal{V}_{\mathbb{U},\pm} = \mathcal{V}_{\mathbb{U}} \cap \mathcal{V}_{\pm}$, and switch the labels of u and v in $\boldsymbol{y}_{\mathbb{U}}$ but keep all the others. Lemma A.1 characterizes the scenarios of failing exact recovery in terms of u and v.

Lemma A.1. *Given some subset $\mathcal{S} \subset \mathcal{V} = [N]$, for vertex $u \in \mathbb{U}$, define the following random variable*

$$W_{m,u}(\mathcal{S}) := y_u \cdot \bigg(\log(a/b) \cdot \sum_{j \in \mathcal{S}} A_{uj} y_j + \frac{2}{N + d/\theta^2} \sum_{j \in \mathcal{S}} \langle \boldsymbol{x}_u, \boldsymbol{x}_j \rangle y_j \bigg). \tag{A.3}$$

Denote by $W_{m,u} := W_{m,u}([N] \setminus \{u\})$ for any $u \in \mathcal{V}_{\mathbb{U}}$. Define the rate function

$$I(t, a_\tau, b_\tau, c_\tau) := \frac{1}{2} \Big(a_\tau - a_\tau \Big(\frac{a_\tau}{b_\tau}\Big)^t + b_\tau - b_\tau \Big(\frac{b_\tau}{a_\tau}\Big)^t \Big) - 2c_\tau(t + t^2). \tag{A.4}$$

Then, it supreme over t is attained at $t^\star = -1/2$,

$$\sup_{t \in \mathbb{R}} I(t, a_\tau, b_\tau, c_\tau) = I(-1/2, a_\tau, b_\tau, c_\tau) = \frac{1}{2}\Big((\sqrt{a_\tau} - \sqrt{b_\tau})^2 + c_\tau \Big) =: I(a_\tau, b_\tau, c_\tau),$$

where the last equality holds as in (3.2).

(a) For any $\varepsilon < \frac{a-b}{2(1-\tau)} \log(a/b) + 2c_\tau$ and $\delta > 0$, there exists some sufficiently large $m_0 > 0$, such that for $I(t, a_\tau, b_\tau, c_\tau)$ in (A.4), the following holds for any $m \geq m_0$

$$\mathbb{P}(W_{m,u} \leq \varepsilon q_m) = (1 + o(1)) \cdot \exp\Big(- q_m \cdot \big(-\delta + \sup_{t \in \mathbb{R}} \{\varepsilon t + I(t, a_\tau, b_\tau, c_\tau)\}\big)\Big).$$

(b) For the pair $u \in \mathcal{V}_{\mathbb{U},+}$ and $v \in \mathcal{V}_{\mathbb{U},-}$, the event $\{W_{m,u} \leq 0\} \cap \{W_{m,v} \leq 0\}$ implies $f(\boldsymbol{y}_{\mathbb{U}}) \leq f(\boldsymbol{\sigma})$ with probability at least $1 - e^{-q_m}$.

However, for any pair $u \in \mathcal{V}_{\mathbb{U},+}$, $v \in \mathcal{V}_{\mathbb{U},-}$, the variables $W_{m,u}$ and $W_{m,v}$ are not independent due to the existence of common random edges. To get rid of the dependency, let $\mathcal{U}$ be a subset of $\mathcal{V}_{\mathbb{U}}$ with cardinality $|\mathcal{U}| = \delta m$ where $\delta = \log^{-3}(m)$, such that $|\mathcal{U} \cap \mathcal{V}_{\mathbb{U},+}| = |\mathcal{U} \cap \mathcal{V}_{\mathbb{U},-}| = \delta m/2$. Define the following random variables

$$U_{m,u} := W_{m,u}([N] \setminus \mathcal{U}), \quad J_{m,u} := W_{m,u}(\mathcal{U} \setminus \{u\}), \quad J_m := \max_{u \in \mathcal{V}_{\mathbb{U}}} J_{m,u} \tag{A.5}$$

Obviously, for some $\zeta_m > 0$, $\{U_{m,u} \le -\zeta_m q_m\} \cap \{J_m \le \zeta_m q_m\}$ implies $\{W_{m,u} \le 0\}$ since $W_{m,u} = U_{m,u} + J_{m,u}$. Furthermore, $\{U_{m,u} \le -\zeta_m q_m\}$ does not reply on $\{J_m \le \zeta_m q_m\}$ since J_m is independent to any vertex in $\mathcal{U}$. Also, $\{U_{m,u}\}_{u \in \mathcal{V}_{\mathbb{U},+} \cap \mathcal{U}}$ is a set of independent random variables since no overlap edges. Thus the failure probability can be lower bounded by

$$\begin{aligned}
\mathbb{P}_{\text{fail}} \ge\, & \mathbb{P}(\exists u \in \mathcal{V}_{\mathbb{U},+},\, v \in \mathcal{V}_{\mathbb{U},-} \text{ s.t. } f(\boldsymbol{y}_{\mathbb{U}}) \le f(\boldsymbol{\sigma})) \\
\ge\, & \mathbb{P}\Big(\cup_{u \in \mathcal{V}_{\mathbb{U},+}} \{W_{m,u} \le 0\} \bigcap \cup_{v \in \mathcal{V}_{\mathbb{U},-}} \{W_{m,v} \le 0\} \Big) \ge \mathbb{P}\Big(\cup_{u \in \mathcal{V}_{\mathbb{U},+} \cap \mathcal{U}} \{W_{m,u} \le 0\} \bigcap \cup_{v \in \mathcal{V}_{\mathbb{U},-} \cap \mathcal{U}} \{W_{m,v} \le 0\} \Big) \\
\ge\, & \mathbb{P}\Big(\cup_{u \in \mathcal{V}_{\mathbb{U},+} \cap \mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\} \bigcap \cup_{v \in \mathcal{V}_{\mathbb{U},-} \cap \mathcal{U}} \{U_{m,v} \le -\zeta_m q_m\} \Big| \{J_m \le \zeta_m q_m\} \Big) \cdot \mathbb{P}(J_m \le \zeta_m q_m) \\
\ge\, & \mathbb{P}\Big(\cup_{u \in \mathcal{V}_{\mathbb{U},+} \cap \mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\} \Big| \{J_m \le \zeta_m q_m\} \Big) \\
& \cdot \mathbb{P}\Big(\cup_{v \in \mathcal{V}_{\mathbb{U},-} \cap \mathcal{U}} \{U_{m,v} \le -\zeta_m q_m\} \Big| \{J_m \le \zeta_m q_m\} \Big) \cdot \mathbb{P}(J_m \le \zeta_m q_m) \\
=\, & \mathbb{P}\Big(\cup_{u \in \mathcal{V}_{\mathbb{U},+} \cap \mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\} \Big) \cdot \mathbb{P}\Big(\cup_{v \in \mathcal{V}_{\mathbb{U},-} \cap \mathcal{U}} \{U_{m,v} \le -\zeta_m q_m\} \Big) \cdot \mathbb{P}(J_m \le \zeta_m q_m).
\end{aligned}$$

Lemma A.2. *For $\zeta_m = (\log \log m)^{-1}$ and $q_m = \log(m)$ and some constant $\widetilde{\delta} > 0$, the following holds*

$$\mathbb{P}(J_m \le \zeta_m q_m) \ge 1 - \log^{-3}(m) \cdot m^{-1+o(1)}, \quad \mathbb{P}\Big(\cup_{u \in \mathcal{V}_{\mathbb{U},+} \cap \mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\} \Big) \ge 1 - \exp\Big(-\frac{m^{1 - I(a_\tau, b_\tau, c_\tau) + \widetilde{\delta}}}{2 \log^3(m)} \Big).$$

With the lower bounds of the three components obtained in Lemma A.2, while $I(a_\tau, b_\tau, c_\tau) = 1 - \varepsilon < 1$ for some $\varepsilon > 0$ and $\widetilde{\delta} > 0$, one has

$$\begin{aligned}
\mathbb{P}_{\text{fail}} \ge\, & \Big[1 - \exp\Big(-\frac{m^{1 - I(a_\tau, b_\tau, c_\tau) + \widetilde{\delta}}}{2 \log^3(m)} \Big) \Big]^2 \cdot \Big(1 - \frac{m^{-1+o(1)}}{\log^3(m)} \Big) \\
\ge\, & 1 - 2 \exp\Big(-\frac{m^{\varepsilon + \widetilde{\delta}}}{2 \log^3(m)} \Big) - \frac{m^{-1+o(1)}}{\log^3(m)} - \exp\Big(-\frac{m^{\varepsilon + \widetilde{\delta}}}{2 \log^3(m)} \Big) \cdot \frac{m^{-1+o(1)}}{\log^3(m)} \overset{m \to \infty}{\to} 1.
\end{aligned}$$

Therefore, the with probability tending to 1, the best estimator MLE (MAP) fails exact recovery, hence no other algorithm could succeed.

A.2. Information-theoretic lower bounds

Proof of Theorem 3.3. For each node $i \in \mathcal{V}_{\mathbb{U}}$, denote $f(\cdot | \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \widetilde{\boldsymbol{y}}_{\mathbb{U},-i}) = \mathbb{P}(y_i = \cdot | \boldsymbol{A} = \widetilde{\boldsymbol{A}}, \boldsymbol{X} = \widetilde{\boldsymbol{X}}, \boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U},-i} = \widetilde{\boldsymbol{y}}_{\mathbb{U},-i})$. Due to the symmetry of the problem, vertices are interchangeable if in the same community, then it suffices to consider the following event

$$\mathcal{A} = \big\{ f(y_1 | \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \widetilde{\boldsymbol{y}}_{\mathbb{U},-1}) < f(-y_1 | \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \widetilde{\boldsymbol{y}}_{\mathbb{U},-1}) \big\}.$$

By Lemma F.3 in (Abbe et al., 2022) and symmetry between vertices, for any sequence of estimators $\widehat{\boldsymbol{y}}_{\mathbb{U}}$, the following holds

$$\mathbb{E} \psi_m(\widehat{\boldsymbol{y}}_{\mathbb{U}}, \boldsymbol{y}_{\mathbb{U}}) \ge \frac{m-1}{3m-1} \cdot \mathbb{P}(\mathcal{A}).$$

Recall the definition of $W_{m,u}$ in (A.3). We denote $W_{m,1}$ by taking $u = 1$ and $\mathbb{S} = [N] \setminus \{u\}$. Define the following two events

$$\mathcal{B}_\varepsilon := \Big\{ \Big| \log\Big(\frac{f(y_1 | \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \widetilde{\boldsymbol{y}}_{\mathbb{U},-1})}{f(-y_1 | \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \widetilde{\boldsymbol{y}}_{\mathbb{U},-1})} \Big) - W_{m,1} \Big| < \varepsilon q_m \Big\}, \quad \mathcal{C}_\varepsilon = \Big\{ W_{m,1} \le -\varepsilon q_m \Big\}.$$

By triangle inequality, $\mathcal{B}_\varepsilon \cap \mathcal{C}_\varepsilon$ implies $\mathcal{A}$, thus $\mathcal{B}_\varepsilon \cap \mathcal{C}_\varepsilon \subset \mathcal{A}$, and

$$\mathbb{E}\psi_m(\widehat{\boldsymbol{y}}_{\mathbb{U}}, \boldsymbol{y}_{\mathbb{U}}) \gtrsim \mathbb{P}(\mathcal{A}) \geq \mathbb{P}(\mathcal{B}_\varepsilon \cap \mathcal{C}_\varepsilon) \geq \mathbb{P}(\mathcal{C}_\varepsilon) - \mathbb{P}(\mathcal{B}_\varepsilon^{\mathrm{c}}).$$

According to Lemma B.1, $\mathbb{P}(\mathcal{B}_\varepsilon^{\mathrm{c}}) \ll e^{-q_m}$. Together with the results above, and by Lemma A.1, we have

$$\liminf_{m\to\infty} q_m^{-1} \log \mathbb{E}\eta_m(\widehat{\boldsymbol{y}}_{\mathbb{U}}, \boldsymbol{y}_{\mathbb{U}}) \geq -\sup_{t\in\mathbb{R}}\{\varepsilon t + I(a, b, c_\tau, \tau)\},$$

and the proof is finished by taking $\varepsilon \to 0$. □

A.3. Deferred proofs

For the sake of convenience, we introduce the following notations for the remaining of this section. For some realization $\boldsymbol{A} = \widetilde{\boldsymbol{A}}$, $\boldsymbol{X} = \widetilde{\boldsymbol{X}}$, $\boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}} \in \{\pm 1\}^n$ and $\boldsymbol{\mu} = \widetilde{\boldsymbol{\mu}}$, we write

$$\mathbb{P}(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{y}}_{\mathbb{L}}) = \mathbb{P}(\boldsymbol{A} = \widetilde{\boldsymbol{A}}, \boldsymbol{X} = \widetilde{\boldsymbol{X}}, \boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}), \quad \mathbb{P}(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{X}} | \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) = \mathbb{P}(\boldsymbol{A} = \widetilde{\boldsymbol{A}}, \boldsymbol{X} = \widetilde{\boldsymbol{X}} | \boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z})$$
$$\mathbb{P}(\widetilde{\boldsymbol{A}} | \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) = \mathbb{P}(\boldsymbol{A} = \widetilde{\boldsymbol{A}} | \boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}), \quad \mathbb{P}(\widetilde{\boldsymbol{X}} | \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) = \mathbb{P}(\boldsymbol{X} = \widetilde{\boldsymbol{X}} | \boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}).$$

Lemma A.3. *The MAP estimator minimizes the* $\mathbb{P}_{\text{fail}}$*, and MAP is equivalent to the MLE* (A.1)*. The quantity that MLE is maximizing is defined in* (A.2)*.*

Proof of Lemma A.3. From model 2.5, independently, $\boldsymbol{y}_{\mathbb{L}}$ and $\boldsymbol{y}_{\mathbb{U}}$ are uniformly distributed over the spaces $\{\pm 1\}^m$ and $\{\pm 1\}^m$ respectively, thus the following factorization holds

$$\mathbb{P}(\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) \coloneqq \mathbb{P}(\boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) = \mathbb{P}(\boldsymbol{y}_{\mathbb{L}} = \widetilde{\boldsymbol{y}}_{\mathbb{L}}) \cdot \mathbb{P}(\boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}),$$

which is some constant irrelevant to $\boldsymbol{z}$. The first sentence of the Lemma can be established by Bayes Theorem, since

$$\begin{aligned}
\widehat{\boldsymbol{y}}_{\mathrm{MAP}} &= \operatorname*{arg\,max}_{\boldsymbol{z}\in\{\pm1\}^m, \mathbf{1}^\top \boldsymbol{z}=0} \mathbb{P}(\boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z} | \boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}}) = \operatorname*{arg\,max}_{\boldsymbol{z}\in\{\pm1\}^m, \mathbf{1}^\top \boldsymbol{z}=0} \frac{\mathbb{P}(\boldsymbol{A}, \boldsymbol{X} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) \cdot \mathbb{P}(\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z})}{\mathbb{P}(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}})} \\
&= \operatorname*{arg\,max}_{\boldsymbol{z}\in\{\pm1\}^m, \mathbf{1}^\top \boldsymbol{z}=0} \mathbb{P}(\boldsymbol{A}, \boldsymbol{X} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) = \operatorname*{arg\,max}_{\boldsymbol{z}\in\{\pm1\}^m, \mathbf{1}^\top \boldsymbol{z}=0} \mathbb{P}(\boldsymbol{A} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z}) \cdot \mathbb{P}(\boldsymbol{X} | \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z})
\end{aligned}$$

where $\mathbb{P}(\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{z})$ and $\mathbb{P}(\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{\mathbb{L}})$ in the first line are factored out since they are irrelevant to $\boldsymbol{z}$, and the last equality holds due to the independence between $\boldsymbol{A}$ and $\boldsymbol{X}$ when given $\boldsymbol{y}$. For the second sentence of the Lemma, the function $f(\boldsymbol{z})$ could be easily obtained by taking the logarithm of the objectve probability. □

Proof of Lemma A.1 (1). By definition of $W_{m,i}$, we have

$$\begin{aligned}
\mathbb{E}[e^{tW_{m,i}} | y_i] =\ & \mathbb{E}\left[\exp\left(\frac{2t\theta^2}{N\theta^2 + d} \sum_{j\in[N]\setminus\{i\}} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle y_j y_i\right) \middle| y_i\right] \\
& \cdot \mathbb{E}\left[\exp\left(t \log(a/b) \sum_{j\in\mathcal{V}_{\mathbb{U}}\setminus\{i\}} A_{ij} y_j y_i\right) \middle| y_i\right] \cdot \mathbb{E}\left[\exp\left(t \log(a/b) \sum_{j\in\mathcal{V}_{\mathbb{L}}} A_{ij} y_j y_i\right) \middle| y_i\right].
\end{aligned}$$

Following the same calculation as Lemma F.2 in (Abbe et al., 2022), we know that

$$\begin{aligned}
\log \mathbb{E}\left[\exp\left(\frac{2t\theta^2}{N\theta^2 + d} \sum_{j\in[N]\setminus\{i\}} \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle y_j y_i\right) \middle| y_i\right] &= 2c_\tau (t + t^2)(1 + o(1)) q_m \\
\log \mathbb{E}\left[\exp\left(t \log(a/b) \sum_{j\in\mathcal{V}_{\mathbb{U}}\setminus\{i\}} A_{ij} y_j y_i\right) \middle| y_i\right] &= \frac{1}{2}\left(-a + a\left(\frac{a}{b}\right)^t - b + b\left(\frac{b}{a}\right)^t\right)(1 + o(1)) q_m.
\end{aligned}$$

Meanwhile, since

$$\mathbb{E}[e^{-tA_{ij}y_iy_j}|y_i] = 1 + \frac{1}{2}\big(\alpha(e^t - 1) + \beta(e^{-t} - 1)\big),$$

we can get

$$\begin{aligned}\mathbb{E}\left[\exp\left(t\log(a/b)\sum_{j\in\mathcal{V}_{\mathbb{L}}} A_{ij}y_jy_i\right)\middle|\, y_i\right] &= n\log\left(1 + \frac{1}{2}\big(\alpha(e^t - 1) + \beta(e^{-t} - 1)\big)\right)\\ &= \frac{n}{2m}\left(a\left(\frac{a}{b}\right)^t - a + b\left(\frac{b}{a}\right)^t - b\right)(1 + o(1))q_m.\end{aligned}$$

