A Theoretical Case Study of Structured Variational Inference for Community Detection

Mingzhang Yin
University of Texas at Austin

Y. X. Rachel Wang
University of Sydney

Purnamrita Sarkar
University of Texas at Austin

Abstract

Mean-field variational inference (MFVI) has been widely applied in large scale Bayesian inference. However, MFVI assumes independent distribution on the latent variables, which often leads to objective functions with many local optima, making optimization algorithms sensitive to initialization. In this paper, we study the advantage of structured variational inference in the context of a simple two-class Stochastic Blockmodel. To facilitate theoretical analysis, the variational distribution is constructed to have a simple pairwise dependency structure on the nodes of the network. We prove that, in a broad density regime and for general random initializations, unlike MFVI, the estimated class labels by structured VI converge to the ground truth with high probability, when the model parameters are known, estimated within a reasonable range or jointly optimized with the variational parameters. In addition, empirically we demonstrate structured VI is more robust compared with MFVI when the graph is sparse and the signal to noise ratio is low. The paper takes a first step towards quantifying the role of added dependency structure in variational inference for community detection.

1 Introduction

Variational inference (VI) is a widely used technique for approximating complex likelihood functions in Bayesian learning (Jordan et al., 1999, Blei et al., 2003, Jaakkola and Jordon, 1999), and is known for its computational scalability. VI reduces an intractable posterior inference problem to an optimization framework by imposing simpler dependence structure and is considered a popular alternative to Markov chain Monte Carlo (MCMC) methods. Similar to the Expectation Maximization (EM) algorithm (Dempster et al., 1977), VI works by the basic principle of constructing a tractable lower bound on the complete log-likelihood of a probabilistic model. One of the simplest forms of approximation is mean-field variational inference (MFVI), where the variational lower bound, also known as ELBO, is computed using the expectation with respect to a product distribution over the latent variables(Blei et al., 2003, 2006, Hoffman et al., 2013). Though VI has achieved great empirical success in probabilistic models, theoretical understanding of its convergence properties is still an open area of research.

Theoretical studies of variational methods (and similar algorithms that involve iteratively maximizing a lower bound) have drawn significant attention recently (see (Balakrishnan et al., 2017, Xu et al., 2016, Yan et al., 2017, Yi et al., 2014, Kwon and Caramanis, 2018) for convergence properties of EM). For VI, the global optimizer of the variational lower bound is shown to be asymptotically consistent for a number of models including Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Gaussian mixture models (Pati et al., 2017). In (Westling and McCormick, 2015) the connection between VI estimates and profile M-estimation is explored and asymptotic consistency is established. In practice, however, it is well known the algorithm is not guaranteed to reach the global optimum and the performance of VI often suffers from local optima (Blei et al., 2017). While in some models, convergence to the global optimum can be achieved with appropriate initialization (Wang et al., 2006, Awasthi and Risteski, 2015), understanding convergence with general initialization and the influence of local optima is less studied with a few exceptions (Xu et al., 2016, Ghorbani et al., 2018, Mukherjee et al., 2018).

In general, despite being computationally scalable, MFVI suffers from many stability issues including symmetry-breaking, multiple local optima, and sen-

Proceedings of the 23^{rd}International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

sitivity to initialization, which are consequences of the non-convexity of typical mean-field problems (Wainwright et al., 2008, Jaakkola, 2001). The independence assumption on latent variables also leads to the underestimation of posterior uncertainty (Blei et al., 2017, Yin and Zhou, 2018). To address these problems, many studies suggest that modeling the latent dependency structure can expand the variational family under consideration and lead to larger ELBO and more stable convergence (Xing et al., 2002, Hoffman and Blei, 2015, Giordano et al., 2015, Tran et al., 2015, Ranganath et al., 2016, Yin and Zhou, 2018, Rezende and Mohamed, 2015, Tran et al., 2017). However, rigorous theoretical analysis with convergence guarantees in this setting remains largely underexplored.

In this paper, we aim to study the effect of added dependency structure in a MFVI framework. Since the behavior of MFVI is well understood for the very simple two class, equal sized Stochastic Blockmodel (SBM) (Mukherjee et al., 2018, Zhang and Zhou, 2017), we propose to add a simple pairwise link structure to MFVI in the context of inference for SBMs. We study how added dependency structure can improve MFVI. In particular, we focus on how random initialization behave for VI with added structure.

The stochastic blockmodel (SBM) (Holland et al., 1983) is a widely used network model for community detection in networks. There are a plethora of algorithms with theoretical guarantees for estimation for SBMs like Spectral methods (Rohe et al., 2011, Coja-Oghlan, 2010), semidefinite relaxation based methods (Guédon and Vershynin, 2016, Perry and Wein, 2017, Amini et al., 2018), likelihood-based methods (Amini et al., 2013), modularity based methods (Snijders and Nowicki, 1997, Newman and Girvan, 2004, Bickel and Chen, 2009). Among these, likelihood-based methods remain important and relevant due to their flexibility in incorporating additional model structures. Examples include mixed membership SBM (Airoldi et al., 2008), networks with node covariates (Razaee et al., 2019), and dynamic networks (Matias and Miele, 2017). Among likelihood based methods, VI provides a tractable approximation to the log-likelihood and is a scalable alternative to more expensive methods like Profile Likelihood (Bickel and Chen, 2009), or MCMC based methods (Snijders and Nowicki, 1997, Newman and Girvan, 2004). Computationally, VI was also shown to scale up well to very large graphs (Gopalan and Blei, 2013).