Thus by using $\log(1 + x) = x$ when $x = o(1)$, we obtain

$$\begin{aligned}&\lim_{m\to\infty} q_m^{-1}\log\mathbb{E}[e^{tW_{m,i}}|y_i] = \lim_{m\to\infty}\left(2c_\tau(t + t^2) + \frac{N}{2m}\left(a\left(\frac{a}{b}\right)^t - a + b\left(\frac{b}{a}\right)^t - b\right)\right)(1 + o(1))\\ &= -I(t, a_\tau, b_\tau, c_\tau)(1 + o(1)).\end{aligned}$$

The proof is then completed by applying Lemma H.5 in (Abbe et al., 2022). □

Proof of Lemma A.1 (2). First, we plug in σ, $\boldsymbol{y}_{\mathbb{U}}$ into (A.2), and consider the effect of u and v,

$$\begin{aligned}f(\boldsymbol{y}_{\mathbb{U}}) - f(\boldsymbol{\sigma}) &= \log\mathbb{P}(\boldsymbol{A}|\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{\sigma}) - \log\mathbb{P}(\boldsymbol{A}|\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}}) + \log\mathbb{P}(\boldsymbol{X}|\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}} = \boldsymbol{\sigma}) - \log\mathbb{P}(\boldsymbol{X}|\boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}})\\ &= \log\mathbb{P}(\boldsymbol{A}|y_u = 1, y_v = -1, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}}) - \log\mathbb{P}(\boldsymbol{A}|y_u = -1, y_v = 1, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}})\\ &\quad + \log\mathbb{P}(\boldsymbol{X}|y_u = 1, y_v = -1, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}}) - \log\mathbb{P}(\boldsymbol{X}|y_u = -1, y_v = 1, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}}).\end{aligned}$$

By Lemma A.4, the term above can be further reformulated as

$$\begin{aligned}f(\boldsymbol{y}_{\mathbb{U}}) - f(\boldsymbol{\sigma}) &= \log\left(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_u, \boldsymbol{y}_{-u})}\right) + \log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{X}}(\boldsymbol{X}|-y_u, \boldsymbol{y}_{-u})}\right)\\ &\quad + \log\left(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_v, \boldsymbol{y}_{-v})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_v, \boldsymbol{y}_{-v})}\right) + \log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_v, \boldsymbol{y}_{-v})}{p_{\boldsymbol{X}}(\boldsymbol{X}|-y_v, \boldsymbol{y}_{-v})}\right).\end{aligned}$$

Note that $y_u^2 = 1$, according to Lemmas A.5 and A.6, with probability at least $1 - e^{-q_m}$, we have

$$\begin{aligned}&\left|\log\left(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_u, \boldsymbol{y}_{-u})}\right) + \log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{X}}(\boldsymbol{X}|-y_u, \boldsymbol{y}_{-u})}\right) - y_u\log\left(\frac{a}{b}\right)\sum_{j\neq i} A_{ij}y_j - y_u\frac{2\theta^2}{N\theta^2 + d}\sum_{j\neq i}\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle y_j\right|\\ \le &\left|\log\left(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_u, \boldsymbol{y}_{-u})}\right) - y_u\log\left(\frac{a}{b}\right)\sum_{j\neq i} A_{ij}y_j\right| + \left|\log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{X}}(\boldsymbol{X}|-y_u, \boldsymbol{y}_{-u})}\right) - y_u\frac{2\theta^2}{N\theta^2 + d}\sum_{j\neq i}\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle y_j\right|\\ \ll &\, q_m.\end{aligned}$$

Consequently by triangle inequality, there exists some large enough constant $c > 0$ such that with probability at least $1 - e^{-cq_m}$,

$$|f(\boldsymbol{y}_{\mathbb{U}}) - f(\boldsymbol{\sigma}) - W_{m,u} - W_{m,v}|/q_m = o(1),$$

The proof is then completed. □

Proof of Lemma A.2. First, $J_m \coloneqq \max_{u\in\mathcal{V}_{\mathbb{U}}} J_{m,u}$ in (A.5), then it suffices to focus on $\mathbb{P}(J_{m,u} > \zeta_m q_m)$, since an argument based on the union bound leads to

$$\mathbb{P}(J_m > \zeta_m q_m) = \mathbb{P}(\exists u \in \mathcal{V}_{\mathbb{U}} \text{ s.t. } J_{m,u} > \zeta_m q_m) \le \sum_{u\in\mathcal{V}_{\mathbb{U}}}\mathbb{P}(J_{m,u} > \zeta_m q_m).$$

We claim the following upper bound with the proof deferred later

$$\mathbb{E}J_{m,u} \le \exp(q_m \log^{-3} m). \tag{A.6}$$

Then by Markov inequality and the fact $q_m = \log(m)$, one has $\mathbb{P}(J_{m,u} > \zeta_m q_m) \le \mathbb{E}J_{m,u}/(\zeta_m q_m) \asymp n^{-2+o(1)}$. Thus by the union bound

$$\mathbb{P}(J_m \le \zeta_m q_m) = 1 - \mathbb{P}(J_m > \zeta_m q_m) \ge 1 - \gamma m \cdot m^{-2+o(1)} = 1 - \log^{-3}(m) \cdot m^{-1+o(1)}.$$

For the second desired inequality, note that the difference between $U_{m,u}$ (A.5) and $W_{m,u}$ (A.3) is relatively negligible since $|\mathcal{U}|/m = \log^{-3}(m) = o(1)$, thus $U_{m,u}$ exhibits the same concentration behavior as $W_{m,u}$. One could follow the same calculation as in Lemma A.1 to figure out that for any $m \ge m_0$ and $\widetilde{\delta} > 0$, with $I(t, a_\tau, b_\tau, c_\tau)$ defined in (A.4), the following holds

$$\mathbb{P}(U_{m,u} \le -\zeta_m q_m) = \exp\big(- q_m \cdot (-\widetilde{\delta} + \sup_{t\in\mathbb{R}}\{-\zeta_m t + I(t, a_\tau, b_\tau, c_\tau)\})\big).$$

Note that $\{U_{m,u}\}_{u\in\mathcal{V}_{\mathbb{U},+}\cap\mathcal{U}}$ is a set of independent for different since no edge overlap, then one has

$$\begin{aligned}\mathbb{P}\Big(\cap_{u\in\mathcal{V}_{\mathbb{U},+}\cap\mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\}\Big) &= \prod_{u\in\mathcal{V}_{\mathbb{U},+}\cap\mathcal{U}} \mathbb{P}(U_{m,u} > -\zeta_m q_m) \\ &= \Big(1 - m^{-I(a_\tau,b_\tau,c_\tau)+\widetilde{\delta}}\Big)^{\delta m/2} \le \exp\Big(- \frac{m^{1-I(a_\tau,b_\tau,c_\tau)+\widetilde{\delta}}}{2\log^3(m)}\Big),\end{aligned}$$

where the last inequality holds since $1 - x \le e^{-x}$, and it leads to our desired result

$$\mathbb{P}\Big(\cup_{u\in\mathcal{V}_{\mathbb{U},+}\cap\mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\}\Big) = 1 - \mathbb{P}\Big(\cap_{u\in\mathcal{V}_{\mathbb{U},+}\cap\mathcal{U}} \{U_{m,u} \le -\zeta_m q_m\}\Big) \ge 1 - \exp\Big(- \frac{m^{1-I(a_\tau,b_\tau,c_\tau)+\widetilde{\delta}}}{2\log^3(m)}\Big).$$

We now establish the proof of (A.6). Recall that $J_{m,u}$ (A.5) is a summation of independent random variables, where the number of such type random variables is at most $|\mathcal{U}| = \gamma m \asymp m\log^{-3}(m)$. Denote $\widehat{\boldsymbol{\mu}}^{(-u)} = \frac{1}{|\mathcal{U}|-1}\sum_{j\in\mathcal{U}\setminus\{u\}} \boldsymbol{x}_j y_j$. Recall $\boldsymbol{x}_i = \theta y_i \boldsymbol{\mu} + \boldsymbol{z}_i$ with $\|\boldsymbol{\mu}\|_2 = 1$ in (2.1), then $y_i\boldsymbol{x}_i \sim \mathcal{N}(\theta\boldsymbol{\mu}, \boldsymbol{I}_d)$ given y_i, and $\sqrt{|\mathcal{U}|-1}\ \widehat{\boldsymbol{\mu}}^{(-u)} \sim \mathcal{N}(\sqrt{|\mathcal{U}|-1}\ \theta\boldsymbol{\mu}, \boldsymbol{I}_d)$, while $y_i\boldsymbol{x}_i$ and $\sqrt{|\mathcal{U}|-1}\ \widehat{\boldsymbol{\mu}}^{(-u)}$ are independent. Following Lemma H.4 in (Abbe et al., 2022), for all $t \in (-\sqrt{|\mathcal{U}|-1}, \sqrt{|\mathcal{U}|-1})$, one has

$$\begin{aligned}&\log\mathbb{E}\big(\exp(t\langle \boldsymbol{x}_u, \widehat{\boldsymbol{\mu}}^{(-u)}\rangle y_u)|y_u\big) = \log\mathbb{E}\big(\exp(t/\sqrt{|\mathcal{U}|-1}\langle \boldsymbol{x}_u, \sqrt{|\mathcal{U}|-1}\ \widehat{\boldsymbol{\mu}}^{(-u)}\rangle y_u)|y_u\big) \\ =& \frac{\frac{t^2}{|\mathcal{U}|-1}}{2(1-\frac{t^2}{|\mathcal{U}|-1})}\big(\theta^2\|\boldsymbol{\mu}\|_2^2 + (|\mathcal{U}|-1)\cdot\theta^2\|\boldsymbol{\mu}\|_2^2\big) + \frac{\frac{t}{\sqrt{|\mathcal{U}|-1}}}{1-\frac{t^2}{|\mathcal{U}|-1}}\theta^2\langle\boldsymbol{\mu}, \sqrt{|\mathcal{U}|-1}\boldsymbol{\mu}\rangle - \frac{d}{2}\log\Big(1 - \frac{t^2}{|\mathcal{U}|-1}\Big) \\ =& \frac{t\theta^2}{1-\frac{t^2}{|\mathcal{U}|-1}}\Big(1 + \frac{|\mathcal{U}|t}{2(|\mathcal{U}|-1)}\Big) - \frac{d}{2}\log\Big(1 - \frac{t^2}{|\mathcal{U}|-1}\Big) = \log\mathbb{E}\big(\exp(t\langle \boldsymbol{x}_u, \widehat{\boldsymbol{\mu}}^{(-u)}\rangle y_u\big),\end{aligned}$$

where the last inequality holds since the result above is independent of y_u. We substitute $s = 2t\tilde{p}/\theta^2$, where $\tilde{p} = \theta^4(|\mathcal{U}|-1)/(N\theta^2+d)$. We focus on the critical case $\theta^2 \asymp q_m \asymp \log(m)$, $|\mathcal{U}| = m\log^{-3} m$, $d/N = \gamma \asymp 1$, thus $s^2/(|\mathcal{U}|-1) = m^{-1}\log^{-3}(m) = o(1)$, $\log(1-x) = -x$ for $x = o(1)$, then

$$\begin{aligned}&\log\mathbb{E}\big(\exp(s\langle \boldsymbol{x}_u, \widehat{\boldsymbol{\mu}}^{(-u)}\rangle y_u\big) = \frac{s\theta^2}{1-\frac{s^2}{|\mathcal{U}|-1}}\Big(1 + \frac{|\mathcal{U}|s}{2(|\mathcal{U}|-1)}\Big) - \frac{d}{2}\log\Big(1 - \frac{s^2}{|\mathcal{U}|-1}\Big) \\ =& [1+o(1)]s\theta^2(1+s/2) + \frac{d}{2}\cdot\frac{s^2}{|\mathcal{U}|-1} = [1+o(1)]\cdot\Big[2t\tilde{p}(1+\frac{t\tilde{p}}{\theta^2}) + \frac{d}{2(|\mathcal{U}|-1)}\cdot\frac{4t^2\tilde{p}^2}{\theta^4}\Big] \\ =& [1+o(1)]\cdot 2\tilde{p}t\Big[1 + \frac{t\tilde{p}}{\theta^2}\Big(1 + \frac{d}{\theta^2(|\mathcal{U}|-1)}\Big)\Big] = [1+o(1)]\cdot 2\tilde{p}t\Big[\Big(1 + t\,\frac{d+\theta^2(|\mathcal{U}|-1)}{d+N\theta^2}\Big)\Big] \\ =& [1+o(1)]\cdot 2\tilde{p}t\big(1 + t(1-\tau)\log^{-3}(m)\big).\end{aligned}$$

By (3.1), we focus on the critical case $\theta^2 \asymp q_m \asymp \log(m)$, $|\mathcal{U}| = m\log^{-3} m$, $d/N = \gamma \asymp 1$, then

$$c_\tau q_m = \frac{\theta^4}{\theta^2 + (1-\tau)d/m} \asymp q_m, \quad \tilde{p} = \frac{\theta^4(|\mathcal{U}|-1)}{N\theta^2 + d} \asymp \log^{-2}(m) \asymp c_\tau q_m \cdot \log^{-3}(m),$$

which leads to

$$\begin{aligned}&\log \mathbb{E}\exp\Big(\frac{2}{N + d/\theta^2}\sum_{j\in\mathcal{U}\setminus\{u\}}\langle \boldsymbol{x}_u, \boldsymbol{x}_j\rangle y_j\Big)\\ &= \log\mathbb{E}\big(\exp(s\langle \boldsymbol{x}_u, \widehat{\boldsymbol{\mu}}^{(-u)}\rangle y_u) = 2c_\tau\big(t + t^2(1-\tau)\log^{-3}(m)\big)\log^{-3}(m)\cdot q_m.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\mathbb{E}[e^{-tA_{uj}y_uy_j}|y_u] &= \frac{1}{2}\mathbb{E}[e^{tA_{uj}}|y_uy_j = 1] + \frac{1}{2}\mathbb{E}[e^{-tA_{uj}}|y_uy_j = -1] = \frac{1}{2}[\alpha e^t + (1-\alpha)] + \frac{1}{2}[\beta e^t + (1-\beta)]\\ &= 1 + \frac{\alpha(e^t-1) + \beta(e^{-t}-1)}{2} = \mathbb{E}[e^{-tA_{uj}y_uy_j}],\end{aligned}$$

where the last equality holds since the result on the second line does not depend on y_u again. Conditioned on y_u, $\{A_{uj}y_uy_j\}_{j\neq u}$ are i.i.d. random variables, then followed by $\alpha = aq_m/m$, $\beta = bq_m/m$ and $\log(1+x) = x$ for $x = o(1)$, we have

$$\begin{aligned}&\log\mathbb{E}\Big[\exp\Big(t\log(a/b)y_u\sum_{j\in\mathcal{U}\setminus\{u\}}A_{uj}y_j\Big)\Big] = (|\mathcal{U}|-1)\cdot\log\Big(1 + \frac{aq_m[(a/b)^t - 1] + bq_m[(b/a)^t-1]}{2m}\Big)\\ &= (1+o(1))\frac{a[(a/b)^t-1] + b[(b/a)^t-1]}{2}\log^{-3}(m)\cdot q_m.\end{aligned}$$

Due to the independence between the graph $\boldsymbol{A}$ and feature vectors $\boldsymbol{X}$ conditioned on y_u, one has

$$\begin{aligned}&q_m^{-1}\log\mathbb{E}e^{tJ_{m,u}} = q_m^{-1}\log\mathbb{E}\exp\Big(\frac{2}{N + d/\theta^2}\sum_{j\in\mathcal{U}\setminus\{u\}}\langle \boldsymbol{x}_u, \boldsymbol{x}_j\rangle y_j\Big) + q_m^{-1}\log\mathbb{E}\Big[\exp\Big(t\log(a/b)y_u\sum_{j\in\mathcal{U}\setminus\{u\}}A_{uj}y_j\Big)\Big]\\ &= [1+o(1)]\cdot\Big[\frac{a[(a/b)^t-1] + b[(b/a)^t-1]}{2} + 2c_\tau\big(t + t^2(1-\tau)\log^{-3}(m)\big)\Big]\log^{-3}(m) \asymp \log^{-3}(m),\end{aligned}$$

where the last line holds since $a, b, c \asymp 1$. The proof of (A.6) is then established once the large deviation results from graph $\boldsymbol{A}$ and feature matrix $\boldsymbol{X}$ are added together. □

Lemma A.4. *Denote by $p_{\boldsymbol{A}}(\cdot|\widetilde{\ell}_i, \widetilde{\boldsymbol{y}}_{-i})$ the conditional probability mass function of $\boldsymbol{A}$ given $y_i = \widetilde{\ell}_i \in \{\pm 1\}$ and $\boldsymbol{y}_{-i} = \widetilde{\boldsymbol{y}}_{-i} \in \{\pm1\}^{N-1}$. Denote by $p_{\boldsymbol{X}}(\cdot|\widetilde{\ell}_i, \widetilde{\boldsymbol{y}}_{-i})$ the conditional probability density function of $\boldsymbol{X}$ given $y_i = \widetilde{\ell}_i \in \{\pm1\}$ and $\boldsymbol{y}_{-i} = \widetilde{\boldsymbol{y}}_{-i} \in \{\pm1\}^{N-1}$. Then*

$$\begin{aligned}\log\Big(\frac{\mathbb{P}(\boldsymbol{A}|y_u, y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}})}{\mathbb{P}(\boldsymbol{A}|-y_u, -y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}})}\Big) &= \log\Big(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_u, \boldsymbol{y}_{-u})}\Big) + \log\Big(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_v, \boldsymbol{y}_{-v})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_v, \boldsymbol{y}_{-v})}\Big),\\ \log\Big(\frac{\mathbb{P}(\boldsymbol{X}|y_u, y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}})}{\mathbb{P}(\boldsymbol{X}|-y_u, -y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}})}\Big) &= \log\Big(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{X}}(\boldsymbol{X}|-y_u, \boldsymbol{y}_{-u})}\Big) + \log\Big(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_v, \boldsymbol{y}_{-v})}{p_{\boldsymbol{X}}(\boldsymbol{X}|-y_v, \boldsymbol{y}_{-v})}\Big).\end{aligned}$$

Proof of Lemma A.4. We start with the graph part. For each vertex $u \in \mathcal{V}_{\mathbb{U}}$, denote $\mathcal{T}_u = \{j \in [N]\setminus u : y_uy_j = 1\}$ and $\mathcal{S}_u = \{j \in [N]\setminus u : A_{uj} = 1\}$, then

$$\begin{aligned}p_{\boldsymbol{A}}(\boldsymbol{A}|y_u, \boldsymbol{y}_{-u}) &\propto \alpha^{|\mathcal{T}_u\cap\mathcal{S}_u|}\cdot(1-\alpha)^{|\mathcal{T}_u\setminus\mathcal{S}_u|}\cdot\beta^{|\mathcal{S}_u\setminus\mathcal{T}_u|}\cdot(1-\beta)^{|[N]\setminus(\mathcal{T}_u\cup\mathcal{S}_u\cup\{u\})|},\\ p_{\boldsymbol{A}}(\boldsymbol{A}|-y_u, \boldsymbol{y}_{-u}) &\propto \alpha^{|\mathcal{S}_u\setminus\mathcal{T}_u|}\cdot(1-\alpha)^{|[N]\setminus(\mathcal{T}_u\cup\mathcal{S}_u\cup\{u\})|}\cdot\beta^{|\mathcal{T}_u\cap\mathcal{S}_u|}\cdot(1-\beta)^{|\mathcal{T}_u\setminus\mathcal{S}_u|},\end{aligned}$$

where $\propto$ hides the factor not involving $\{A_{uj}\}_{j=1}^N$ and y_u. Then

$$\log\Big(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{A}}(\boldsymbol{A}|-y_u, \boldsymbol{y}_{-u})}\Big) = \Big(|\mathcal{T}_u\cap\mathcal{S}_u| - |\mathcal{S}_u\setminus\mathcal{T}_u|\Big)\log\Big(\frac{\alpha}{\beta}\Big) + \Big(|\mathcal{T}_u\setminus\mathcal{S}_u| - |[N]\setminus(\mathcal{T}_u\cup\mathcal{S}_u\cup\{u\})|\Big)\log\Big(\frac{1-\alpha}{1-\beta}\Big).$$

For the left hand side, we assume $y_u = 1, y_v = -1$ and factor out the terms irrelevant to u and v, then

$$\mathbb{P}(\boldsymbol{A}|y_u, y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}}) \propto \beta^{A_{uv}}(1-\beta)^{1-A_{uv}} \cdot \alpha^{|\mathcal{T}_u \cap \mathcal{S}_u|} \cdot (1-\alpha)^{|\mathcal{T}_u \setminus \mathcal{S}_u|} \cdot \beta^{|\mathcal{S}_u \setminus \mathcal{T}_u|} \cdot (1-\beta)^{|[N]\setminus(\mathcal{T}_u \cup \mathcal{S}_u \cup \{u\})|} \cdot \alpha^{|\mathcal{T}_v \cap \mathcal{S}_v|} \cdot (1-\alpha)^{|\mathcal{T}_v \setminus \mathcal{S}_v|} \cdot \beta^{|\mathcal{S}_v \setminus \mathcal{T}_v|} \cdot (1-\beta)^{|[N]\setminus(\mathcal{T}_v \cup \mathcal{S}_v \cup \{v\})|}.$$

We perform the same calculation under the assumption $y_u = -1, y_v = 1$, which gives

$$\mathbb{P}(\boldsymbol{A}| - y_u, -y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}}) \propto \beta^{A_{uv}}(1-\beta)^{1-A_{uv}} \cdot \alpha^{|\mathcal{S}_u \setminus \mathcal{T}_u|} \cdot (1-\alpha)^{|[N]\setminus(\mathcal{T}_u \cup \mathcal{S}_u \cup \{u\})|} \cdot \beta^{|\mathcal{T}_u \cap \mathcal{S}_u|} \cdot (1-\beta)^{|\mathcal{T}_u \setminus \mathcal{S}_u|} \cdot \alpha^{|\mathcal{S}_v \setminus \mathcal{T}_v|} \cdot (1-\alpha)^{|[N]\setminus(\mathcal{T}_v \cup \mathcal{S}_v \cup \{v\})|} \cdot \beta^{|\mathcal{T}_v \cap \mathcal{S}_v|} \cdot (1-\beta)^{|\mathcal{T}_v \setminus \mathcal{S}_v|},$$

where the probability of generating edge (u, v) remains unchanged when flipping the signs of u and v at the same time. The proof follows easily by rearranging and separating relevant terms.

For the second part, note that

$$p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{y}) \propto \mathbb{E}_{\boldsymbol{\mu}} \exp\Big(-\frac{1}{2}\sum_{j\in\mathcal{V}} \|\boldsymbol{x}_j - y_j\boldsymbol{\mu}\|_2^2\Big) \propto \mathbb{E}_{\boldsymbol{\mu}} \exp\Big(\Big\langle \sum_{j\in\mathcal{V}} \boldsymbol{x}_j y_j, \boldsymbol{\mu}\Big\rangle\Big),$$

where $\propto$ hides quantities that do not depend on $\boldsymbol{y}$. Consequently,

$$\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_u, \boldsymbol{y}_{-u})}{p_{\boldsymbol{X}}(\boldsymbol{X}| - y_u, \boldsymbol{y}_{-u})} = \mathbb{E}_{\boldsymbol{\mu}} \exp(2y_u \boldsymbol{x}_u^\top \boldsymbol{\mu}).$$

For the left hand side, similarly, let $\propto$ hide the quantities independent of u, v, then

$$\mathbb{P}(\boldsymbol{X}|y_u, y_v, \boldsymbol{y}_{\mathbb{L}}, \boldsymbol{y}_{\mathbb{U}\setminus\{u,v\}}) \propto \mathbb{E}_{\boldsymbol{\mu}} \exp(y_u \boldsymbol{x}_u^\top \boldsymbol{\mu} + y_v \boldsymbol{x}_v^\top \boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{\mu}} \exp(y_u \boldsymbol{x}_u^\top \boldsymbol{\mu}) \cdot \mathbb{E}_{\boldsymbol{\mu}} \exp(y_v \boldsymbol{x}_v^\top \boldsymbol{\mu}).$$

The conclusion follows easily by the linearity of expectation. □

Lemma A.5 (Lemma F.4, (Abbe et al., 2022))**.** *Denote by $p_{\boldsymbol{X}}(\cdot|\widetilde{\ell}_i, \widetilde{\boldsymbol{y}}_{-i})$ the conditional probability density function of $\boldsymbol{X}$ given $y_i = \widetilde{\ell}_i \in \{\pm 1\}$ and $\boldsymbol{y}_{-i} = \widetilde{\boldsymbol{y}}_{-i} \in \{\pm 1\}^{N-1}$. Then there exists some large enough constant $c > 0$ such that for each $i \in [N]$, with probability at least $1 - e^{-cq_N}$,*

$$\left| y_i \log\left(\frac{p_{\boldsymbol{X}}(\boldsymbol{X}|y_i, \boldsymbol{y}_{-i})}{p_{\boldsymbol{X}}(\boldsymbol{X}|y_i, \boldsymbol{y}_{-i})}\right) - \frac{2\theta^2}{N\theta^2 + d}\sum_{j\neq i}\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle y_j \right| / q_N = o(1).$$

Lemma A.6 (Lemma F.5, (Abbe et al., 2022))**.** *Denote by $p_{\boldsymbol{A}}(\cdot|\widetilde{\ell}_i, \widetilde{\boldsymbol{y}}_{-i})$ the conditional probability mass function of $\boldsymbol{A}$ given $y_i = \widetilde{\ell}_i \in \{\pm 1\}$ and $\boldsymbol{y}_{-i} = \widetilde{\boldsymbol{y}}_{-i} \in \{\pm 1\}^{N-1}$. Then there exists some large enough constant $c > 0$ such that for each $i \in [N]$, with probability at least $1 - e^{-cq_N}$,*

$$\left| y_i \log\left(\frac{p_{\boldsymbol{A}}(\boldsymbol{A}|y_i, \boldsymbol{y}_{-i})}{p_{\boldsymbol{X}}(\boldsymbol{A}|y_i, \boldsymbol{y}_{-i})}\right) - \log\left(\frac{a}{b}\right)\sum_{j\neq i} A_{ij} y_j \right| / q_N = o(1).$$

B. Performance of optimal spectral estimator

According to (Abbe et al., 2022) and the discussion in Section 3.1, the ideal estimator for the label of the node $i \in \mathcal{V}_{\mathbb{U}}$ could be derived from

$$\widehat{y}_i^{\text{genie}} = \arg\max_{y=\pm 1} \mathbb{P}(y_i = y|\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{-i})$$

Lemma B.1. *For each given $i \in \mathcal{V}_{\mathbb{U}}$, following the $o_{\mathbb{P}}(q_m; q_m)$ notation in (Abbe et al., 2022), we have*

$$\left| \log\Big(\frac{\mathbb{P}(y_i = 1|\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{-i})}{\mathbb{P}(y_i = -1|\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{-i})}\Big) - \Big[\log\Big(\frac{a}{b}\Big)\boldsymbol{A}\boldsymbol{y} + \frac{2}{N + d/\theta^2}\boldsymbol{G}\boldsymbol{y}\Big]_i \right| = o_{\mathbb{P}}(q_m; q_m).$$

Proof of Lemma B.1. By definition of conditional probability and the independence between $\boldsymbol{A}|\boldsymbol{y}$ and $\boldsymbol{X}|\boldsymbol{y}$, for vertex $i \in \mathcal{V}_{\mathbb{U}}$, one has

$$\log\Big(\frac{\mathbb{P}(y_i = 1|\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{-i})}{\mathbb{P}(y_i = -1|\boldsymbol{A}, \boldsymbol{X}, \boldsymbol{y}_{-i})}\Big) = \log\Big(\frac{\mathbb{P}(\boldsymbol{A}, \boldsymbol{X}|y_i = 1, \boldsymbol{y}_{-i})}{\mathbb{P}(\boldsymbol{A}, \boldsymbol{X}|y_i = -1, \boldsymbol{y}_{-i})}\Big)$$
$$= \log\Big(\frac{\mathbb{P}(\boldsymbol{A}|y_i = 1, \boldsymbol{y}_{-i})}{\mathbb{P}(\boldsymbol{A}|y_i = -1, \boldsymbol{y}_{-i})}\Big) + \log\Big(\frac{\mathbb{P}(\boldsymbol{X}|y_i = 1, \boldsymbol{y}_{-i})}{\mathbb{P}(\boldsymbol{X}|y_i = -1, \boldsymbol{y}_{-i})}\Big).$$

Then, one could apply Lemmas F.4, F.5 in (Abbe et al., 2022) separately to conclude the results for the two terms above. □

From Lemma B.1 above, the ideal estimator $(\widehat{y}_1^{\mathrm{genie}}, \dots, \widehat{y}_m^{\mathrm{genie}})^\top$ can be approximated by

$$\mathrm{sign}\left(\log(a/b)(\boldsymbol{A}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}} + \boldsymbol{A}_{\mathbb{U}}\boldsymbol{y}_{\mathbb{U}}) + \frac{2}{N + d/\theta^2}(\boldsymbol{G}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}} + \boldsymbol{G}_{\mathbb{U}}\boldsymbol{y}_{\mathbb{U}})\right). \tag{B.2}$$

Note that $\boldsymbol{A}_{\mathbb{UL}}, \boldsymbol{y}_{\mathbb{L}}$ and $\boldsymbol{G}_{\mathbb{UL}}$ are accessible for us in semi-supervised setting. Below, Lemma B.2 indicates that a scaled version of (B.2) is entrywisely close to $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ in (3.4) up to a global sign flip.

Lemma B.2. *Denote $\bar{\boldsymbol{u}} \coloneqq \boldsymbol{y}_{\mathbb{U}}/\sqrt{m}$. For each $i \in \mathbb{U}$, define*

$$\boldsymbol{w} \coloneqq \log(a/b)\Big(\boldsymbol{A}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}}/\sqrt{m} + \boldsymbol{A}_{\mathbb{U}}\bar{\boldsymbol{u}}\Big) + \frac{2}{N + d/\theta^2}(\boldsymbol{G}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}}/\sqrt{m} + \boldsymbol{G}_{\mathbb{U}}\bar{\boldsymbol{u}}).$$

Then for $\widehat{\boldsymbol{y}}_{\mathrm{PCA}}$ in (3.4), there exists some sequence $\{\varepsilon_m\}_m$ going to 0 such that

$$\mathbb{P}(\min_{c=\pm 1}\|c\widehat{\boldsymbol{y}}_{\mathrm{PCA}} - \boldsymbol{w}\|_\infty \geq \varepsilon_m\, m^{-1/2}\log(m)) \lesssim m^{-2}.$$

Proof of Lemma B.2. Define the following intermediate-term

$$\boldsymbol{v} = \log\Big(\frac{\alpha}{\beta}\Big)\cdot\left(\boldsymbol{A}_{\mathbb{UL}}\frac{1}{\sqrt{m}}\boldsymbol{y}_{\mathbb{L}} + \frac{m(\alpha-\beta)}{2}\cdot\boldsymbol{u}_2(\boldsymbol{A}_{\mathbb{U}})\right) + \frac{2\theta^2}{N\theta^2 + d}\left(\boldsymbol{G}_{\mathbb{UL}}\frac{1}{\sqrt{m}}\boldsymbol{y}_{\mathbb{L}} + m\theta^2\cdot\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}})\right).$$

By definition of α and β in Assumption 3.1, $\alpha/\beta = a/b$. We focus on the case $a > b$, then

$$\|\boldsymbol{v} - \boldsymbol{w}\|_\infty \leq \log(a/b)\cdot\|\boldsymbol{A}_{\mathbb{U}}\bar{\boldsymbol{u}} - \frac{(a-b)q_m}{2}\boldsymbol{u}_2(\boldsymbol{A}_{\mathbb{U}})\|_\infty + \frac{2\theta^2}{N\theta^2 + d}\|\boldsymbol{G}_{\mathbb{U}}\bar{\boldsymbol{u}} - m\theta^2\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}})\|_\infty$$
$$\begin{aligned}\|\boldsymbol{v} - \widehat{\boldsymbol{y}}_{\mathrm{PCA}}\|_\infty \leq{}& \left|\log\Big(\frac{\lambda_1(\boldsymbol{A}) + \lambda_2(\boldsymbol{A})}{\lambda_1(\boldsymbol{A}) - \lambda_2(\boldsymbol{A})}\Big) - \log(\alpha/\beta)\right|\cdot\frac{1}{\sqrt{m}}\|\boldsymbol{A}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}}\|_\infty\\ &+ \left|\log\Big(\frac{\lambda_1(\boldsymbol{A}) + \lambda_2(\boldsymbol{A})}{\lambda_1(\boldsymbol{A}) - \lambda_2(\boldsymbol{A})}\Big)\lambda_2(\boldsymbol{A}_{\mathbb{U}}) - \log(a/b)\frac{(a-b)q_m}{2}\right|\cdot\|\boldsymbol{u}_2(\boldsymbol{A}_{\mathbb{U}})\|_\infty\\ &+ \frac{1}{\sqrt{m}}\left|\frac{2\lambda_1(\boldsymbol{G})}{N\lambda_1(\boldsymbol{G}) + dN} - \frac{2\theta^2}{N\theta^2 + d}\right|\cdot\|\boldsymbol{G}_{\mathbb{UL}}\boldsymbol{y}_{\mathbb{L}}\|_\infty,\\ &+ \left|\frac{2\lambda_1(\boldsymbol{G})\lambda_1(\boldsymbol{G}_{\mathbb{U}})}{N\lambda_1(\boldsymbol{G}) + dN} - \frac{2m\theta^4}{N\theta^2 + d}\right|\cdot\|\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}})\|_\infty.\end{aligned}$$

Without loss of generality, we assume $\langle\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}}, \bar{\boldsymbol{u}})\rangle \geq 0$ and $\langle\boldsymbol{u}_2(\boldsymbol{A}_{\mathbb{U}}, \bar{\boldsymbol{u}})\rangle \geq 0$. Also, by Lemma B.1, Theorem 2.1 in (Abbe et al., 2022), with probability at least $1 - e^{-n}$,

$$\lambda_1(\boldsymbol{G}) = (1 + o(1))\cdot N\theta^2, \quad \lambda_1(\boldsymbol{G}_{\mathbb{U}}) = (1 + o(1))\cdot m\theta^2,$$

and for some large constant $c > 4$, with probability at least $1 - n^{-c}$, there exists some vanishing sequence $\{\varepsilon_m\}_m$ such that

$$\|\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}}) - \boldsymbol{G}_{\mathbb{U}}\bar{\boldsymbol{u}}/(m\theta^2)\|_\infty \lesssim \varepsilon_m\cdot m^{-1/2}, \quad \|\boldsymbol{u}_1(\boldsymbol{G}_{\mathbb{U}})\|_\infty \lesssim m^{-1/2}.$$

One can also obtain the upper bounds for $\|\boldsymbol{A}_{\mathbb{UL}}\|_{2\to\infty}$ and $\|\boldsymbol{G}_{\mathbb{UL}}\|_{2\to\infty}$. The remaining procedure follows similarly as Lemma F.1 in (Abbe et al., 2022) and Corollary 3.1 in (Abbe et al., 2020). □

Proof of Theorem 3.4 (1). First, for each node $i \in \mathcal{V}_{\mathbb{U}}$, if there exists some positive constant ξ such that $q_m^{-1}\sqrt{m} y_i(\widehat{\boldsymbol{y}}_{\text{PCA}})_i \geq \xi$, then the estimator $\text{sign}(\widehat{\boldsymbol{y}}_{\text{PCA}})$ recovers the label of each node correctly. Thus a sufficient condition for exact recovery is

$$q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i(\widehat{\boldsymbol{y}}_{\text{PCA}})_i \geq \xi, \quad \text{for some positive constant } \xi.$$

Remind the result of Lemma B.2, for some vanishing positive sequence $\{\varepsilon_m\}_m$, we have $\min_{c=\pm 1}\|c\widehat{\boldsymbol{y}}_{\text{PCA}} - \boldsymbol{w}\|_\infty \geq \varepsilon_m\, m^{-1/2} q_m$ with probability at most m^{-2}. Denote $\hat{c} := \arg\min_{c=\pm 1}\|c\widehat{\boldsymbol{y}}_{\text{PCA}} - \boldsymbol{w}\|_\infty$ and $\hat{\boldsymbol{v}} = \hat{c}\cdot\widehat{\boldsymbol{y}}_{\text{PCA}}$. Based on the facts above, the sufficient condition for exact recovery can be further simplified as

$$q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i\hat{v}_i \geq q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i w_i - \varepsilon_m \geq \xi,$$

where the last inequality holds since ε_m vanishes to 0. Then we have

$$\begin{aligned}
\mathbb{P}(\psi_m = 0) =& \mathbb{P}(\text{sign}(\widehat{\boldsymbol{y}}_{\text{PCA}}) = \boldsymbol{y}) \geq \mathbb{P}(q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i\cdot(\widehat{\boldsymbol{y}}_{\text{PCA}})_i \geq \xi)\\
\geq& \mathbb{P}(q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i\hat{v}_i \geq \xi,\ q_m^{-1}\sqrt{m}\|\hat{\boldsymbol{v}} - \boldsymbol{w}\|_\infty < \varepsilon_m)\\
\geq& \mathbb{P}(q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i w_i \geq \xi,\ q_m^{-1}\sqrt{m}\|\hat{\boldsymbol{v}} - \boldsymbol{w}\|_\infty < \varepsilon_m)\\
\geq& \mathbb{P}(q_m^{-1}\sqrt{m}\min_{i\in\mathbb{U}} y_i w_i \geq \xi) - \mathbb{P}(q_m^{-1}\sqrt{m}\|\hat{\boldsymbol{v}} - \boldsymbol{w}\|_\infty \geq \varepsilon_m)\\
\geq& 1 - \sum_{i\in\mathbb{U}}\mathbb{P}(q_m^{-1}\sqrt{m}y_i w_i < \xi) - m^{-2} = 1 - m\cdot\mathbb{P}(q_m^{-1}\sqrt{m}y_i w_i < \xi) - m^{-2}.
\end{aligned}$$