On the theoretical front, (Bickel et al., 2013) proved that the global optimum of MFVI behaves optimally in the dense degree regime. In terms of algorithm convergence, (Zhang and Zhou, 2017) showed the batch coordinate ascent algorithm (BCAVI) for optimizing the mean-field objective has guaranteed convergence if the initialization is sufficiently close to the ground truth. (Mukherjee et al., 2018) fully characterized the optimization landscape and convergence regions of BCAVI for a simple two-class SBM with random initializations. It is shown that uninformative initializations can indeed converge to suboptimal local optima, demonstrating the limitations of the MFVI objective function.

Coming back to structured variational inference, it is important to note that, if one added dependencies between the posterior of each node, the natural approximate inference method is the belief propagation (BP) algorithm (Pearl, 1982, 2014, Wilinski et al., 2019). Based on empirical evidence, it has been conjectured in (Decelle et al., 2011a) that BP is asymptotically optimal for a simple two-class SBM. In the sparse setting where phase transition occurs, (Mossel et al., 2016) analyzed a local variant of BP and showed it is optimal given a specific initialization. In other parameter regions, rigorous theoretical understanding of BP, in particular, how adding dependence structure can improve convergence with general initializations is still an open problem.

Motivated by the above observations, we present a theoretical case study of structured variational inference for SBM. We emphasize here that our primary contribution *does not* lie in proposing a new estimation algorithm that outperforms state-of-the-art methods; rather we use this algorithm as an example to understand the interplay between a non-convex objective function and an iterative optimization algorithm with respect to random initializations, and compare it with MFVI. We consider a two-class SBM with equal class size, an assumption commonly used in theoretical work (Mossel et al., 2016, Mukherjee et al., 2018) where the analysis for the simplest case is nontrivial.

We study structured VI by introducing a simple pairwise dependence structure between randomly paired nodes. By carefully bounding the variational parameters in each iteration using a combination of concentration and Littlewood-Offord type anti-concentration arguments (Erdös, 1945), we prove that in a broad density regime and under a fairly general random initialization scheme, the Variational Inference algorithm with Pairwise Structure (VIPS) can converge to the ground truth with probability tending to one, when the parameters are known, estimated within a reasonable range, or updated appropriately (Section 3). This is in contrast to MFVI, where convergence only happens for a narrower range of initializations. In addition, VIPS can escape from certain local optima that exist in the MFVI objective. These results highlight the theoretical advantage of the added dependence structure. Empirically, we demonstrate that VIPS is more robust compared to MFVI when the graph is sparse and the signal to noise ratio is low (Section 4). We observe similar trends hold in more general models with unbalanced class sizes and more than two classes. We hope that our analysis for the simple blockmodel setting can shed light on theoretical analysis of algorithms with more general dependence structure such as BP.

The paper is organized as follows. Section 2 contains the model definition and introduces VIPS. We present our theoretical results in Section 3. Finally in Section 4, we demonstrate the empirical performance of VIPS in contrast to MFVI and other algorithms. We conclude with a discussion on possible generalizations, accompanied by promising empirical results in Section 5.

2 Preliminaries and Proposed Work

2.1 Preliminaries

The stochastic block model (SBM) is a generative network model with community structure. A K-community SBM for n nodes is generated as follows: each node is assigned to one of the communities in $\{1,\dots,K\}$. These memberships are represented by $U \in \{0,1\}^{n\times K}$, where each row follows an independent Multinomial $(1;\pi)$ distribution with parameter π. We have $U_{ik}=1$ if node i belongs to community k and $\sum_{k=1}^{K} U_{ik}=1$. Given the community memberships, links between pairs of nodes are generated according to the entries in a $K\times K$ connectivity matrix B. That is, if A denotes the $n\times n$ binary symmetric adjacency matrix, then, for $i \neq j$,

$$P(A_{ij}=1|U_{ik}=1,U_{j\ell}=1)=B_{k\ell}. \tag{1}$$

We consider undirected networks, where both B and A are symmetric. Given an observed A, the goal is to infer the latent community labels U and the model parameters (π, B). Since the data likelihood $P(A;B,\pi)$ requires summing over K^n possible labels, approximations such as MFVI are often needed to produce computationally tractable algorithms.

Throughout the rest of the paper, we will use $\mathbf{1}_n$ to denote the all-one vector of length n. When it is clear from the context, we will drop the subscript n. Let I be the identity matrix and $J=\mathbf{1}\mathbf{1}^T$. $\mathbf{1}_C$ denotes a vector where the i-th element is 1 if $i \in C$ and 0 otherwise, where C is some index set. Similar to (Mukherjee et al., 2018), we consider a two-class SBM with equal class size, where $K=2$, $\pi=1/2$, and B takes the form $B_{11}=B_{22}=p$, $B_{12}=B_{21}=q$, with $p>q$. We denote the two true underlying communities by G_1 and G_2, where G_1, G_2 form a partition of $\{1,2,\dots,n\}$ and $|G_1|=|G_2|$. (For convenience, we assume n is even.) As will become clear, the full analysis of structured VI in this simple case is highly nontrivial.

2.2 Variational inference with pairwise structure (VIPS)

The well-known MFVI approximates the likelihood by assuming a product distribution over the latent variables. In other words, the posterior label distribution of the nodes is assumed to be independent in the variational distribution. To investigate how introducing dependence structure can help with the inference, we focus on a simple setting of linked pairs which are independent of each other. To be concrete, we randomly partition the n nodes into two sets: $P_1=\{z_1,\cdots,z_m\}$, $P_2=\{y_1,\cdots,y_m\}$, with $m=n/2$. Here $z_k, y_k \in \{1,\dots,n\}$ are the node indices. In our structured variational distribution, we label pairs of nodes (z_k, y_k) using index $k \in \{1,\dots,m\}$ and assume there is dependence within each pair. The corresponding membership matrices for P_1 and P_2 are denoted by Z and Y respectively, which are both $m\times 2$ sub-matrices of the full membership matrix U. More explicitly, the k^{th} row of matrix Z encodes the membership of node z_k in P_1, and similarly for Y. For convenience, we permute both the rows and columns of A based on the node ordering in P_1 followed by that in P_2 to create a partitioned matrix: $A=\left[\begin{array}{c|c} A^{zz} & A^{zy} \\ \hline A^{yz} & A^{yy}\end{array}\right]$, where each block is an $m\times m$ matrix. Given the latent membership variable (Z,Y), by Eq. (1) the likelihood of A is given by

$$\begin{aligned}
&P(A^{zz}_{ij}|Z,B)=\textstyle\prod_{a,b}[B_{ab}^{A^{zz}_{ij}}(1-B_{ab})^{1-A^{zz}_{ij}}]^{Z_{ia}Z_{jb}}\\
&P(A^{zy}_{ij}|Y,Z,B)=\textstyle\prod_{a,b}[B_{ab}^{A^{zy}_{ij}}(1-B_{ab})^{1-A^{zy}_{ij}}]^{Z_{ia}Y_{jb}}\\
&P(A^{yy}_{ij}|Y,B)=\textstyle\prod_{a,b}[B_{ab}^{A^{yy}_{ij}}(1-B_{ab})^{1-A^{yy}_{ij}}]^{Y_{ia}Y_{jb}}
\end{aligned} \tag{2}$$

where $a,b \in \{1,2\}$ and $A^{zy}=(A^{yz})^T$.

A simple illustration of the partition and how ordered pairs of nodes are linked to incorporate dependence is given in Figure 1, where the the true underlying communities G_1 and G_2 are shaded differently. After the partition, we have m pairs of linked nodes indexed from 1 to m. For convenience of analysis, we define the following sets for these pairs of linked nodes, as illustrated in Figure 1.

Define C_1, (C_1') as the set of indices i of pairs (z_i, y_i) with $z_i \in G_1$, $(y_i \in G_1)$. Similarly, C_2, (C_2') is the set of indices of pairs (z_i, y_i) with $z_i \in G_2$, $(y_i \in G_2)$. We will also make use of the sets $C_{ab} := C_a \cap C_b'$, where $a,b \in \{1,2\}$. In Figure 1, these sets correspond to different combinations of shading, i.e. community memberships, of the linked pairs, e.g. C_{12} is the index set of pairs (z_i, y_i) with $z_i \in G_1, y_i \in G_2$.

We define the variational distribution for the latent membership matrix (Z, Y) as $Q(Z, Y)$, which we assume takes the form

$$Q(Z, Y) = \prod_{i=1}^{m} Q(Z_i, Y_i), \tag{3}$$

where Z_i denotes the i^{th} row of Z, and $Q(Z_i, Y_i)$ is a general categorical distribution with variational parameters defined as follows.

$$\psi_i^{cd} := Q(Z_{i,c+1} = 1, Y_{i,d+1} = 1),$$

for $i \in \{1, \ldots, m\}, c, d \in \{0, 1\}$. This allows one to encode more dependence structure between the posteriors at different nodes than vanilla MFVI, since we allow for dependence within each linked pair of nodes while keeping independence between different pairs. We define the marginal probabilities as:

$$\begin{aligned} \phi_i &:= Q(Z_{i1} = 1) = \psi_i^{10} + \psi_i^{11} \\ \xi_i &:= Q(Y_{i1} = 1) = \psi_i^{01} + \psi_i^{11}. \end{aligned} \tag{4}$$

Next we derive the ELBO on the data log-likelihood $\log P(A)$ using $Q(Z, Y)$. For pairwise structured variational inference (VIPS), ELBO takes the form