Note that $\sqrt{m}w_i y_i = W_{n,i}([N])$ defined in Lemma A.1. We take $0 < \varepsilon < \frac{a-b}{2(1-\tau)}\log(a/b) + 2c_\tau$, then for any $\delta > 0$, there exists some large enough positive constant M such that for $m \geq M$, $\varepsilon_m < \xi$, it follows that

$$\mathbb{P}(\sqrt{m}w_i y_i \leq \xi q_m) \leq \exp\Big(-\Big(\sup_{t\in\mathbb{R}}\Big\{\xi t + I(t, a_\tau, b_\tau, c_\tau))\Big\} + \delta\Big)\cdot\log(m)\Big).$$

By combining the arguments above, the probability of accomplishing exact recovery is lower bounded by

$$\mathbb{P}(\psi_m = 0) \geq 1 - m^{1-\sup_{t\in\mathbb{R}}\{\xi t + I(t,a_\tau,b_\tau,c_\tau))\}+\delta} - m^{-2} \overset{m\to\infty}{\longrightarrow} 1,$$

since $I(a_\tau, b_\tau, c_\tau) = \sup_{t\in\mathbb{R}}\{\xi t + I(t, a_\tau, b_\tau, c_\tau))\} > 1$ by assumption when choosing small enough ξ and δ. □

Proof of Theorem 3.4 (2). The proof procedure follows similarly to Theorem 4.2 in (Abbe et al., 2022), where we should apply the large deviation results Lemma A.1 and Lemma B.2 instead. □

Proof of Theorem 3.5. The proof procedure follows similarly to Theorem 4.4 in (Abbe et al., 2022), where we should apply the new large deviation result Lemma A.1 instead. The proof is simplified since $\boldsymbol{y}_{\mathbb{L}}$ is accessible under the semi-supervised learning regime. □

C. The analysis of the ridge regression on linear graph convolution

For CSBM$(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$, in this section, we focus on analyzing how these parameters c_τ, a_τ and b_τ defined in Assumption 3.1 affect the learning performances of the *linear* graph convolutional networks. We consider a graph convolutional kernel $h(\boldsymbol{X}) \in \mathbb{R}^{N\times d}$ which is a function of data matrix $\boldsymbol{X}$ and adjacency matrix $\boldsymbol{A}$ sampled from CSBM$(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$. We add self-loop and define the new adjacency matrix $\boldsymbol{A}_\rho := \boldsymbol{A} + \rho\boldsymbol{I}_N$, where $\rho \in \mathbb{R}$ represents the intensity. Let $\boldsymbol{D}_\rho$ be the diagonal matrix whose diagonals are the average degree for $\boldsymbol{A}_\rho$, i.e., $[\boldsymbol{D}_\rho]_{ii} = \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N(\boldsymbol{A}_\rho)_{ij}$ for each $i \in [N]$. Denote $\boldsymbol{D} := \boldsymbol{D}_0$, indicating no self-loop added. Recall that the normalization we applied for the linear graph convolutional layer is

$$h(\boldsymbol{X}) = \frac{1}{\sqrt{Nq_m}}\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho\boldsymbol{X}.$$

Let us denote the expectation of the average degree of $\boldsymbol{A}_\rho$ by

$$\widetilde{d} \coloneqq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{D}_\rho]_{ii} = \frac{a_\tau + b_\tau}{2} q_m + \rho. \tag{C.1}$$

However, $h(\boldsymbol{X})$ is hard to deal with when we consider its large deviation principle. Instead, we use the following $\widetilde{h}(\boldsymbol{X})$ for analysis

$$\widetilde{h}(\boldsymbol{X}) = \frac{1}{\widetilde{d}\cdot\sqrt{Nq_m}}\boldsymbol{A}_\rho \boldsymbol{X}.$$

C.1. Large Deviation Results for Ridge Regression Estimators

For any $i \in \mathcal{V}$, we denote $\mathcal{N}_i \subset \mathcal{V}$ as the neighborhood of vertex $i \in \mathcal{V}$. We consider the case that the feature learned by the GCN is $\zeta\sqrt{q_m}\,\boldsymbol{\mu}$ for some constant ζ, i.e., $\boldsymbol{w} = \zeta\sqrt{q_m}\,\boldsymbol{\mu}$. Notice that

$$\begin{aligned} h_i =& y_i\zeta\sqrt{q_m}(\mathbb{E}\boldsymbol{D}_\rho)^{-1}(\boldsymbol{A}_\rho\boldsymbol{X})_{i:}\boldsymbol{\mu} \\ =& \frac{y_i\zeta\sqrt{q_m}}{\widetilde{d}}\Big(\sum_{j\in\mathcal{N}_i}\boldsymbol{\mu}^\top(\theta y_j\boldsymbol{\mu} + \boldsymbol{z}_j) + \rho\boldsymbol{\mu}^\top(\theta y_i\boldsymbol{\mu} + \boldsymbol{z}_i)\Big) \\ =& \underbrace{\frac{\rho\zeta\theta\sqrt{q_m}y_i^2\|\boldsymbol{\mu}\|_2^2}{\widetilde{d}}}_{I_1} + \underbrace{\frac{\rho\zeta\sqrt{q_m}y_i\boldsymbol{\mu}^\top\boldsymbol{z}_i}{\widetilde{d}}}_{I_2} + \underbrace{\frac{\zeta\theta\sqrt{q_m}\|\boldsymbol{\mu}\|_2^2}{\widetilde{d}}\sum_{j\neq i}\boldsymbol{A}_{ij}y_iy_j}_{I_3} + \underbrace{\frac{\zeta\sqrt{q_m}}{\widetilde{d}}\sum_{j\neq i}y_i\boldsymbol{A}_{ij}\boldsymbol{\mu}^\top\boldsymbol{z}_j}_{I_4}. \end{aligned} \tag{C.2}$$

Here $\mathcal{N}_i$ is the neighborhood of vertex $i \in \mathcal{V}_{\mathbb{U}}$.

Proposition C.1 (LDP for Ridge Regression)**.** *Under the Assumption 3.1 with* $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ *and* $\rho = sq_m$ *for some constant* $s \in \mathbb{R}$. *Assume* $d/N \ll q_m$, *then for any fixed* $i \in \mathcal{V}_{\mathbb{U}}$ *and constant* $\zeta > 0$, *we have*

$$\lim_{m\to\infty} q_m^{-1}\log\mathbb{P}(y_i\zeta(\mathbb{E}\boldsymbol{D}_\rho)^{-1}(\boldsymbol{A}_\rho\boldsymbol{X})_{i:}\boldsymbol{\mu} \le \varepsilon\sqrt{q_m}) = -\sup_{t\in\mathbb{R}}\{\varepsilon t + g(a,b,c,\tau,\zeta,s,t)\}$$

for sufficiently small $\varepsilon > 0$ *and all large* m*, where*

$$\begin{aligned} g(a_\tau, b_\tau, c_\tau, \zeta, s, t) =& g_1(t) + g_2(t), \\ g_1(t) =& -\frac{2ts\zeta\sqrt{c_\tau}}{a_\tau + b_\tau + 2s} - \frac{2t^2s^2\zeta^2}{(a_\tau + b_\tau + 2s)^2}, \\ g_2(t) =& -\frac{a_\tau}{2}\Big[\exp\Big(\frac{2t\zeta\sqrt{c_\tau}}{a_\tau + b_\tau + 2s}\Big) - 1\Big] - \frac{b_\tau}{2}\Big[\exp\Big(-\frac{2t\zeta\sqrt{c_\tau}}{a_\tau + b_\tau + 2s}\Big) - 1\Big]. \end{aligned}$$

Consequently, for any sufficiently small $\varepsilon > 0$ *and any* $\delta > 0$*, there exists some* $N_0 > 0$ *such that for all* $N \ge N_0$*, we have*

$$\mathbb{P}(y_i\zeta(\mathbb{E}\boldsymbol{D}_\rho)^{-1}(\boldsymbol{A}_\rho\boldsymbol{X})_{i:}\boldsymbol{\mu} \le \varepsilon\sqrt{q_m}) = \exp\left(-q_m[\sup_{t\in\mathbb{R}}\{\varepsilon t + g(a_\tau, b_\tau, c_\tau, \zeta, s, t)\} - \delta]\right)$$

Proof of Proposition C.1. Our goal is to calculate the following moment-generating function

$$\mathbb{E}[\exp(th_i)] \coloneqq \mathbb{E}_{\boldsymbol{A}}[\mathbb{E}_{\boldsymbol{X}}[\exp(th_i)|\boldsymbol{A}]].$$

First, since $\|\boldsymbol{\mu}\|_2 = 1$, $y_i^2 = 1$, then in (C.2), $I_1 = \rho\zeta\theta\sqrt{q_m}/\widetilde{d}$, and it is deterministic. Second, $\boldsymbol{\mu}^\top\boldsymbol{z}_i \sim \mathcal{N}(0,1)$, then $I_2 \sim \mathcal{N}(0, \rho^2\zeta^2 q_m/\widetilde{d}^2)$, and

$$\mathbb{E}_{\boldsymbol{X}}[\exp(tI_2)|y_i] = \exp\Big(\frac{t^2\rho^2\zeta^2q_m}{2\widetilde{d}^2}\Big) = \mathbb{E}[\exp(tI_2)],$$

where the last equality holds since the result we obtained is independent of $y_i, \boldsymbol{A}$. Let $\mathcal{N}_i$ denote the set of neighbors of node i and $|\mathcal{N}_i|$ denote its cardinality.

Conditioned on $\boldsymbol{A}, \boldsymbol{y}, \boldsymbol{\mu}$, then $I_4 \sim \mathcal{N}(0, |\mathcal{N}_i|\zeta^2 q_m/\widetilde{d}^2)$, and

$$\mathbb{E}_{\boldsymbol{X}}[\exp(tI_4)|\boldsymbol{A}, y_i, \boldsymbol{\mu}] = \exp\Big(\frac{t^2\zeta^2|\mathcal{N}_i|q_m}{2\widetilde{d}^2}\Big).$$

Note that $|\mathcal{N}_i| = \sum_{j=1}^N A_{ij}$, and I_3 is independent of $\boldsymbol{X}$, then

$$\mathbb{E}_{\boldsymbol{X}}[\exp\big(t(I_3+I_4)\big)|\boldsymbol{A}, y_i, \boldsymbol{\mu}] = \exp\Bigg(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\sum_{j\neq i} A_{ij}\Big(y_iy_j + \frac{t\zeta\sqrt{q_m}}{2\widetilde{d}\theta}\Big)\Bigg)$$

One could take the expectation over $\boldsymbol{A}$ conditioned on $\boldsymbol{y}$, then

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{A}}\Big[\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}A_{ij}\Big(y_iy_j + \frac{t\zeta\sqrt{q_m}}{2\widetilde{d}\theta}\Big)\Big)\Big|y_i\Big]\\
=&\frac{1}{2}\mathbb{E}_{\boldsymbol{A}}\Big[\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}A_{ij}\Big(1 + \frac{t\zeta\sqrt{q_m}}{2\widetilde{d}\theta}\Big)\Big)\Big|y_iy_j=1\Big] + \frac{1}{2}\mathbb{E}_{\boldsymbol{A}}\Big[\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}A_{ij}\Big(-1 + \frac{t\zeta\sqrt{q_m}}{2\widetilde{d}\theta}\Big)\Big)\Big|y_iy_j=-1\Big]\\
=&\frac{\alpha}{2}\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}} + \frac{t^2\zeta^2q_m}{2\widetilde{d}^2}\Big) + \frac{1-\alpha}{2} + \frac{\beta}{2}\exp\Big(-\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}} + \frac{t^2\zeta^2q_m}{2\widetilde{d}^2}\Big) + \frac{1-\beta}{2}\\
=&1 + \frac{\alpha}{2}\Big(\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}} + \frac{t^2\zeta^2q_m}{2\widetilde{d}^2}\Big) - 1\Big) + \frac{\beta}{2}\Big(\exp\Big(-\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}} + \frac{t^2\zeta^2q_m}{2\widetilde{d}^2}\Big) - 1\Big)\\
=&1 + \frac{\alpha}{2}\Big(\exp\Big((1+o(1))\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big) + \frac{\beta}{2}\Big(\exp\Big(-(1+o(1))\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big),
\end{aligned}$$

where the last equality holds since $\widetilde{d} \asymp q_m$, $\theta \asymp \sqrt{q_m}$, $\zeta \asymp 1$, and for some fix t, $1 \asymp |\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}| \gg \frac{t^2\zeta^2q_m}{2\widetilde{d}^2} \asymp q_m^{-1} = o(1)$ for sufficiently large m. At the same time, the result above is again independent of $y_i, \boldsymbol{\mu}$. Recall $\alpha = aq_m/m = o(1)$, $\beta = bq_m/m = o(1)$, $\frac{\theta^4}{\theta^2 + d/N} = c_\tau q_m$ in Assumption 3.1. By using $\log(1+x) = x$ for $x = o(1)$, we then have

$$\begin{aligned}
q_m^{-1}\log\mathbb{E}_{\boldsymbol{A}}\big[\mathbb{E}_{\boldsymbol{X}}[\exp\{t(I_3+I_4)\}]\big] &= \frac{1}{q_m}\sum_{j\in[N]\setminus\{i\}}\log\Big(1 + \frac{\alpha}{2}\Big[\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big] + \frac{\beta}{2}\Big[\exp\Big(-\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big]\Big)\\
&= \frac{N-1}{q_m}\cdot\Big(\frac{aq_m}{2m}\Big[\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big] + \frac{bq_m}{2m}\Big[\exp\Big(-\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big]\Big)\\
&= \frac{a}{2(1-\tau)}\Big[\exp\Big(\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big] + \frac{b}{2(1-\tau)}\Big[\exp\Big(-\frac{t\zeta\theta\sqrt{q_m}}{\widetilde{d}}\Big) - 1\Big]\\
&= \frac{a}{2(1-\tau)}\Big[\exp\Big(\frac{2t\zeta c_\tau}{a+b+2s}\Big) - 1\Big] + \frac{b}{2(1-\tau)}\Big[\exp\Big(-\frac{2t\zeta c_\tau}{a+b+2s}\Big) - 1\Big],
\end{aligned}$$

where the last equality holds since we apply $\widetilde{d} = (\frac{a+b}{2} + s)q_m$ and $\theta^2 = (1+o(1))c_\tau q_m$ since $d/N \ll q_m$ by assumption. Combining the calculations above, we then compute the following rate function

$$\begin{aligned}
g(a_\tau, b_\tau, c_\tau, \zeta, s, t) =& -q_m^{-1}\log\mathbb{E}[\exp(th_i)] = -q_m^{-1}\log\mathbb{E}_{\boldsymbol{A}}\big[\mathbb{E}_{\boldsymbol{X}}[\exp(t(I_1+I_2+I_3+I_4))|\boldsymbol{A}]\big]\\
=& -(g_1(t) + g_2(t)).
\end{aligned}$$

Then, we can apply Lemma H.5 in (Abbe et al., 2022) to conclude our results in this proposition. □

Lemma C.2. *For function $g(a_\tau, b_\tau, c_\tau, \zeta, s, t)$ defined in Proposition C.1, we know that*

$$J(a_\tau, b_\tau, c_\tau, \zeta, s) := \sup_{t\in\mathbb{R}} g(a_\tau, b_\tau, c_\tau, \zeta, s, t) \leq I(a_\tau, b_\tau, c_\tau)$$

and the equality is attained when $s = \frac{2c_\tau}{\log(a_\tau/b_\tau)} = \frac{2c_\tau}{\log(a/b)}$.

Moreover, if $s = 0$, then $g_1(t) \equiv 0$ and

$$J(a_\tau, b_\tau, c_\tau, \zeta, 0) = I(a_\tau, b_\tau, 0) \leq J(a_\tau, b_\tau, c_\tau, \zeta, s).$$

Proof of Lemma C.2. Notice that both $g_1(t)$ and $g_2(t)$ are concave. First, $g_1(t)$ achieves its maximum at the point $t_1 := -\sqrt{c_\tau}(a+b+2s)/(2s\zeta)$, and $g_2(t)$ achieves its maximum at the point $t_2 := (a+b+2s)\log(b/a)/(4\zeta\sqrt{c_\tau})$. Note that

$$\sup_{t\in\mathbb{R}} g(a_\tau, b_\tau, c_\tau, \zeta, s, t) \le \max_{t\in\mathbb{R}} g_1(t) + \max_{t\in\mathbb{R}} g_2(t) = g_1(t_1) + g_2(t_2), \tag{C.3}$$

where

$$g_1(t_1) = c_\tau/2, \quad \text{and } g_2(t_2) = (\sqrt{a} - \sqrt{b})^2/(2(1-\tau)).$$

Thus, this proves the upper bound on $J(a_\tau, b_\tau, c_\tau, \zeta, s)$.