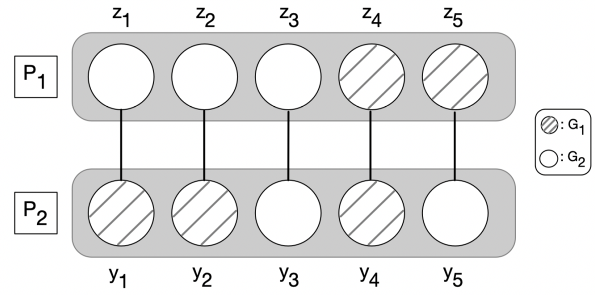


Figure 1: An illustration of a partition for $n = 10$. The shaded nodes belong to community G_1 and unshaded nodes belong to community G_2. The nodes are randomly partitioned into two sets P_1 and P_2, and pairs of nodes are linked from index 1 to m. Dependence structure within each linked pair is incorporated into the variational distribution $Q(Z, Y)$. For this partition and pair linking, $C_1 = \{4, 5\}$, $C_2 = \{1, 2, 3\}$, $C_1' = \{1, 2, 4\}$, $C_2' = \{3, 5\}$; $C_{11} = \{4\}$, $C_{12} = \{5\}$, $C_{21} = \{1, 2\}$, $C_{22} = \{3\}$.

$$\mathcal{L}(Q; \pi, B) = \mathbb{E}_{Z,Y \sim Q(Z,Y)} \log P(A|Z, Y) - \mathrm{KL}(Q||P),$$

where $P(Z, Y)$ is the probability of community labels from SBM and follows independent Bernoulli (π) distribution, $\mathrm{KL}(\cdot||\cdot)$ denotes the usual Kullback–Leibler divergence between two distributions. Using the likelihood in Eq. (2), the ELBO becomes

$$\begin{aligned} \mathcal{L}(Q; \pi, B) = \frac{1}{2}\mathbb{E}_Q \Bigg\{ &\sum_{i \neq j, a, b} [Z_{ia} Z_{jb} (A_{ij}^{zz} \alpha_{ab} + f(\alpha_{ab})) \\ &+ \frac{1}{2} Y_{ia} Y_{jb} (A_{ij}^{yy} \alpha_{ab} + f(\alpha_{ab})) + Z_{ia} Y_{jb} (A_{ij}^{zy} \alpha_{ab} + f(\alpha_{ab}))] \\ &+ \sum_{i,a,b} Z_{ia} Y_{ib} (A_{ii}^{zy} \alpha_{ab} + f(\alpha_{ab})) \Bigg\} \\ &- \sum_{i=1}^{m} \mathrm{KL}(Q(z_i, y_i)||P(z_i)P(y_i)), \end{aligned} \tag{5}$$

where $\alpha_{ab} = \log(B_{ab}/(1 - B_{ab}))$ and $f(\alpha) = -\log(1 + e^{\alpha})$. The KL regularization term can be computed as

$$\begin{aligned} &\mathrm{KL}(Q(z_i, y_i)||P(z_i)P(y_i)) \\ =&\textstyle\sum_{0 \leq c, d \leq 1} \psi_i^{cd} \log(\psi_i^{cd})/(\pi^c \pi^d (1 - \pi)^{1-c} (1 - \pi)^{1-d}). \end{aligned}$$

Our goal is to maximize $\mathcal{L}(Q; \pi, B)$ with respect to the variational parameters ψ_i^{cd} for $1 \leq i \leq m$. Since $\sum_{c,d} \psi_i^{cd} = 1$ for each i, it suffices to consider ψ_i^{10}, ψ_i^{01} and ψ_i^{11}. By taking derivatives, we can derive a batch coordinate ascent algorithm for updating $\psi^{cd} = (\psi_1^{cd}, \ldots, \psi_m^{cd})$. Detailed calculation of the derivatives can be found in Section A of the Appendix. Recall that $\pi = \frac{1}{2}$. Also, define

$$t := \frac{1}{2} \log \frac{p/(1-p)}{q/(1-q)} \qquad \lambda := \frac{1}{2t} \log \frac{1-q}{1-p}, \tag{6}$$

$$\theta^{cd} := \log \frac{\psi^{cd}}{1 - \psi^{01} - \psi^{10} - \psi^{11}}, \tag{7}$$

where θ^{cd} are logits, $c, d \in \{0, 1\}$ and all the operations are defined *element-wise.*

Given the model parameters p, q, the current values of ψ^{cd} and the marginals $\phi = \psi^{10} + \psi^{11}$, $\xi = \psi^{01} + \psi^{11}$ as defined in Eq. (4), the updates for θ^{cd} are given by:

$$\begin{aligned} \theta^{10} =& 4t[A^{zz} - \lambda(J - I)](\phi - \tfrac{1}{2}\mathbf{1}_m) \\ &+ 4t[A^{zy} - \lambda(J - I) - \mathrm{diag}(A^{zy})](\xi - \frac{1}{2}\mathbf{1}_m) \\ &- 2t(\mathrm{diag}(A^{zy}) - \lambda I)\mathbf{1}_m, \end{aligned} \tag{8}$$

$$\begin{aligned} \theta^{01} =& 4t[A^{yy} - \lambda(J - I)](\xi - \tfrac{1}{2}\mathbf{1}_m) \\ &+ 4t[A^{yz} - \lambda(J - I) - \mathrm{diag}(A^{yz})](\phi - \frac{1}{2}\mathbf{1}_m) \\ &- 2t(\mathrm{diag}(A^{yz}) - \lambda I)\mathbf{1}_m, \end{aligned} \tag{9}$$

$$\begin{aligned} \theta^{11} =& 4t[A^{zz} - \lambda(J - I)](\phi - \tfrac{1}{2}\mathbf{1}_m) \\ &+ 4t[A^{zy} - \lambda(J - I) - \mathrm{diag}(A^{zy})](\xi - \frac{1}{2}\mathbf{1}_m) \\ &+ 4t[A^{yz} - \lambda(J - I) - \mathrm{diag}(A^{yz})](\phi - \frac{1}{2}\mathbf{1}_m) \\ &+ 4t[A^{yy} - \lambda(J - I)](\xi - \frac{1}{2}\mathbf{1}_m). \end{aligned} \tag{10}$$

Given θ^{cd}, we can update the current values of ψ^{cd} and the corresponding marginal probabilities ϕ, ξ using element-wise operations as follows:

$$\psi^{cd} = \frac{e^{\theta^{cd}}}{1 + e^{\theta^{01}} + e^{\theta^{11}} + e^{\theta^{10}}}, \quad u := (\phi, \xi)$$

$$\phi = \frac{e^{\theta^{10}} + e^{\theta^{11}}}{1 + e^{\theta^{10}} + e^{\theta^{01}} + e^{\theta^{11}}}, \quad \xi = \frac{e^{\theta^{01}} + e^{\theta^{11}}}{1 + e^{\theta^{10}} + e^{\theta^{01}} + e^{\theta^{11}}} \tag{11}$$

where $(c, d) = (1, 0), (0, 1), (1, 1)$. The marginal probabilities are concatenated as $u = (\phi, \xi) \in [0, 1]^n$. Thus u can be interpreted as the estimated posterior membership probability of all the nodes.

Since θ^{cd} determines ψ^{cd} in the categorical distribution and u represents the corresponding marginals, one can think of θ^{cd} and u as the local and global parameters respectively. It has been empirically shown that the structured variational methods can achieve better convergence property by iteratively updating the local and global parameters (Blei et al., 2003, Hoffman et al., 2013, Hoffman and Blei, 2015). In the same spirit, we update the parameters θ^{cd} and u iteratively by (8)–(11), following the order

$$\theta^{10} \to u \to \theta^{01} \to u \to \theta^{11} \to u \to \theta^{10} \cdots . \tag{12}$$

We call a full update of all the parameters $\theta^{10}, \theta^{01}, \theta^{11}, u$ in (12) as one *meta iteration* which consists of three inner iterations of u updates. We use $u_j^{(k)}$ ($j = 1, 2, 3$) to denote the update in the j-th iteration of the k-th meta iteration, and $u^{(0)}$ to denote the initialization. Algorithm 1 gives the full algorithm when the model parameters are known.

Algorithm 1 Variational Inference with Pairwise Structure (VIPS)

```
input  : Adjacency matrix $A \in \{0,1\}^{n \times n}$, model parameter $p, q, \pi = 1/2$.
output : The estimated node membership vector $u$.
Initialize the elements of $u$ i.i.d. from an arbitrary distribution $f_\mu$ defined on $[0,1]$ with mean $\mu$. Initialize $\theta^{10} = \theta^{01} = \theta^{11} = \mathbf{0}$;
Randomly select $n/2$ nodes as $P_1$ and the other $n/2$ nodes as $P_2$;
while not converged do
    Update $\theta^{10}$ by (8); update $u = (\phi, \xi)$ by (11).
    Update $\theta^{01}$ by (9); update $u = (\phi, \xi)$ by (11).
    Update $\theta^{11}$ by (10); update $u = (\phi, \xi)$ by (11).
end
```

Remark 1. *So far we have derived the updates and described the optimization algorithm when the true parameters p, q are known. When they are unknown, they can be updated jointly with the variational parameters after each meta iteration as*

$$p = \frac{\begin{array}{c}(\mathbf{1}_n - u)^T A (\mathbf{1}_n - u) + u^T A u \\ + 2(\mathbf{1}_m - \psi^{10} - \psi^{01})^T diag(A^{zy}) \mathbf{1}_m\end{array}}{\begin{array}{c}(\mathbf{1}_n - u)^T (J - I)(\mathbf{1}_n - u) \\ + u^T (J - I) u + 2(\mathbf{1}_m - \psi^{10} - \psi^{01})^T \mathbf{1}_m\end{array}}$$

$$q = \frac{(\mathbf{1}_n - u)^T A u + (\psi^{10} + \psi^{01})^T diag(A^{zy}) \mathbf{1}_m}{(\mathbf{1}_n - u)^T (J - I) u_n + (\psi^{10} + \psi^{01})^T \mathbf{1}_m} \tag{13}$$

Although it is typical to update p, q and u jointly, as shown in (Mukherjee et al., 2018), analyzing MFVI updates with known parameters can shed light on the convergence behavior of the algorithm. Initializing u randomly while jointly updating p, q always leads MFVI to an uninformative local optima. For this reason, in what follows we will analyze Algorithm 1 in the context of both fixed and updating parameters p, q.

3 Main results

In this section, we present theoretical analysis of the algorithm in three settings: (i) When the parameters are set to the true model parameters p, q; (ii) When the parameters are not too far from the true values, and are held fixed throughout the updates; (iii) Starting from some reasonable guesses of the parameters, they are jointly updated with latent membership estimates.