Notice that the equality in (C.3) holds if $t_1 = t_2$. It turns out that when $s = \frac{2c_\tau}{\log(a/b)}$, then $t_1 = t_2$, and $g_1(t_1) = c_\tau/2$, $g_2(t_2) = (\sqrt{a} - \sqrt{b})^2/(2(1-\tau))$. Therefore, in this case, we have

$$\max_{t\in\mathbb{R}} g(a, b, c, \tau, \zeta, t) = \frac{(1-\tau)^{-1}(\sqrt{a} - \sqrt{b})^2 + c_\tau}{2} = I(a_\tau, b_\tau, c_\tau).$$

□

C.2. Preliminary Lemmas on Ridge Regression Estimator

Note the facts that $(\boldsymbol{B}^\top\boldsymbol{B} + \boldsymbol{I}_d)^{-1}\boldsymbol{B}^\top = \boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{I}_N)^{-1}$ for any matrix $\boldsymbol{B}^{N\times d}$, $\boldsymbol{P}_{\mathbb{U}}^2 = \boldsymbol{P}_{\mathbb{U}}$, $\boldsymbol{P}_{\mathbb{L}}^2 = \boldsymbol{P}_{\mathbb{L}}$ and $\boldsymbol{P}_{\mathbb{U}} = \boldsymbol{I}_N - \boldsymbol{P}_{\mathbb{L}}$, then,

$$\begin{aligned} h(\boldsymbol{X})\widehat{\boldsymbol{\beta}} &= h(\boldsymbol{X})(h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} \boldsymbol{y} \\ &= h(\boldsymbol{X})[(\boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}))^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d]^{-1} (\boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}))^\top \boldsymbol{y} \\ &= h(\boldsymbol{X}) h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} [\boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} + \lambda \boldsymbol{I}_N]^{-1} \boldsymbol{y}. \end{aligned}$$

Consequently, the test risk can be written as

$$\mathcal{R}(\lambda) = \frac{1}{m} \boldsymbol{y}^\top \boldsymbol{Q}^\top \boldsymbol{P}_{\mathbb{U}} \boldsymbol{Q} \boldsymbol{y}, \ \text{ where } \boldsymbol{Q} := h(\boldsymbol{X}) h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} [\boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} + \lambda \boldsymbol{I}_N]^{-1} - \boldsymbol{I}_N$$

Lemma C.3. *Assume that $|\rho| \lesssim q_m$ and $q_m \gtrsim \log N$. Let $\boldsymbol{D}_\rho$ be the diagonal matrix where each diagonal represents the average degree of the graph $\boldsymbol{A}_\rho$ after adding the self-loop ρ and let $\widetilde{d}$ denote the expected average degree of $\boldsymbol{A}_\rho$. Then $\|\boldsymbol{D}_\rho^{-1} - (\widetilde{d})^{-1}\| \lesssim q_m^{-3/2}$ with probability at least $1 - e^{-N}$. Furthermore, with probability at least $1 - 2N^{-10} - 2e^{-N}$,*

$$\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho - \mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\| \lesssim q_m^{-1/2}.$$

Consequently, $\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho\| \le C$ with probability at least $1 - 2N^{-10} - 2e^{-N}$ for some constant $C > 0$.

Proof of Lemma C.3. First, for any $i \in [N]$, note that $[\boldsymbol{D}_\rho]_{ii} = \frac{1}{N}\sum_{i=1}^N(\rho + \sum_{j\neq i}\boldsymbol{A}_{ij}) = \rho + \frac{1}{N}\sum_{i=1}^N\sum_{j\neq i}\boldsymbol{A}_{ij}$, and $\widetilde{d} = \mathbb{E}[\boldsymbol{D}_\rho]_{ii} = \frac{a_\tau + b_\tau}{2}q_m + \rho$, then by Bernstein inequality,

$$\mathbb{P}\Big(\big|[\boldsymbol{D}_\rho]_{ii} - \widetilde{d}\big| \ge \sqrt{q_m}\Big) = \mathbb{P}\Big(\Big|\sum_{i=1}^N\sum_{j\neq i}(\boldsymbol{A}_{ij} - \mathbb{E}\boldsymbol{A}_{ij})\Big| \ge N\sqrt{q_m}\Big) \lesssim \exp(-N).$$

Thus by comparing the entrywise difference of $[\boldsymbol{D}_\rho]_{ii} - \widetilde{d}$, with probability at least $1 - e^{-N}$, we have

$$\|\boldsymbol{D}_\rho^{-1} - (\widetilde{d})^{-1}\| \lesssim q_m^{-3/2}. \tag{C.4}$$

For the second part of the statement, by the triangle inequality, we have

$$\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho - \mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\| \le \|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho - \boldsymbol{D}_\rho^{-1}\mathbb{E}\boldsymbol{A}_\rho\| + \|\boldsymbol{D}_\rho^{-1}\mathbb{E}\boldsymbol{A}_\rho - \mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\|$$

For the first term, we proved that with probability at least $1-e^{-N}$, $[D_\rho]_{ii}=(1+o(1))\widetilde{d}\asymp q_m$ with deviation at most $\sqrt{q_m}$, then according to Lemma E.2, with probability at least $1-2N^{-10}-2e^{-N}$, when $q_m\gtrsim \log(N)$, the following is bounded by

$$\|\boldsymbol{D}_\rho^{-1}(\boldsymbol{A}_\rho-\mathbb{E}\boldsymbol{A}_\rho)\|\leq(\|\widetilde{d}^{-1}\|+\|\boldsymbol{D}_\rho^{-1}-\widetilde{d}^{-1}\|)\cdot\|(\boldsymbol{A}_\rho-\mathbb{E}\boldsymbol{A}_\rho)\|\leq(q_m^{-1}+q_m^{-3/2})\sqrt{q_m}\asymp q_m^{-1/2}.$$

For the second term, note that $\|\mathbb{E}\boldsymbol{A}_\rho\|\lesssim q_m$, then by results above

$$\|\left(\boldsymbol{D}_\rho^{-1}-\widetilde{d}^{-1}\right)\mathbb{E}\boldsymbol{A}_\rho\|\leq\|\boldsymbol{D}_\rho^{-1}-(\widetilde{d})^{-1}\|\cdot\|\mathbb{E}\boldsymbol{A}_\rho\|\lesssim\frac{1}{\sqrt{q_m}}.$$

Therefore, with probability at least $1-2N^{-10}-2e^{-N}$,

$$\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho-\mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\|\lesssim q_m^{-1/2}.$$

For the last part, $\|\mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\|\lesssim 1$ since

$$\mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}=\frac{1}{N}\boldsymbol{1}\boldsymbol{1}^\top+\frac{(a-b)q_m+2\rho}{(a+b)q_m+2\rho}\frac{1}{N}\boldsymbol{y}\boldsymbol{y}^\top.$$

Then the proof is completed by triangle inequality since $q_m\gg 1$, and

$$\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho\|\leq\|\mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\|+\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho-\mathbb{E}\boldsymbol{A}_\rho/\widetilde{d}\|\lesssim 1+q_m^{-1/2}\lesssim 1.$$

□

Lemma C.4. *Consider $\boldsymbol{X}\sim\mathrm{GMM}(\boldsymbol{\mu},\boldsymbol{y},\theta)$. Suppose that $d\lesssim N$, then we have*

$$\left\|\frac{1}{\sqrt{Nq_m}}\boldsymbol{X}-\frac{\theta}{\sqrt{Nq_m}}\boldsymbol{y}\boldsymbol{\mu}^\top\right\|\leq\frac{C}{\sqrt{q_m}},$$

with probability at least $1-2e^{-cN}$ for some constants $C,c>0$.

Proof of Lemma C.4. Recall the concentration on the operator norm of the Gaussian random matrix for $\boldsymbol{Z}\in\mathbb{R}^{N\times d}$ (see (Vershynin, 2018)). Then for every $t>0$, there exists some constant $c>0$ such that

$$\mathbb{P}(\|\boldsymbol{Z}\|\geq\sqrt{N}+\sqrt{d}+t)\leq 2\exp\left(-ct^2\right). \tag{C.5}$$

Then, we know that $\frac{\|\boldsymbol{Z}\|}{\sqrt{N}}\lesssim 1$ with probability at least $1-2e^{-cN}$ by taking $t=\sqrt{N}$. Then we have

$$\left\|\frac{1}{\sqrt{Nq_m}}\boldsymbol{X}-\frac{\theta}{\sqrt{Nq_m}}\boldsymbol{y}\boldsymbol{\mu}^\top\right\|\leq\frac{\|\boldsymbol{Z}\|}{\sqrt{Nq_m}}\lesssim\frac{\sqrt{N}+\sqrt{d}}{\sqrt{Nq_m}}\asymp\frac{1}{\sqrt{q_m}},$$

with probability at least $1-2e^{-cN}$. □

Lemma C.5. *Consider $(\boldsymbol{A},\boldsymbol{X})\sim\mathrm{CSBM}(\boldsymbol{y},\boldsymbol{\mu},\alpha,\beta,\theta)$. Under the Assumption 3.1, when $d\lesssim N$, we have that*

$$\|h(\boldsymbol{X})-\boldsymbol{H}\|\leq\frac{C}{\sqrt{q_m}},$$

with probability at least $1-cN^{-10}$, where $\boldsymbol{H}:=\frac{\kappa_m}{\sqrt{N}}\boldsymbol{y}\boldsymbol{\mu}^\top$ and $\kappa_m:=\frac{\alpha-\beta+2\rho}{\alpha+\beta+2\rho}\cdot\frac{\theta}{\sqrt{q_m}}$, for all large m and n and some constants $c,C>0$.

Proof of Lemma C.5. Notice that $\boldsymbol{H}=\frac{\theta}{\widetilde{d}\sqrt{Nq_m}}(\mathbb{E}\boldsymbol{A}_\rho)\boldsymbol{y}\boldsymbol{\mu}^\top$. We apply Lemmas C.3 and C.4 to derive that

$$\begin{aligned}\|h(\boldsymbol{X})-\boldsymbol{H}\|\leq&\left\|h(\boldsymbol{X})-\frac{1}{\widetilde{d}}(\mathbb{E}\boldsymbol{A}_\rho)\frac{\boldsymbol{X}}{\sqrt{Nq_m}}\right\|+\left\|\frac{1}{\widetilde{d}}(\mathbb{E}\boldsymbol{A}_\rho)\frac{\boldsymbol{X}}{\sqrt{Nq_m}}-\boldsymbol{H}\right\|\\ \leq&\frac{1}{\sqrt{q_m}}(\theta+2+\sqrt{N/d})\cdot\left\|\boldsymbol{D}_\rho^{-1}\boldsymbol{A}_\rho-\frac{1}{\widetilde{d}}(\mathbb{E}\boldsymbol{A}_\rho)\right\|+\frac{\|\mathbb{E}\boldsymbol{A}_\rho\|}{\widetilde{d}}\left\|\frac{\boldsymbol{X}}{\sqrt{Nq_m}}-\frac{\theta}{\sqrt{Nq_m}}\boldsymbol{y}\boldsymbol{\mu}^\top\right\|\lesssim 1/\sqrt{q_m},\end{aligned}$$

with probability at least $1-cN^{-10}$. Here we apply the fact that $\theta/\sqrt{q_m}\lesssim 1$. □

Lemma C.6. *Consider* $(\boldsymbol{A}, \boldsymbol{X}) \sim \mathrm{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$. *Under the Assumption 3.1 with* $d \lesssim N$, *the ridge regression solution* $\widehat{\boldsymbol{\beta}}$ *defined in (3.7) satisfies*

$$\frac{1}{\sqrt{N}}\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\| \leq \frac{C}{\sqrt{q_m}},$$

with probability at least $1 - cN^{-10}$, *where* $\widetilde{\boldsymbol{\beta}} := \frac{\sqrt{N}\kappa_m \tau}{\kappa_m^2 \tau + \lambda}\boldsymbol{\mu}$ *and* $\kappa_m = \frac{\alpha - \beta + 2\rho}{\alpha + \beta + 2\rho} \cdot \frac{\theta}{\sqrt{q_m}}$, *for all large* m *and* n *and some constants* $c, C > 0$. *Moreover,* $\|\widehat{\boldsymbol{\beta}}\| \lesssim \sqrt{N}$ *with probability at least* $1 - cN^{-10}$.

Proof. Notice that $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{H}^\top \boldsymbol{P}_{\mathbb{L}} \boldsymbol{H} + \lambda \boldsymbol{I}_d)^{-1}\boldsymbol{H}^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{y}$, where $\boldsymbol{H}$ is defined by Lemma C.5. From Lemma C.3, we know that $\|h(\boldsymbol{X})\| \lesssim 1$ and $\|\boldsymbol{H}\| \lesssim 1$ with probability at least $1 - 2N^{-10}$. Moreover, $\|(\boldsymbol{H}^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{H} + \lambda \boldsymbol{I}_d)^{-1}\| \leq \lambda^{-1}$ and $\|(h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1}\| \leq \lambda^{-1}$. Therefore, applying Lemma C.5, we derive that

$$\begin{aligned}
\frac{1}{\sqrt{N}}\|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\| \leq\ & \|(\boldsymbol{H}^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{H} + \lambda \boldsymbol{I}_d)^{-1}\boldsymbol{H}^\top - (h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h(\boldsymbol{X})^\top\| \cdot \|\boldsymbol{P}_{\mathbb{L}}\boldsymbol{y}\|/\sqrt{N} \\
\leq\ & \|(\boldsymbol{H}^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{H} + \lambda \boldsymbol{I}_d)^{-1}\|\|(\boldsymbol{H} - h(\boldsymbol{X}))\| + \|(\boldsymbol{H}^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{H} + \lambda \boldsymbol{I}_d)^{-1}\| \\
& \cdot \|\boldsymbol{H} - h(\boldsymbol{X})\| \cdot (\|\boldsymbol{H}\| + \|h(\boldsymbol{X})\|) \cdot \|(h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1}\| \cdot \|h(\boldsymbol{X})\| \\
\lesssim\ & \|\boldsymbol{H} - h(\boldsymbol{X})\| \lesssim 1/\sqrt{q_m},
\end{aligned}$$

with at least $1 - cN^{-10}$, for some constant $c > 0$. □

C.3. Exact Recovery Threshold for Ridge Regression on Linear GCN

Lemma C.7. *Let* $h(\boldsymbol{X})^\top = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N]$ *and* $\widetilde{h}(\boldsymbol{X})^\top = [\bar{\boldsymbol{h}}_1, \ldots, \bar{\boldsymbol{h}}_N]$. *For any* $i \in [N]$ *and deterministic unit vector* $\boldsymbol{u} \in \mathbb{R}^d$, *there exists some* $c, C > 0$ *such that*

$$\mathbb{P}(|(\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i)^\top \boldsymbol{u}| \leq C/\sqrt{Nq_m}) \geq\ 1 - cN^{-10}, \tag{C.6}$$

$$\mathbb{P}(|\bar{\boldsymbol{h}}_i^\top \boldsymbol{u}| \leq C/\sqrt{N}) \geq\ 1 - cN^{-10}, \tag{C.7}$$

$$\mathbb{P}\left(\|\boldsymbol{h}_i\| \leq C\sqrt{d/(Nq_m)}\right) \geq\ 1 - cN^{-10}, \tag{C.8}$$

for all large n *and* m.

Proof. For any unit vector $\boldsymbol{u} \in \mathbb{R}^d$ and $i \in [N]$, conditioning on event (C.4), we have

$$\left|(\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i)^\top \boldsymbol{u}\right| \leq\ \frac{1}{\sqrt{Nq_m}}\left\|(\widetilde{d})^{-1}\boldsymbol{I}_N - \boldsymbol{D}_\rho^{-1}\right\| \cdot |(\boldsymbol{A}_\rho \boldsymbol{X})_{i:}\boldsymbol{u}| \lesssim \frac{1}{q_m^2\sqrt{N}}\left|(\boldsymbol{A}_\rho \boldsymbol{X})_{i:}\boldsymbol{u}\right|,$$

where we employ Lemma C.3. Then, for any $i \in [N]$, we can further have

$$\begin{aligned}
|(\boldsymbol{A}_\rho \boldsymbol{X})_{i:}\boldsymbol{u}| =\ & \left|\sum_{j \in \mathcal{N}_i}(\theta y_j \boldsymbol{\mu}^\top \boldsymbol{u} + \boldsymbol{z}_j^\top \boldsymbol{u}) + \theta\rho\boldsymbol{\mu}^\top \boldsymbol{u} + \rho \boldsymbol{z}_i^\top \boldsymbol{u}\right| \\
\leq\ & \theta(|\mathcal{N}_i| + \rho) + \left|\sum_{j \in \mathcal{N}_i}\boldsymbol{z}_j^\top \boldsymbol{u} + \rho\boldsymbol{z}_i^\top \boldsymbol{u}\right|
\end{aligned} \tag{C.9}$$

Based on Lemma 3.3 in (Alt et al., 2021), we can upper bound the degree of each vertex by

$$\mathbb{P}(|\mathcal{N}_i| \leq C\log N) \geq 1 - C_D N^{-D}, \tag{C.10}$$

for any $i \in [N]$, some constants $C, C_D > 0$ and sufficiently large constant $D > 0$. Meanwhile, since each $\boldsymbol{z}_j^\top \boldsymbol{u} \sim \mathcal{N}(0, 1)$, by applying Hoeffding's inequality (Theorem 2.6.2 in (Vershynin, 2018)), we can deduce that

$$\mathbb{P}\Big(\Big|\sum_{j \in \mathcal{N}_i}\boldsymbol{z}_j^\top \boldsymbol{u} + \rho\boldsymbol{z}_i^\top \boldsymbol{u}\Big| \leq t\Big||\mathcal{N}_i| = k\Big) \geq 1 - 2\exp\Big(-\frac{ct^2}{k + \rho^2}\Big), \tag{C.11}$$

for any $k \in \mathbb{N}$, $t > 0$, and some constant $c > 0$. Then combining (C.10) and (C.11), for any large $D > 0$, there exists some constants $C, C_D > 0$ such that

$$\mathbb{P}\Big(\Big|\sum_{j \in \mathcal{N}_i} \boldsymbol{z}_j^\top \boldsymbol{u} + \rho \boldsymbol{z}_i^\top \boldsymbol{u}\Big| \le C \log N,\ |\mathcal{N}_i| \le C \log N\Big) \ge 1 - 2C_D N^{-D}. \tag{C.12}$$

Thus, with (C.9) and $\rho \asymp \log N$, we can conclude that $|(\boldsymbol{A}_\rho \boldsymbol{X})_{i:} \boldsymbol{u}| \lesssim q_m^{3/2}$ with probability at least $1 - 2C_D N^{-D}$. Following with (C.9), we can conclude that

$$\mathbb{P}\left(\left|(\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i)^\top \boldsymbol{u}\right| \le C/\sqrt{q_m N}\right) \ge 1 - cN^{-10}.$$

For the second part, we can analogously get $|\bar{\boldsymbol{h}}_i^\top \boldsymbol{u}| \lesssim \frac{1}{q_m^{3/2}\sqrt{N}} |(\boldsymbol{A}_\rho \boldsymbol{X})_{i:} \boldsymbol{u}|$. Then, we can apply (C.9) and (C.12) to conclude (C.7).