In the following analysis, we will frequently use the eigen-decomposition of the expected adjacency matrix $P = \mathbb{E}[A|U] = \frac{p+q}{2} \mathbf{1}_n \mathbf{1}_n^T + \frac{p-q}{2} v_2 v_2^T - pI$ where $v_2 = (v_{21}, v_{22})^T = (\mathbf{1}_{C_1} - \mathbf{1}_{C_2}, \mathbf{1}_{C_1'} - \mathbf{1}_{C_2'})^T$ is the second eigenvector. Since the second eigenvector is just a shifted and scaled version of the membership vector, the projection $|\langle u, v_2 \rangle|$ is equivalent to the ℓ_1 error from true label z^* (up-to label permutation) by $\|u - z^*\|_1 = m - |\langle u, v_2 \rangle|$. We consider the parametrization $p \asymp q \asymp \rho_n$, where the density $\rho_n \to 0$ at some rate and $p - q = \Omega(\rho_n)$.

When the true parameters p, q are known, without dependency structure, MFVI with random initializations converges to the stationary points with non-negligible probability (Sarkar et al., 2019). When the variational distribution has a simple pairwise dependency structure as VIPS, we show a stronger result. To be concrete, in this setting, we establish that convergence happens with probability approaching 1. In addition, unlike MFVI, the convergence holds for general random initializations. We will first consider the situation when $u^{(0)}$ is initialized from a distribution centered at $\mu = \frac{1}{2}$ and show the results for $\mu \neq \frac{1}{2}$ in Corollary 1.

Theorem 1 (Sample behavior for known parameters)**.** *Assume $\theta^{10}, \theta^{01}, \theta^{11}$ are initialized as $\mathbf{0}$ and the elements of $u^{(0)} = (\phi^{(0)}, \xi^{(0)})$ are initialized i.i.d. from Bernoulli($\frac{1}{2}$). When $p \asymp q \asymp \rho_n$, $p - q = \Omega(\rho_n)$, and $\sqrt{n}\rho_n = \Omega(\log(n))$, Algorithm 1 converges to the true labels asymptotically after the second meta iteration, in the sense that*

$$\|u_3^{(2)} - z^*\|_1 = n \exp(-\Omega_P(n\rho_n)))$$

z^* are the true labels with $z^* = \mathbf{1}_{G_1}$ or $\mathbf{1}_{G_2}$. The same convergence holds for all the later iterations.

Remark 2. *It is important to note that there are many algorithms (see (Abbe, 2017) for a survey) which re-*

cover the memberships exactly in this regime. We do not compare our theoretical results with those or to well known thresholds for exact recovery (Abbe et al., 2015), because our goal is not to design a new algorithm with an improved theoretical guarantees. Rather, we show that by introducing the simplest possible pairwise dependence structure, variational inference for a simple setting of a SBM improves over MFVI which has no such structure. The density regime simply makes the analysis somewhat easier.

Proof. We provide a proof sketch here and defer the details to Section B of the Appendix. We assume for the first six iterations, we randomly partition A into six $A^{(i)}, i = 0, \ldots, 5$ by assigning each edge to one of the six subgraphs with equal probability. For the later iterations, we can use the whole graph A. Then $A^{(i)}$'s are independent with population matrix $P/6$. Although not used in Algorithm 1, the graph splitting is a widely used technique for theoretical convenience (McSherry, 2001, Chaudhuri et al., 2012) and allows us to bound the noise in each iteration more easily. The main arguments involve lower bounding the size of the projection $|\langle u, v_2\rangle|$ in each iteration as it increases towards $n/2$, at which point the algorithm achieves strong consistency. For ease of exposition, we will scale everything by 6 so that p, q, λ correspond to the parameters for the full un-split matrix P. This does not affect the analysis in any way.

In each iteration, we decompose the intermediate $\theta^{10}, \theta^{01}, \theta^{11}$ into blockwise constant signal and random noise using the spectral property of the population matrix P. As an illustration, in the first meta iteration, we write the update in (8)–(10) as signal plus noise,

$$\begin{aligned}\theta_i^{10} &= 4t(s_1\mathbf{1}_{C_1} + s_2\mathbf{1}_{C_2} + r_i^{(0)})\\ \theta_i^{01} &= 4t(x_1\mathbf{1}_{C_1'} + x_2\mathbf{1}_{C_2'} + r_i^{(1)})\\ \theta_i^{11} &= 4t(y_1\mathbf{1}_{C_1} + y_2\mathbf{1}_{C_2} + y_1\mathbf{1}_{C_1'} + y_2\mathbf{1}_{C_2'} + r_i^{(2)})\end{aligned}$$

where t is a constant and the noise has the form

$$r^{(i)} = R^{(i)}(u_j^{(k)} - \frac{1}{2}\mathbf{1}) \tag{14}$$

for appropriate j, k, where $R^{(i)}$ arises from the sample noise in the adjacency matrix. We handle the noise from the first iteration $r^{(0)}$ with a Berry-Esseen bound conditional on $u^{(0)}$, and the later $r^{(i)}$ with a uniform bound. The blockwise constant signals s_1, x_1, y_1 are updated as $(\frac{p+q}{2} - \lambda)(\langle u, \mathbf{1}_n\rangle - m) + (\frac{p-q}{2})\langle u, v_2\rangle$ and s_2, x_2, y_2 are updated as $(\frac{p+q}{2} - \lambda)(\langle u, \mathbf{1}_n\rangle - m) - (\frac{p-q}{2})\langle u, v_2\rangle$. As $\langle u, v_2\rangle$ increases throughout the iterations, the signals become increasingly separated for the two communities. Using Littlewood-Offord type anti-concentration, we show in the first meta iteration,

$$\begin{aligned}&\langle u_1^{(1)}, v_2\rangle = \Omega_P(n\sqrt{\rho_n}), \quad \langle u_1^{(1)}, \mathbf{1}\rangle - m = 0\\ &\langle u_2^{(1)}, v_2\rangle \geq \frac{n}{8} - o_P(n), \quad \langle u_2^{(1)}, \mathbf{1}\rangle - m = 0\\ &\langle u_3^{(1)}, v_2\rangle \geq \frac{1}{4}n + o_P(n),\\ &-\frac{n}{8} - o_P(n) \leq \langle u_3^{(1)}, \mathbf{1}\rangle - m \leq \frac{n}{4} + o_P(n)\end{aligned} \tag{15}$$

After the second meta iteration we have

$$\begin{gathered}s_1^{(2)}, x_1^{(2)}, y_1^{(2)} = \Omega_P(n\rho_n),\ s_2^{(2)}, x_2^{(2)}, y_2^{(2)} = -\Omega_P(n\rho_n);\\ 2y_1^{(2)} - s_1^{(2)} = \Omega_P(n\rho_n),\ 2y_1^{(2)} - x_1^{(2)} = \Omega_P(n\rho_n);\\ s_1^{(2)} - (y_1^{(2)} + y_2^{(2)}) = \Omega_P(n\rho_n),\ x_1^{(2)} - (y_1^{(2)} + y_2^{(2)}) = \Omega_P(n\rho_n).\end{gathered}$$

Plugging equations above to (11), we have the desired convergence after the second meta iteration. □

The next corollary shows the same convergence holds when we use a general random initialization not centered at 1/2. In contrast, MFVI converges to stationary points $\mathbf{0}_n$ or $\mathbf{1}_n$ with such initializations.

Corollary 1. *Assume the elements of $u^{(0)}$ are i.i.d. sampled from a distribution with mean $\mu \neq 0.5$. Under the conditions in Theorem 1, applying Algorithm 1 with known p, q, we have $\|u_1^{(3)} - z^*\|_1 = n\exp(-\Omega_P(n\rho_n)))$. The same order holds for all the later iterations.*

The proof relies on showing after the first iteration, $u_1^{(1)}$ behaves like nearly independent Bernoulli($\frac{1}{2}$), the details of which can be found in Appendix B.

The next proposition focuses on the behavior of special points in the optimization space for u. In particular, we show that Algorithm 1 enables us to move away from the stationary points $\mathbf{0}_n$ and $\mathbf{1}_n$, whereas in MFVI, the optimization algorithm gets trapped in these stationary points (Mukherjee et al., 2018).

Proposition 1 (Escaping from stationary points)**.**

(i) $(\psi^{00}, \psi^{01}, \psi^{10}, \psi^{11}) = (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0})$, $(\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1})$ *(these vectors are m-dimensional) are the stationary points of the pairwise structured ELBO when p, q are known, which maps to $u = \mathbf{0}_n$ and $\mathbf{1}_n$ respectively.*

(ii) With the updates in Algorithm 1, when $u^{(0)} = \mathbf{0}_n$, $\mathbf{1}_n$, VIPS converges to the true labels with $\|u_1^{(3)} - z^*\|_1 = n\exp(-\Omega_P(n\rho_n)))$.

The above results requires knowing the true p and q. The next corollary shows that, even if we do not have access to the true parameters, as long as some reasonable estimates can be obtained, the same convergence as in Theorem 1 holds thus demonstrating robustness to misspecified parameters. Here we hold the parameters fixed and only update u as in Algorithm 1. When $\hat{p}, \hat{q} \asymp \rho_n$, we need $\hat{p} - \hat{q} = \Omega(\rho_n)$ and $\hat{p}, \hat{q}$ not too far from the true values to achieve convergence. The proof is deferred to the Appendix.

Proposition 2 (Parameter robustness)**.** *If we replace true p, q with some estimation $\hat{p}, \hat{q}$ in Algorithm 1, the same conclusion as in Theorem 1 holds if*

$$1.\ \frac{p+q}{2} > \hat{\lambda}, \qquad 2.\ \hat{\lambda} - q = \Omega(\rho_n), \qquad 3.\ \hat{t} = \Omega(1).$$

where $\hat{t} = \frac{1}{2}\log\dfrac{\hat{p}/(1-\hat{p})}{\hat{q}/(1-\hat{q})}$, $\hat{\lambda} = \frac{1}{2\hat{t}}\log\dfrac{1-\hat{q}}{1-\hat{p}}$.

Finally, we consider updating the parameters jointly with u (as explained in Remark 1) by first initializing the algorithm with some reasonable $p^{(0)}, q^{(0)}$.