Finally, notice that

$$\begin{aligned}\|(\boldsymbol{A}_\rho \boldsymbol{X})_{i:}\| &= \left\|\sum_{j \in \mathcal{N}_i} (\theta y_j \boldsymbol{\mu} + \boldsymbol{z}_j) + \theta \rho \boldsymbol{\mu} + \rho \boldsymbol{z}_i\right\| \\ &\le\ \theta(|\mathcal{N}_i| + \rho) + \left\|\sum_{j \in \mathcal{N}_i} \boldsymbol{z}_j + \rho \boldsymbol{z}_i\right\|.\end{aligned}$$

Applying Theorem 3.1.1 in (Vershynin, 2018), we know that

$$\mathbb{P}(\|\boldsymbol{z}_i\| \le 2\sqrt{d}) \ge 1 - 2\exp(-cd)$$

for some constant $c > 0$ and any $i \in [N]$. Thus, combining (C.10) and Lemma E.1, we have that with probability at least $1 - cN^{-10}$,

$$\|\boldsymbol{h}_i\| \le \frac{1}{q_m^{3/2}\sqrt{N}} \|(\boldsymbol{A}_\rho \boldsymbol{X})_{i:}\| \lesssim \sqrt{\frac{d}{Nq_m}}$$

because of the fact that $q_m \lesssim d$ and $N \asymp m$. This completes the proof of this lemma. □

Inspired by (Abbe et al., 2020; 2022), we now apply a general version of leave-one-out analysis for $\widehat{\boldsymbol{\beta}}$ by defining the following approximated estimator. For any $i \in \mathcal{V}_{\mathbb{U}}$, denote by

$$\widehat{\boldsymbol{\beta}}^{(i)} = (h^{(i)}(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h^{(i)}(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h^{(i)}(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} \boldsymbol{y}, \tag{C.13}$$

where $h^{(i)}(\boldsymbol{X}) := \frac{1}{\sqrt{Nq_m}} \boldsymbol{D}_\rho^{-1} \boldsymbol{A}_\rho (\boldsymbol{X} - \boldsymbol{Z}^{(i)})$ and $\boldsymbol{Z}^{(i)} := [\boldsymbol{z}_1 \mathbf{1}_{1 \in \mathcal{N}_i \cup \{i\}}, \ldots, \boldsymbol{z}_k \mathbf{1}_{k \in \mathcal{N}_i \cup \{i\}}, \ldots, \boldsymbol{z}_N \mathbf{1}_{N \in \mathcal{N}_i \cup \{i\}}]^\top \in \mathbb{R}^{N \times d}$. Here, the difference between $h^{(i)}(\boldsymbol{X})$ and $h(\boldsymbol{X})$ is that we turn off the feature noises $\boldsymbol{z}_i$ for vertices $\mathcal{N}_i \cup \{i\}$. In this case, conditional on $\boldsymbol{y}$ and $\boldsymbol{\mu}$, both $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}^{(i)}$ are independent with $\boldsymbol{h}_i$ and $\bar{\boldsymbol{h}}_i$ given any $i \in \mathcal{V}_{\mathbb{U}}$. Next, we present the following properties for $\widehat{\boldsymbol{\beta}}^{(i)}$.

Lemma C.8. *Assume that $q_m \ll d \ll Nq_m$. For (3.7) and (C.13), we have*

$$\frac{1}{\sqrt{N}} \left\|\widehat{\boldsymbol{\beta}}^{(i)} - \widehat{\boldsymbol{\beta}}\right\| \le\ C\sqrt{\frac{d}{q_m N}},$$
$$\|\widehat{\boldsymbol{\beta}}^{(i)}\| \le\ C\sqrt{N},$$

with a probability at least $1 - cN^{-10}$, for some constants $c, C > 0$

Proof. Based on Lemma C.3, we have that

$$\|h^{(i)}(\boldsymbol{X}) - h(\boldsymbol{X})\| = \frac{1}{\sqrt{Nq_m}} \|\boldsymbol{D}_\rho^{-1} \boldsymbol{A}_\rho \boldsymbol{Z}^{(i)}\| \le \frac{1}{\sqrt{q_m N}} \|\boldsymbol{Z}^{(i)}\| \lesssim \sqrt{\frac{d}{q_m N}}$$

with probability at least $1 - cN^{-10}$, where we utilize (C.10) and (C.5) for $\boldsymbol{Z}^{(i)}$ as well. Thus, we know that $\|h^{(i)}(\boldsymbol{X})\| \lesssim 1$ with probability at least $1 - cN^{-10}$. Then, analogously with Lemma C.6, we have

$$\begin{aligned}
\frac{1}{\sqrt{N}}\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)}\| \leq\ & \|(h^{(i)}(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h^{(i)}(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h^{(i)}(\boldsymbol{X})^\top - (h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1} h(\boldsymbol{X})^\top\| \\
\leq\ & \|(h^{(i)}(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h^{(i)}(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1}\| \cdot \|(h^{(i)}(\boldsymbol{X}) - h(\boldsymbol{X}))\| + \|(h^{(i)}(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h^{(i)}(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1}\| \\
& \cdot \|h^{(i)}(\boldsymbol{X}) - h(\boldsymbol{X})\| \cdot (\|h^{(i)}(\boldsymbol{X})\| + \|h(\boldsymbol{X})\|) \cdot \|(h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}} h(\boldsymbol{X}) + \lambda \boldsymbol{I}_d)^{-1}\| \cdot \|h(\boldsymbol{X})\| \\
\lesssim\ & \|h^{(i)}(\boldsymbol{X}) - h(\boldsymbol{X})\| \lesssim \sqrt{\frac{d}{q_m N}},
\end{aligned}$$

with a probability at least $1 - cN^{-10}$, for some constant $c > 0$. Also, with Lemma C.6, we can show that $\|\widehat{\boldsymbol{\beta}}^{(i)}\| \lesssim \sqrt{N}$ with very high probability. □

Lemma C.9. *Under the same assumption as Lemma C.6, for each $i \in \mathcal{V}_{\mathbb{U}}$ the estimator $\widehat{\boldsymbol{\beta}}^{(i)}$ defined in (C.13) satisfies*

$$\frac{1}{\sqrt{N}}\|\widehat{\boldsymbol{\beta}}^{(i)} - \widetilde{\boldsymbol{\beta}}\| \leq \frac{C}{\sqrt{q_m}},$$

with probability at least $1 - cN^{-10}$, where $\widetilde{\boldsymbol{\beta}}$ is defined in Lemma C.6, for all large m and n and some constants $c, C > 0$.

The proof of this lemma is the same as Lemma C.6, so we ignore the details here.

Proof of Theorem 3.6. Recall that

$$\zeta = \frac{\kappa\tau}{\kappa^2\tau + \lambda} \quad \text{and} \quad \kappa = \sqrt{c_\tau} \cdot \frac{a_\tau - b_\tau + 2s}{a_\tau + b_\tau + 2s},$$

where both ζ and κ are some constants in $\mathbb{R}$. Hence, we know that

$$\frac{\kappa_m \tau}{\kappa_m^2 \tau + \lambda} = \zeta(1 + o(1)).$$

Then, $\widetilde{\boldsymbol{\beta}}/\sqrt{N} = \zeta\boldsymbol{\mu} + o(1/\sqrt{N})$. Denote $\boldsymbol{y}_{\mathbb{U},i}$ as the i-th entry of label $\boldsymbol{y}_{\mathbb{U}}$. Firstly, we consider the general case when $\rho = sq_m$ for some fixed constant $s \in \mathbb{R}$. For each $i \in \mathcal{V}_{\mathbb{U}}$, we can utilize Lemmas C.7, C.8, and C.9 to obtain that

$$\begin{aligned}
\left|\frac{y_i}{\sqrt{N}}\bar{\boldsymbol{h}}_i^\top \widetilde{\boldsymbol{\beta}} - \frac{y_i}{\sqrt{N}}\boldsymbol{h}_i^\top \widehat{\boldsymbol{\beta}}\right| \leq\ & \frac{1}{\sqrt{N}}\left|(\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i)^\top \widetilde{\boldsymbol{\beta}}\right| + \frac{1}{\sqrt{N}}\left|(\bar{\boldsymbol{h}}_i - \boldsymbol{h}_i)^\top (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})\right| \\
& + \frac{1}{\sqrt{N}}\left|\bar{\boldsymbol{h}}_i^\top (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})\right| + \frac{1}{\sqrt{N}}\left|\boldsymbol{h}_i^\top (\widehat{\boldsymbol{\beta}}^{(i)} - \widehat{\boldsymbol{\beta}})\right| \\
\leq\ & \frac{C}{\sqrt{Nq_m}} + C\frac{d}{Nq_m},
\end{aligned}$$

with probability at least $1 - cN^{-10}$ for some constants $c, C > 0$. Here, we applied (C.6) when $\boldsymbol{u} = \widetilde{\boldsymbol{\beta}}/\sqrt{N}$ and $\boldsymbol{u} = (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})/\sqrt{N}$, (C.7) when $\boldsymbol{u} = (\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}^{(i)})/\sqrt{N}$, and (C.8). Then, if $d \lesssim \sqrt{Nq_m}$, we can conclude that

$$\left|\frac{y_i}{\sqrt{N}}\bar{\boldsymbol{h}}_i^\top \widetilde{\boldsymbol{\beta}} - \frac{y_i}{\sqrt{N}}\boldsymbol{h}_i^\top \widehat{\boldsymbol{\beta}}\right| \leq \frac{C}{\sqrt{Nq_m}},$$

with very high probability for some universal constant $C > 0$.

Therefore, we can take $\varepsilon_m = 1/\sqrt{q_m}$ to get

$$\begin{aligned}
\mathbb{P}(\psi_m(\operatorname{sign}(\widehat{\boldsymbol{y}}_{\mathbb{U}}), \boldsymbol{y}_{\mathbb{U}}) = 0) =\ & \mathbb{P}\big(\min_{i\in[m]} \boldsymbol{y}_{\mathbb{U},i} \cdot \widehat{\boldsymbol{y}}_{\mathbb{U},i} > 0\big) = \mathbb{P}\Big(\min_{i\in\mathcal{V}_{\mathbb{U}}} \frac{y_i \cdot \boldsymbol{h}_i^\top \widehat{\boldsymbol{\beta}}}{\sqrt{N}} > 0\Big) \\
\geq\ & \mathbb{P}\Big(\min_{i\in\mathcal{V}_{\mathbb{U}}} \frac{y_i}{\sqrt{N}} \cdot \bar{\boldsymbol{h}}_i^\top \tilde{\boldsymbol{\beta}} > \frac{C\varepsilon_m}{\sqrt{N}}\Big) - \sum_{i\in\mathcal{V}_{\mathbb{U}}} \mathbb{P}\Big(\Big|\frac{y_i}{\sqrt{N}} \bar{\boldsymbol{h}}_i^\top \tilde{\boldsymbol{\beta}} - \frac{y_i}{\sqrt{N}} \boldsymbol{h}_i^\top \widehat{\boldsymbol{\beta}}\Big| > \frac{C\varepsilon_m}{\sqrt{N}}\Big) \\
\geq\ & \mathbb{P}\big(\min_{i\in\mathcal{V}_{\mathbb{U}}} y_i \zeta \cdot \frac{1}{\tilde{d}} (\boldsymbol{A}_\rho \boldsymbol{X})_{i:} \boldsymbol{\mu} > C\sqrt{q_m}\varepsilon_m\big) - Cm^{-2} \\
\geq\ & 1 - \sum_{i\in\mathcal{V}_{\mathbb{U}}} \mathbb{P}\Big(y_i \cdot \frac{\zeta}{\tilde{d}} (\boldsymbol{A}_\rho \boldsymbol{X})_{i:} \boldsymbol{\mu} \leq C\sqrt{q_m}\varepsilon_m\Big) - Cm^{-2} \\
\geq\ & 1 - m \cdot \mathbb{P}\Big(y_i \cdot \frac{\zeta}{\tilde{d}} (\boldsymbol{A}_\rho \boldsymbol{X})_{i:} \boldsymbol{\mu} \leq C\sqrt{q_m}\varepsilon_m\Big) - Cm^{-2} \\
\geq\ & 1 - m^{1-\sup_{t\in\mathbb{R}}\{\varepsilon_m t + g(a,b,c,\tau,\zeta,s,t)\}+\delta} - Cm^{-2},
\end{aligned}$$

for any $\delta > 0$ and sufficiently large m, where in the last line we employ Proposition C.1. Thus, applying Lemma C.2, we know that when $J(a_\tau, b_\tau, c_\tau, \zeta, s) > 1$, $\mathbb{P}(\psi_m(\operatorname{sign}(\widehat{\boldsymbol{y}}_{\mathbb{U}}), \boldsymbol{y}_{\mathbb{U}}) = 0) \to 1$ as $m \to \infty$.

When $s = \frac{2c_\tau}{\log(a/b)}$, Lemma C.2 implies that $J(a_\tau, b_\tau, c_\tau, \zeta, s) = I(a_\tau, b_\tau, c_\tau)$ defined in (3.2). Notice that $J(a_\tau, b_\tau, c_\tau, \zeta, s) \leq I(a_\tau, b_\tau, c_\tau)$ for any $s \in \mathbb{R}$. Whereas $s = 0$, Lemma C.2 implies that $J(a_\tau, b_\tau, c_\tau, \zeta, s) = I(a_\tau, b_\tau, 0)$. Hence, this completes the proof of this theorem. □

C.4. Asymptotic Errors for Ridge Regression on Linear GCN

Lemma C.10. *Under the Assumption 3.1, there exist some constant $c, C > 0$ such that with probability at least $1 - cN^{-2}$,*

$$\begin{aligned}
|\overline{\mathcal{R}}(\lambda) - \mathcal{R}(\lambda)| \leq\ & \frac{C}{\sqrt{q_m}}, \\
|\overline{\mathcal{E}}(\lambda) - \mathcal{E}(\lambda)| \leq\ & \frac{C}{\sqrt{q_m}},
\end{aligned}$$

where

$$\begin{aligned}
\overline{\mathcal{R}}(\lambda) :=\ & \frac{1}{m}(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y})^\top \boldsymbol{P}_{\mathbb{U}} (\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y}), \\
\overline{\mathcal{E}}(\lambda) :=\ & \frac{1}{n}(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y})^\top \boldsymbol{P}_{\mathbb{L}} (\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y}).
\end{aligned}$$

Proof. From Lemmas C.5 and C.6, we know that

$$\frac{1}{\sqrt{m}}\|\boldsymbol{H}\tilde{\boldsymbol{\beta}} - h(\boldsymbol{X})\widehat{\boldsymbol{\beta}}\| \leq \frac{1}{\sqrt{m}}\|\boldsymbol{H}\| \cdot \|\tilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\| + \frac{1}{\sqrt{m}}\|\boldsymbol{H} - h(\boldsymbol{X})\| \cdot \|\widehat{\boldsymbol{\beta}}\| \lesssim \frac{1}{\sqrt{q_m}},$$

with probability at least $1 - CN^{-2}$ for some constant $C > 0$. Since $\|\boldsymbol{P}_{\mathbb{L}}\|$ and $\|\boldsymbol{P}_{\mathbb{U}}\|$ are both upper bounded by one, we can directly conclude Lemma C.10. □

Proof of Theorem 3.7. Based on Lemma C.10, we can instead compute $\overline{\mathcal{E}}(\lambda)$ and $\overline{\mathcal{R}}(\lambda)$. Recall that $\boldsymbol{H} = \frac{\kappa_m}{\sqrt{N}}\boldsymbol{y}\boldsymbol{\mu}^\top$ and $\frac{1}{\sqrt{N}}\widetilde{\boldsymbol{\beta}} = \frac{\kappa_m \tau}{\kappa_m^2 \tau + \lambda}\boldsymbol{\mu}$. Thus, $\boldsymbol{H}\tilde{\boldsymbol{\beta}} = \frac{\kappa_m^2 \tau}{\kappa_m^2 \tau + \lambda}\boldsymbol{y}$. Then, since $\frac{1}{m}\boldsymbol{y}^\top \boldsymbol{P}_{\mathbb{U}} \boldsymbol{y} = \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{P}_{\mathbb{L}} \boldsymbol{y} = 1$, we have

$$\begin{aligned}
\overline{\mathcal{R}}(\lambda) =\ & \frac{1}{m}(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y})^\top \boldsymbol{P}_{\mathbb{U}} (\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y}) = \Big(1 - \frac{\kappa_m^2 \tau_n}{\kappa_m^2 \tau_n + \lambda}\Big)^2 = \frac{\lambda^2}{(\kappa^2 \tau + \lambda)^2} + o(1), \\
\overline{\mathcal{E}}(\lambda) =\ & \frac{1}{n}(\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y})^\top \boldsymbol{P}_{\mathbb{L}} (\boldsymbol{H}\tilde{\boldsymbol{\beta}} - \boldsymbol{y}) = \Big(1 - \frac{\kappa_m^2 \tau_n}{\kappa_m^2 \tau_n + \lambda}\Big)^2 = \frac{\lambda^2}{(\kappa^2 \tau + \lambda)^2} + o(1).
\end{aligned}$$

Then taking $m \to \infty$, we can get the results of this lemma. □

D. Feature Learning of Graph Convolutional Networks

In this section, we complete the proof of Theorem 3.9 in Section 3.4. Recall that

$$\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} - \eta_1 \Big(\nabla_{\boldsymbol{W}^{(0)}} \mathcal{L}(\boldsymbol{W}^{(0)}, s^{(0)}) + \lambda_1 \boldsymbol{W}^{(0)} \Big). \tag{D.1}$$

and

$$s^{(1)} = \frac{2}{n^2 q_m} \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{X}_{\mathbb{L}} \boldsymbol{W}^{(1)} \boldsymbol{a} \Big/ \log \left(\frac{N \cdot D_0 + \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}} \boldsymbol{y}_{\mathbb{L}}}{N \cdot D_0 - \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}} \boldsymbol{y}_{\mathbb{L}}} \right), \tag{D.2}$$

The algorithm we applied in Theorem 3.9 is given by Algorithm 1. Below, we first construct an optimal solution for this problem and present the LDP analysis. Then, we present will use (Ba et al., 2022; Damian et al., 2022) to analyze the feature learned from $\boldsymbol{W}^{(1)}$. Finally, inspired by the optimal solution, we will prove $s^{(1)}$ is close to the optimal s in (3.11). Meanwhile, we also present an additional gradient-based method in Algorithm 2 to approach the optimal s in (3.11). We will leave the theoretical analysis for Algorithm 2 as a future direction to explore.

Algorithm 1 Gradient-based training for GCN in Theorem 3.9

Input: Learning rates η_1, weight decay λ_1
Initialization: $s^{(0)} = 0$, $\sqrt{K} \cdot [\boldsymbol{W}_0]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\sqrt{K} \cdot [\boldsymbol{a}]_j \overset{\text{i.i.d.}}{\sim} \text{Unif}\{\pm 1\}$, for all $i \in [d], j \in [K]$.
Training Stage 1:
Set $\sigma(x) = x$ in (3.12)
$\boldsymbol{W}^{(1)} \leftarrow \boldsymbol{W}^{(0)} - \eta_1 (\nabla_{\boldsymbol{W}^{(0)}} \mathcal{L}(\boldsymbol{W}^{(0)}, s^{(0)}) + \lambda_1 \boldsymbol{W}^{(0)})$
Training Stage 2:
$s^{(1)} \leftarrow s^{(0)} + \frac{2}{n^2 q_m} \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{X}_{\mathbb{L}} \boldsymbol{W}^{(1)} \boldsymbol{a} \Big/ \log \left(\frac{N \cdot D_0 + \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}} \boldsymbol{y}_{\mathbb{L}}}{N \cdot D_0 - \boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}} \boldsymbol{y}_{\mathbb{L}}} \right)$
return Prediction function for unknown labels: $\text{sign}(\boldsymbol{S}_{\mathbb{U}} \boldsymbol{D}_{s^{(1)}}^{-1} \boldsymbol{A}_{s^{(1)}} \boldsymbol{W}^{(1)} \boldsymbol{a})$

D.1. Thresholds for GCNs and LDP analysis

Consider $(\boldsymbol{A}, \boldsymbol{X}) \sim \text{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$. Let $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ denote the adjacency matrix of the graph G and let us define the degree matrix by $\boldsymbol{D}_0 := \text{diag}\{D_0, \ldots, D_0\} \in \mathbb{R}^{N \times N}$ where $D_0 = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^N \boldsymbol{A}_{ij}$. Let $\rho = s q_m$ denote the self-loop intensity (Kipf & Welling, 2017; Wu et al., 2019; Shi et al., 2024) for some $s \in \mathbb{R}$, and $\boldsymbol{A}_s = \boldsymbol{A} + \rho \boldsymbol{I}$, $\boldsymbol{D}_s = \boldsymbol{D} + \rho \boldsymbol{I}$ denote the adjacency, average degree matrices of the graph after adding self-loops, respectively.

The convolutional feature vector is $\widetilde{\boldsymbol{x}}_i := \big((\boldsymbol{D}_s)^{-1} \boldsymbol{A}_s \boldsymbol{X} \big)_{i:}$. Ideally, our goal is to prove that the convoluted feature vectors are linearly separable, i.e., find some $\boldsymbol{w} \in \mathbb{R}^d$ such that

$$\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{w} + b > 0 \text{ if } y_i = 1, \quad \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{w} + b < 0 \text{ if } y_i = -1,$$

for some $b \geq 0$. We consider the case that the feature learned by the GCN is exactly $\boldsymbol{\mu}$ in GMM, i.e., the optimal margin is $\boldsymbol{w} = \boldsymbol{\mu}$. Under this setting, we show the LDP results for this estimator. The proof is similar to Proposition C.1.

Proposition D.1 (LDP for GCNs). *For $(\boldsymbol{A}, \boldsymbol{X}) \sim \text{CSBM}(\boldsymbol{y}, \boldsymbol{\mu}, \alpha, \beta, \theta)$ with Assumption 3.1, when $s = 2c_\tau / \log(\frac{a}{b})$ and $c_\tau = \theta^2 / q_m + o(1)$, we have that*

$$\lim_{m \to \infty} q_m^{-1} \log \mathbb{P}(y_i \cdot \sqrt{q_m} (\mathbb{E} \boldsymbol{D}_s)^{-1} (\boldsymbol{A}_s \boldsymbol{X})_{i:} \boldsymbol{\mu} \leq \varepsilon q_m) = -\sup_{t \in \mathbb{R}} \{ \varepsilon t + I(a_\tau, b_\tau, c_\tau, t), \}$$

where $\sup_{t \in \mathbb{R}} I(a_\tau, b_\tau, c_\tau, t) = I(a_\tau, b_\tau, c_\tau)$ defined in (3.2).

Proof. For simplicity, let $\widetilde{d} := \frac{1}{N}\sum_{i=1}^{N}\widetilde{D}_i = \frac{a+b}{2}q_m + \rho$, denoting the expected degree of each node $i \in [N]$. Then

$$\begin{aligned} h_i :=& y_i\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{w} = y_i\theta(\mathbb{E}\widetilde{\boldsymbol{D}})^{-1}(\widetilde{\boldsymbol{A}}\boldsymbol{X})_{i:}\boldsymbol{\mu} \\ =& \frac{y_i\theta}{\widetilde{d}}\Big(\sum_{j\in\mathcal{N}_i}\boldsymbol{\mu}^\top(\theta y_j\boldsymbol{\mu} + \boldsymbol{z}_j) + \rho\boldsymbol{\mu}^\top(\theta y_i\boldsymbol{\mu} + \boldsymbol{z}_i)\Big) \\ =& \underbrace{\frac{\rho\theta^2 y_i^2\|\boldsymbol{\mu}\|_2^2}{\widetilde{d}}}_{I_1} + \underbrace{\frac{\rho\theta y_i\boldsymbol{\mu}^\top\boldsymbol{z}_i}{\widetilde{d}}}_{I_2} + \underbrace{\frac{\theta^2\|\boldsymbol{\mu}\|_2^2}{\widetilde{d}}\sum_{j\neq i}A_{ij}y_iy_j}_{I_3} + \underbrace{\frac{\theta}{\widetilde{d}}\sum_{j\neq i}y_iA_{ij}\boldsymbol{\mu}^\top\boldsymbol{z}_j}_{I_4}. \end{aligned}$$

Our goal is to calculate the following moment-generating function

$$\mathbb{E}[\exp(th_i)] := \mathbb{E}_{\boldsymbol{A}}[\mathbb{E}_{\boldsymbol{X}}[\exp(th_i)|\boldsymbol{A}]].$$

First, since $\|\boldsymbol{\mu}\|_2 = 1$, $y_i^2 = 1$, $I_1 = \rho\theta^2/\widetilde{d}^2$, and it is deterministic. Second, $\boldsymbol{\mu}^\top\boldsymbol{z}_i \sim \mathcal{N}(0,1)$, then $I_2 \sim \mathcal{N}(0, \rho^2\theta^2/\widetilde{d}^2)$, and

$$\mathbb{E}_{\boldsymbol{X}}[\exp(tI_2)|y_i] = \exp\Big(\frac{t^2\rho^2\theta^2}{2\widetilde{d}^2}\Big) = \mathbb{E}[\exp(tI_2)],$$

where the last equality holds since the result we obtained is independent of y_i. Let $\mathcal{N}_i$ denote the set of neighbors of node i and $|\mathcal{N}_i|$ denote its cardinality. Conditioned on $\boldsymbol{A}, \boldsymbol{y}, \boldsymbol{\mu}$, $I_4 \sim \mathcal{N}(0, |\mathcal{N}_i|\theta^2/\widetilde{d}^2)$, and

$$\mathbb{E}_{\boldsymbol{X}}[\exp(tI_4)|\boldsymbol{A}, y_i, \boldsymbol{\mu}] = \exp\Big(\frac{t^2\theta^2|\mathcal{N}_i|}{2\widetilde{d}^2}\Big).$$

Note that $|\mathcal{N}_i| = \sum_{j=1}^{N}A_{ij}$, and I_3 is independent of $\boldsymbol{X}$, then

$$\mathbb{E}_{\boldsymbol{X}}[\exp\big(t(I_3 + I_4)\big)|\boldsymbol{A}, y_i, \boldsymbol{\mu}] = \exp\Bigg(\frac{t\theta^2}{\widetilde{d}}\sum_{j\neq i}A_{ij}\Big(y_iy_j + \frac{t}{2\widetilde{d}}\Big)\Bigg)$$

One could take the expectation over $\boldsymbol{A}$ conditioned on $\boldsymbol{y}$, then

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{A}}\Big[\exp\Big(\frac{t\theta^2}{\widetilde{d}}A_{ij}\Big(y_iy_j + \frac{t}{2\widetilde{d}}\Big)\Big)\Big|y_i\Big] \\ =& \frac{1}{2}\mathbb{E}_{\boldsymbol{A}}\Big[\exp\Big(\frac{t\theta^2}{\widetilde{d}}A_{ij}\Big(y_iy_j + \frac{t}{2\widetilde{d}}\Big)\Big)\Big|y_iy_j = 1\Big] + \frac{1}{2}\mathbb{E}_{\boldsymbol{A}}\Big[\exp\Big(\frac{t\theta^2}{\widetilde{d}}A_{ij}\Big(y_iy_j + \frac{t}{2\widetilde{d}}\Big)\Big)\Big|y_iy_j = -1\Big] \\ =& \frac{1}{2}\Big[\alpha\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} + \frac{t\theta^2}{\widetilde{d}}\Big) + (1-\alpha) + \beta\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} - \frac{t\theta^2}{\widetilde{d}}\Big) + (1-\beta)\Big] \\ =& 1 + \frac{\alpha}{2}\Big[\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} + \frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big] + \frac{\beta}{2}\Big[\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} - \frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big], \end{aligned}$$

where the result is again independent of $y_i, \boldsymbol{\mu}$. Recall $\alpha = aq_m/m = o(1)$, $\beta = bq_m/m = o(1)$, $\frac{\theta^4}{\theta^2 + (1-\tau)d/m} = c_\tau q_m$ in Assumption 3.1, thus $\theta^2 = (1 + o(1))c_\tau q_m$. By using $\log(1+x) = x$ for $x = o(1)$, we then have

$$\begin{aligned} q_m^{-1}\log\mathbb{E}_{\boldsymbol{A}}[\mathbb{E}_{\boldsymbol{X}}[\exp\{t(I_3 + I_4)\}]] =& \log\Bigg(1 + \frac{\alpha}{2}\Big[\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} + \frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big] + \frac{\beta}{2}\Big[\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} - \frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big]\Bigg) \\ =& \frac{a}{2}\Big[\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} + \frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big] + \frac{b}{2}\Big[\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} - \frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big] \\ =& (1 + o(1))\frac{a}{2}\Big[\exp\Big(\frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big] + (1 + o(1))\frac{b}{2}\Big[\exp\Big(-\frac{t\theta^2}{\widetilde{d}}\Big) - 1\Big] \end{aligned}$$

Combining the calculations above, compute the following rate function

$$g(a, b, c_\tau, t) := -q_m^{-1}\log\mathbb{E}[\exp(th_i)].$$

Recall $\alpha = aq_m/m = o(1)$, $\beta = bq_m/m = o(1)$, $\frac{\theta^4}{\theta^2+(1-\tau)d/m} = c_\tau q_m$ in Assumption 3.1, thus $\theta^2 = (1+o(1))c_\tau q_m$. By using $\log(1+x) = x$ for $x = o(1)$, the rate function $g(a,b,c_\tau,t)$ can be calculated as

$$g(a,b,c_\tau,t) = -\frac{t\rho\theta^2}{q_m\widetilde{d}} - \frac{t^2\rho^2\theta^2}{2q_m\widetilde{d}^2} + \frac{(N-1)}{2m}\Big[a - a\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} + \frac{t\theta^2}{\widetilde{d}}\Big) + b - b\exp\Big(\frac{t^2\theta^2}{2\widetilde{d}^2} - \frac{t\theta^2}{\widetilde{d}}\Big)\Big]$$
$$= \frac{1}{2(1-\tau)}\Big[a\Big(1-\exp\Big(\frac{2c_\tau t}{a+b+2s}\Big)\Big) + b\Big(1-\exp\Big(-\frac{2c_\tau t}{a+b+2s}\Big)\Big)\Big] - \frac{2c_\tau st}{a+b+2s} - \frac{2c_\tau s^2t^2}{(a+b+2s)^2},$$

where in the last line, we used $\rho = sq_m$, $\widetilde{d} = (\frac{a+b}{2}+s)q_m$. By choosing $s = \frac{2c_\tau}{\log(a/b)}$, we can conclude that

$$I^\star = \sup_{t\in\mathbb{R}} I(a_\tau,b_\tau,c_\tau,t) = \frac{(1-\tau)^{-1}(\sqrt{a}-\sqrt{b})^2 + c_\tau}{2} \equiv I(a_\tau,b_\tau,c_\tau),$$

which completes the proof. □

D.2. Gradient descent for the first layer weight matrix $\boldsymbol{W}$

For simplicity, we denote $\widetilde{\boldsymbol{X}} = \boldsymbol{D}_s^{-1}\boldsymbol{A}_s\boldsymbol{X} = (\widetilde{\boldsymbol{x}}_1,\ldots,\widetilde{\boldsymbol{x}}_N)^\top$ where $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^d$ for $i\in[N]$ and $s=0$. In this case, we will explore the feature learning on $\boldsymbol{W}$. Below, we will always fix $\boldsymbol{a}$ (at initialization in Assumption 3.8) and perform gradient descent on $\boldsymbol{W}$ in (D.1). To ease the notions, we write the initialized first-layer weights as $\boldsymbol{W}^{(0)} = \boldsymbol{W}_0$, and the weights after one gradient step as $\boldsymbol{W}^{(1)} = \boldsymbol{W}_1$, where the learning rate of the first gradient descent is $\eta_1 > 0$. Let $s^{(0)} = 0$. Following the notions in (Ba et al., 2022), we denote that

$$\boldsymbol{G}_1 := -\nabla_{\boldsymbol{W}_0}\mathcal{L}(\boldsymbol{W}_0, s^{(0)}) = \frac{1}{n}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\left[\left(\frac{1}{\sqrt{K}}\left(\boldsymbol{y}_{\mathbb{L}} - \frac{1}{\sqrt{K}}\sigma(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\boldsymbol{a}\right)\boldsymbol{a}^\top\right)\odot\sigma'(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\right],$$

where $\widetilde{\boldsymbol{X}}_{\mathbb{L}} = \boldsymbol{S}_{\mathbb{L}}\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n\times d}$, $\odot$ is the Hadamard product, and σ' is the derivative of σ (acting entry-wise). Here K represents the number of neurons in the hidden GCN layer in (3.12). Then, from (D.1) with $\lambda_1 = \eta_1^{-1}$, we have

$$\boldsymbol{W}_1 = \boldsymbol{W}_0 + \eta_1\cdot\boldsymbol{G}_1 - \boldsymbol{W}_0 = \eta_1\cdot\boldsymbol{G}_1.$$

Thus, our target is to analyze the gradient matrix $\boldsymbol{G}_1$. The following proposition is similar to Proposition 2 in (Ba et al., 2022), implies that this gradient matrix is approximately rank one.

Proposition D.2. *Under the same assumption as Theorem 3.9, we have that*

$$\left\|\boldsymbol{G}_1 - \frac{1}{n\sqrt{K}}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\boldsymbol{y}_{\mathbb{L}}\boldsymbol{a}^\top\right\|_F \le \frac{Cq_m}{K},$$

with probability at least $1-\exp(-c\log^2 N)$, *for some constant* $c, C > 0$.

Proof. First of all, analogously to the proof of Lemma C.5, we can show that

$$\left\|\widetilde{\boldsymbol{X}}_{\mathbb{L}}\right\| \le \sqrt{q_mN}, \tag{D.3}$$

with very high probability, since $d \lesssim N$. Moreover, $\|\boldsymbol{y}_{\mathbb{L}}\| = \sqrt{n}$ and we can always view $\boldsymbol{y}$ as a deterministic vector in $\mathbb{R}^N$. By the definition, the gradient matrix $\boldsymbol{G}_1$ under the MSE can be simplified as follows

$$\boldsymbol{G}_1 = -\frac{1}{n}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\left[\left(\frac{1}{\sqrt{K}}\left(\frac{1}{\sqrt{K}}\sigma(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\boldsymbol{a} - \boldsymbol{y}_{\mathbb{L}}\right)\boldsymbol{a}^\top\right)\odot\sigma'(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\right]$$
$$= \frac{1}{n}\cdot\frac{\mu_1}{\sqrt{K}}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\left(\boldsymbol{y}_{\mathbb{L}} - \frac{1}{\sqrt{K}}\sigma(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\boldsymbol{a}\right)\boldsymbol{a}^\top + \frac{1}{n}\cdot\frac{1}{\sqrt{K}}\boldsymbol{X}^\top\left(\left(\boldsymbol{y}_{\mathbb{L}} - \frac{1}{\sqrt{N}}\sigma(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\boldsymbol{a}\right)\boldsymbol{a}^\top\odot\sigma'_\perp(\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0)\right),$$

where we utilized the orthogonal decomposition: $\sigma'(z) = \mu_1 + \sigma'_\perp(z)$. By Stein's lemma, we know that $\mathbb{E}[z\sigma(z)] = \mathbb{E}[\sigma'(z)] = \mu_1$, and hence $\mathbb{E}[\sigma'_\perp(z)] = 0$ for $z\sim\mathcal{N}(0,1)$. Notice that we consider $\sigma(x) = x$, hence $\mu_1 = 1$ and $\sigma'_\perp(z)\equiv 0$. Therefore, we have

$$\boldsymbol{G}_1 = \frac{1}{n}\cdot\frac{1}{\sqrt{K}}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\boldsymbol{y}_{\mathbb{L}}\boldsymbol{a}^\top - \underbrace{\frac{1}{nK}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\widetilde{\boldsymbol{X}}_{\mathbb{L}}\boldsymbol{W}_0\boldsymbol{a}\boldsymbol{a}^\top}_{\boldsymbol{\Delta}}.$$

Notice that $\|\boldsymbol{a}\| = 1$ and we can apply (C.5) for Gaussian random matrix $\boldsymbol{W}_0$. Thus, because of Lemma C.5, $d \lesssim n \asymp N$ and $d \lesssim K$, we have that

$$\|\boldsymbol{\Delta}\|_F = \|\boldsymbol{\Delta}\| \leq \frac{Cq_m}{K}(1 + \sqrt{d/K}),$$

with very high probability, which completes the proof of this proposition.

□

This proposition shows that for $\boldsymbol{W}$ at Gaussian initialization, the corresponding gradient matrix can be approximated in operator norm by the **rank-1** matrix only related to labels $\boldsymbol{y}_{\mathbb{L}}$, feature matrix $\widetilde{\boldsymbol{X}}_{\mathbb{L}}$, and $\boldsymbol{a}$.

In the following, we will use the parameter

$$\zeta := \sqrt{c_\tau}\frac{\eta_1\sqrt{q_m}}{K}\frac{\alpha - \beta}{\alpha + \beta}.$$

Notice that $\zeta = \Theta(1)$ if $K/\eta_1 = \Theta(\sqrt{q_m})$. Then we can tune the learning rate η_1 to ensure that this trained and normalized weight matrix $\frac{1}{\sqrt{K}}\boldsymbol{W}^{(1)}$ can be aligned with $\boldsymbol{\mu}$ perfectly.

Lemma D.3. *Under the assumption as Theorem 3.9, we have that*

$$\left\| \frac{1}{\sqrt{K}}\boldsymbol{W}^{(1)}\boldsymbol{a} - \sqrt{c_\tau}\frac{\eta_1\sqrt{q_m}}{K}\frac{\alpha - \beta}{\alpha + \beta}\boldsymbol{\mu} \right\| = O(\frac{\eta_1}{K}),$$

with a probability at least $1 - cN^{-10}$, for some constants $c, C > 0$.

Proof. Notice that $\boldsymbol{W}_1 = \eta_1 \cdot \boldsymbol{G}_1$ and $\boldsymbol{a}^\top\boldsymbol{a} = 1$. Notice that $\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top \boldsymbol{y}_{\mathbb{L}} = \sqrt{Nq_m} \cdot h(\boldsymbol{X})^\top \boldsymbol{P}_{\mathbb{L}}\boldsymbol{y}$. Following from Proposition D.2 and Lemma C.7, we can have

$$\begin{aligned} \left\| \frac{1}{\sqrt{K}}\boldsymbol{W}^{(1)}\boldsymbol{a} - \zeta\boldsymbol{\mu} \right\| \leq &\ \frac{1}{\sqrt{K}}\left\|\eta_1\boldsymbol{\Delta}\boldsymbol{a}\right\| + \left\|\zeta\boldsymbol{\mu} - \frac{\eta_1}{nK}\widetilde{\boldsymbol{X}}_{\mathbb{L}}^\top\boldsymbol{y}_{\mathbb{L}}\boldsymbol{a}^\top\boldsymbol{a}\right\| \\ \lesssim &\ \frac{\eta_1}{K} + \frac{\eta_1 q_m}{K^{3/2}}, \end{aligned}$$

with very high probability. Notice that here $\boldsymbol{a}^\top\boldsymbol{a} = 1$. Then, we can assume $q_m^2 \lesssim K$ to finish this proof. □

D.3. Learning the optimal self-loop weight

Lemma D.4. *Under Assumption 3.1, we know that*

$$\left|D_0 - \bar{d}\right| \leq \frac{C}{q_m^{1/2}},$$

with probability at least $1 - ce^{-N}$ for some constants $c, C > 0$, where $\bar{d} := \frac{a_\tau + b_\tau}{2}q_m$.

This is straightforward based on the proof of Lemma C.3, hence we ignore the proof here.

Lemma D.5. *Under the assumption as Theorem 3.9, we have that*

$$\left|\frac{2}{n^2 q_m}\boldsymbol{y}_{\mathbb{L}}^\top\boldsymbol{X}_{\mathbb{L}}\boldsymbol{W}^{(1)}\boldsymbol{a} - 2c_\tau\right| \leq \frac{C}{n\sqrt{q_m}}$$

with probability at least $1 - cN^{-10}$, for some constants $c, C > 0$.

Proof. By Proposition D.2 and Lemma D.3, we can replace $\boldsymbol{W}^{(1)}\boldsymbol{a}$ by $\zeta\boldsymbol{\mu}$. Notice that (D.3) and Lemma D.3 indicate that

$$\left|\frac{2}{n^2 q_m}\boldsymbol{y}_{\mathbb{L}}^\top\boldsymbol{X}_{\mathbb{L}}(\boldsymbol{W}^{(1)}\boldsymbol{a} - \zeta\boldsymbol{\mu})\right| \leq \frac{2}{n^2 q_m}\|\boldsymbol{y}_{\mathbb{L}}\| \cdot \|\boldsymbol{X}_{\mathbb{L}}\| \cdot \|\boldsymbol{W}^{(1)}\boldsymbol{a} - \zeta\boldsymbol{\mu}\| \lesssim 1/(nq_m),$$

Algorithm 2 Gradient-based training for both $\boldsymbol{W}$ and s in GCN

Input: Learning rates η_t, weight decay λ_t, number of steps T
Initialization: $s^{(0)} \sim \mathrm{Unif}([-1,1])$, $\sqrt{K}\cdot[\boldsymbol{W}_0]_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $\sqrt{K}\cdot[\boldsymbol{a}]_j \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{\pm 1\}$, for all $i\in[d], j\in[K]$.
Training Stage 1:
Set $\sigma(x)=x$ in (3.12)
$\boldsymbol{W}^{(1)} \leftarrow \boldsymbol{W}^{(0)} - \eta_1(\nabla_{\boldsymbol{W}^{(0)}}\mathcal{L}(\boldsymbol{W}^{(0)}, s^{(0)}) + \lambda_1 \boldsymbol{W}^{(0)})$
$s^{(1)} \leftarrow s^{(0)}$
$\boldsymbol{W}^{(1)} \leftarrow \boldsymbol{W}^{(1)}\boldsymbol{a}$
$\boldsymbol{a} \leftarrow 1$
Training Stage 2:
Set $\sigma(x)=\tanh(x)$ in (3.12)
for $t=2$ **to** T **do**
  $s^{(t)} \leftarrow s^{(t-1)} - \eta_t \nabla_{s^{(t)}}\mathcal{L}(\boldsymbol{W}^{(1)}, s^{(t-1)}) + \lambda_t s^{(t-1)}$
end for
return Prediction function for unknown labels: $\mathrm{sign}(\boldsymbol{S}_{\mathbb{U}}\boldsymbol{D}_{s^{(T)}}^{-1}\boldsymbol{A}_{s^{(T)}}\boldsymbol{W}^{(1)}\boldsymbol{a})$

with very high probability. for $s=0$. Then, we can apply Lemmas E.1 and C.4 to conclude that

$$\left|\frac{1}{n^2 q_m}\boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{X}_{\mathbb{L}}\zeta\boldsymbol{\mu} - c_\tau\right| \lesssim \frac{1}{n\sqrt{q_m}},$$

with a very high probability for sufficiently large n and m. □

Lemma D.6. *Following the assumptions in Theorem 3.9, we have that*

$$\left|\frac{1}{n}\boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}}\boldsymbol{y}_{\mathbb{L}} - \frac{a_\tau - b_\tau}{2}q_m\right| = o(q_m),$$

with a probability of at least $1-cN^{-10}$*, for some constants* $c, C>0$*.*

Proof. This lemma follows from Lemma F.6 in (Abbe et al., 2022) and Corollary 3.1 in (Abbe et al., 2020). Notice that

$$\begin{aligned}\frac{1}{n}\boldsymbol{y}_{\mathbb{L}}^\top \boldsymbol{A}_{\mathbb{L}}\boldsymbol{y}_{\mathbb{L}} &= \frac{1}{n}\sum_{i,j\in\mathcal{V}_{\mathbb{L}}}\boldsymbol{A}_{ij}y_i y_j\\ &= \frac{1}{n}\sum_{i,j \text{ in same block of } \mathcal{V}_{\mathbb{L}}}\boldsymbol{A}_{ij} - \sum_{i,j \text{ in different blocks of } \mathcal{V}_{\mathbb{L}}}\boldsymbol{A}_{ij}.\end{aligned}$$

Then, we can apply the proof of (F.23) in (Abbe et al., 2022) to conclude this lemma. □

Combining all the above lemmas in this section, we can derive the following lemma.

Lemma D.7. *Following the assumptions in Theorem 3.9, we have that*

$$\left|s^{(1)} - \frac{2c_\tau}{\log\left(\frac{a_\tau}{b_\tau}\right)}\right| \le \frac{C}{\sqrt{q_m}},$$

with probability at least $1-cN^{-10}$*, for some constants* $c, C>0$*, where* $s^{(1)}$ *is given by* (D.2)*.*

D.4. Proof of Theorem 3.9

Let us recall that we consider $K/\eta_1 \asymp \sqrt{q_m}$ and $d=o(q_m^2)$ with $\boldsymbol{W}_{i,j}\sim\mathcal{N}(0,1/K)$. Finally, we complete the proof of Theorem 3.9 for Algorithm 1 as follows. Recall that $\boldsymbol{D}_s := (D_0 + sq_m)\boldsymbol{I}\in\mathbb{R}^{n\times n}$, for any $s\in\mathbb{R}$, where D_0 is the average degree of the graph. Denote that

$$
\begin{aligned}
s_{\text{opt}} &:= \frac{2c_\tau}{\log\left(\frac{a_\tau}{b_\tau}\right)} \\
\boldsymbol{D}_{s^{(1)}}^{-1}\boldsymbol{A}_{s^{(1)}}\boldsymbol{X} &=: [\widehat{\boldsymbol{g}}_1,\dots,\widehat{\boldsymbol{g}}_N]^\top \in \mathbb{R}^{N\times d}, \\
\boldsymbol{D}_{s_{\text{opt}}}^{-1}\boldsymbol{A}_{s_{\text{opt}}}\boldsymbol{X} &=: [\bar{\boldsymbol{g}}_1,\dots,\bar{\boldsymbol{g}}_N]^\top \in \mathbb{R}^{N\times d}, \\
(\tilde{d}+s_{\text{opt}}\cdot q_m)^{-1}\boldsymbol{A}_{s_{\text{opt}}}\boldsymbol{X} &=: [\tilde{\boldsymbol{g}}_1,\dots,\tilde{\boldsymbol{g}}_N]^\top \in \mathbb{R}^{N\times d}, \\
\frac{1}{\sqrt{K}}\boldsymbol{W}^{(1)}\boldsymbol{a} &=: \widehat{\boldsymbol{\mu}}.
\end{aligned}
$$

Then, by definition, $\widehat{\boldsymbol{y}}_{\text{GCN},i} = \widehat{\boldsymbol{g}}_i^\top\widehat{\boldsymbol{\mu}}$ for $i\in\mathcal{V}_{\mathbb{U}}$. As a remark, notice that Lemma C.7 verifies that with high probability $\|\widehat{\boldsymbol{g}}_i\| \lesssim \sqrt{d}$. Because of this bound, we can only consider the regime when $d = o(q_m^2)$ for our following analysis. To improve this to a high dimensional regime, e.g., $d \asymp N$, we improve the following concentration without simply using the bound of $\|\widehat{\boldsymbol{g}}_i\|$. Similarly with the proof of Lemma C.8 in ridge regression of linear GCN part, we need to do certain leave-one-out analysis to achieve a larger regime for d.

Next, based on the above decomposition, we follow the proof idea of Theorem 3.6 to complete the proof of Theorem 3.9. Combining Lemmas C.7, D.3 and D.7, for each $i\in\mathcal{V}_{\mathbb{U}}$, we can obtain that

$$
|y_i\cdot\widehat{\boldsymbol{g}}_i^\top\widehat{\boldsymbol{\mu}} - y_i\cdot\tilde{\boldsymbol{g}}_i^\top\boldsymbol{\mu}| \le \left|(\tilde{\boldsymbol{g}}_i-\bar{\boldsymbol{g}}_i)^\top\boldsymbol{\mu}\right| + \left|(\bar{\boldsymbol{g}}_i-\widehat{\boldsymbol{g}}_i)^\top\boldsymbol{\mu}\right| + \left|\widehat{\boldsymbol{g}}_i^\top(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right| = o(\sqrt{q_m}),
$$

with probability at least $1-cN^{-10}$ for some constants $c, C>0$. Therefore, we can take $\zeta=1$, $\varepsilon_m = o(1)$ and $\rho = s_{\text{opt}}\cdot q_m$ to get

$$
\begin{aligned}
\mathbb{P}(\psi_m(\text{sign}(\widehat{\boldsymbol{y}}_{\text{GCN}}),\boldsymbol{y}_{\mathbb{U}})=0) =\ & \mathbb{P}\big(\min_{i\in[m]}\boldsymbol{y}_{\mathbb{U},i}\cdot\widehat{\boldsymbol{y}}_{\text{GCN},i}>0\big) = \mathbb{P}\Big(\min_{i\in\mathcal{V}_{\mathbb{U}}} y_i\cdot\widehat{\boldsymbol{g}}_i^\top\widehat{\boldsymbol{\mu}}>0\Big) \\
\ge\ & \mathbb{P}\Big(\min_{i\in\mathcal{V}_{\mathbb{U}}} y_i\cdot\tilde{\boldsymbol{g}}_i^\top\boldsymbol{\mu} > \varepsilon_m\sqrt{q_m},\ |y_i\cdot\widehat{\boldsymbol{g}}_i^\top\widehat{\boldsymbol{\mu}} - y_i\cdot\tilde{\boldsymbol{g}}_i^\top\boldsymbol{\mu}| \le \varepsilon_m\sqrt{q_m},\ \forall i\in\mathcal{V}_{\mathbb{U}}\Big) \\
\ge\ & \mathbb{P}\Big(\min_{i\in\mathcal{V}_{\mathbb{U}}} y_i\cdot\tilde{\boldsymbol{g}}_i^\top\boldsymbol{\mu} > \varepsilon_m\sqrt{q_m}\Big) - \sum_{i\in\mathcal{V}_{\mathbb{U}}}\mathbb{P}\Big(|y_i\cdot\widehat{\boldsymbol{g}}_i^\top\widehat{\boldsymbol{\mu}} - y_i\cdot\tilde{\boldsymbol{g}}_i^\top\boldsymbol{\mu}| > \varepsilon_m\sqrt{q_m}\Big) \\
\ge\ & \mathbb{P}\big(\min_{i\in\mathcal{V}_{\mathbb{U}}} y_i\zeta\cdot\frac{1}{\tilde{d}}(\boldsymbol{A}_\rho\boldsymbol{X})_{i:}\boldsymbol{\mu} > C\sqrt{q_m}\varepsilon_m\big) - Cm^{-2} \\
\ge\ & 1-\sum_{i\in\mathcal{V}_{\mathbb{U}}}\mathbb{P}\Big(y_i\cdot\frac{\zeta}{\tilde{d}}(\boldsymbol{A}_\rho\boldsymbol{X})_{i:}\boldsymbol{\mu} \le C\sqrt{q_m}\varepsilon_m\Big) - Cm^{-2} \\
\ge\ & 1-m\mathbb{P}\Big(y_i\cdot\frac{\zeta}{\tilde{d}}(\boldsymbol{A}_\rho\boldsymbol{X})_{i:}\boldsymbol{\mu} \le C\sqrt{q_m}\varepsilon_m\Big) - Cm^{-2} \\
\ge\ & 1-m^{1-\sup_{t\in\mathbb{R}}\{\varepsilon_m t + g(a,b,c,\tau,1,s_{\text{opt}},t)\}+\delta} - Cm^{-2},
\end{aligned}
$$

for any $\delta>0$ and sufficiently large m, where in the last line we employ Proposition D.1. Thus, applying Lemma C.2, we know that when $J(a_\tau,b_\tau,c_\tau,1,s_{\text{opt}}) = I(a_\tau,b_\tau,c_\tau)>1$, $\mathbb{P}(\psi_m(\text{sign}(\widehat{\boldsymbol{y}}_{\text{GCN}}),\boldsymbol{y}_{\mathbb{U}})=0)\to 1$ as $m\to\infty$.

E. Auxiliary Lemmas and Proofs

Lemma E.1. *Let $\boldsymbol{Z}\in\mathbb{R}^{N\times d}$ defined in (2.1). Then, there exists some constant $c, K>0$ such that for any $t>0$*

$$
\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{\sqrt{N}}\boldsymbol{1}^\top\boldsymbol{Z}\boldsymbol{\mu}\right|\ge t\right) &\le 2\exp(-ct^2 d), \\
\mathbb{P}\left(\left|\frac{1}{N}\boldsymbol{1}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{y}\right|\ge t\right) &\le 2\exp\left(-cd\min\left\{\frac{t^2}{K^2},\frac{t}{K}\right\}\right), \\
\mathbb{P}\left(\left|\frac{1}{N}\boldsymbol{1}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{1}-1\right|\ge t\right) &\le 2\exp\left(-cd\min\left\{\frac{t^2}{K^2},\frac{t}{K}\right\}\right).
\end{aligned}
$$

Proof. Based on general Hoeffding's inequality Theorem 2.6.3 in (Vershynin, 2018), we can get

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{N}}\mathbf{1}^\top \boldsymbol{Z}\boldsymbol{\mu}\right| \leq t\right) = \mathbb{P}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\boldsymbol{z}_i^\top \boldsymbol{\mu}\right| \leq t\right) \leq 1 - 2\exp\left(-ct^2 d\right).$$

Similarly, by Bernstein's inequality Theorem 2.8.2 in (Vershynin, 2018), we have

$$\mathbb{P}\left(\left|\frac{1}{N}\mathbf{1}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \boldsymbol{y}\right| \leq t\right) \geq 1 - 2\exp\left(-cd\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right),$$

$$\mathbb{P}\left(\left|\frac{1}{N}\mathbf{1}^\top \boldsymbol{Z}\boldsymbol{Z}^\top \mathbf{1} - 1\right| \leq t\right) \geq 1 - 2\exp\left(-cd\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right),$$

where $K = \|\xi\|_{\psi_2}^2$ for $\xi \sim \mathcal{N}(0,1)$. □

Lemma E.2 (Theorem 3.1 in (Dumitriu & Wang, 2023))**.** *Let $G = ([N], E)$ be an inhomogeneous Erdős-Rényi graph associated with the probability matrix $\boldsymbol{P}$, that is, each edge $e = \{i,j\} \subset [N]^2$ is sampled from $\mathrm{Ber}(P_{ij})$, namely, $\mathbb{P}(A_{ij} = 1) = P_{ij}$. Let $\boldsymbol{A}$ denote the adjacency matrix of G. Denote $P_{\max} := \max_{i,j\in[N]} P_{ij}$. Suppose that*

$$N \cdot P_{\max} \geq c\log N\,,$$

for some positive constant c, then with probability at least $1 - 2n^{-10} - 2e^{-N}$, adjacency matrix $\boldsymbol{A}$ satisfies

$$\|\boldsymbol{A} - \mathbb{E}\boldsymbol{A}\| \leq \mathrm{C}_{(\mathrm{E}.1)} \cdot \sqrt{N \cdot P_{\max}}\,, \tag{E.1}$$

Lemma E.3 (Bernstein's inequality, Theorem 2.8.4 of (Vershynin, 2018))**.** *Let $X_1, \ldots, X_n$ be independent mean-zero random variables such that $|X_i| \leq K$ for all i. Let $\sigma^2 = \sum_{i=1}^{n}\mathbb{E}X_i^2$. Then for every $t \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$