Theorem 2 (Updating parameters and u simultaneously)**.** *Suppose we initialize with some estimates of true (p, q) as $\hat{p} = p^{(0)}$, $\hat{q} = q^{(0)}$ satisfying the conditions in Proposition 2 and apply two meta iterations in Algorithm 1 to update u before updating $\hat{p} = p^{(1)}, \hat{q} = q^{(1)}$. After this, we alternate between updating u and the parameters after each meta iteration. Then*

$$p^{(1)} = p + O_P(\sqrt{\rho_n}/n), \qquad q^{(1)} = q + O_P(\sqrt{\rho_n}/n),$$
$$\|u_3^{(2)} - z^*\|_1 = n\exp(-\Omega(n\rho_n)),$$

and the same holds for all the later iterations.

4 Experiments

In this section, we present some numerical results. In Figures 2 to 4 we show the effectiveness of VIPS in our theoretical setting of two equal sized communities. In Figures 5 (a) and (b) we show that empirically the advantage of VIPS holds even for unbalanced community sizes and $K > 2$. Our goal is two-fold: (i) we demonstrate that the empirical convergence behavior of VIPS coincides well with our theoretical analysis in Section 3; (ii) in practice VIPS has superior performance over MFVI in both the simple setting we have analyzed and more general settings, thus confirming the advantage of the added dependence structure. For the sake of completeness, we also include comparisons with other popular algorithms, even though it is not our goal to show VIPS outperforms these methods.

In Figure 2, we compare the convergence property of VIPS with MFVI for initialization from independent Bernoulli's with means $\mu = 0.1, 0.5$, and 0.9. We randomly generate a graph with $n = 3000$ nodes with parameters $p_0 = 0.2, q_0 = 0.01$ and show results from 20 random trials. We plot $\min(\|u - z^*\|_1, \|u - (\mathbf{1} - z^*)\|_1)$, or the ℓ_1 distance of the estimated label u to the ground truth z^* on the Y axis versus the iteration number on the X axis. In this experiments, both VIPS and MFVI were run with the true p_0, q_0 values. As shown in Figure 2, when $\mu = \frac{1}{2}$, VIPS converges to z^* after two meta iterations (6 iterations) for all the random initializations. In contrast, for MFVI, a fraction of the random initializations converge to $\mathbf{0}_n$ and $\mathbf{1}_n$. When $\mu \neq \frac{1}{2}$, VIPS converges to the ground truth after three meta iterations, whereas MFVI stays at the stationary points $\mathbf{0}_n$ and $\mathbf{1}_n$. This is consistent with our theoretical results in Theorem 1 and Corollary 1, and those in (Mukherjee et al., 2018).

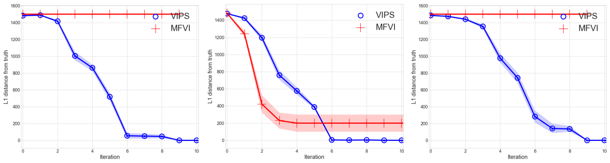
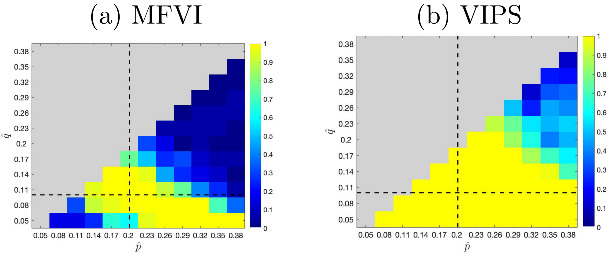


Figure 2: ℓ_1 distance from ground truth (Y axis) vs. number of iterations (X axis). The line is the mean of 20 random trials and the shaded area shows the standard deviation. u is initialized from i.i.d. Bernoulli with mean $\mu = 0.1, 0.5, 0.9$ from the left to right.

In Figure 3, we show when the true p, q are unknown, the dependence structure makes the algorithm more robust to estimation errors in $\hat{p}, \hat{q}$. The heatmap represents the normalized mutual information (NMI) (Romano et al., 2014) between u and z^*, with $\hat{p}$ on the X axis and $\hat{q}$ on the Y axis. We only examine pairs with $\hat{p} > \hat{q}$. Both VIPS and MFVI were run with $\hat{p}$ and $\hat{q}$, which were held fixed and differ from the true values to varying extent. The dashed line represents the true p, q used to generate the graph. For each $\hat{p}, \hat{q}$ pair, the mean NMI for 20 random initializations from i.i.d Bernoulli($\frac{1}{2}$) is shown. VIPS recovers the ground truth in a wider range of $\hat{p}, \hat{q}$ values than MFVI. We show in Section D of the Appendix that similar results also hold for $K = 2$ with unbalanced community sizes.


(a) MFVI

(b) VIPS

Figure 3: NMI averaged over 20 random initializations for each $\hat{p}$, $\hat{q}$ ($\hat{p} > \hat{q}$). The true parameters are $(p_0, q_0) = (0.2, 0.1)$, $\pi = 0.5$ and $n = 2000$. The dashed lines indicate the true parameter values.

In Figure 4, we compare VIPS with MFVI under different network sparsities and signal-to-noise ratios (SNR) as defined by $r_0 = p_0/q_0$. For the sake of completeness, we also include two other popular algorithms, Belief Propagation (BP) (Decelle et al., 2011b) and Spectral

Clustering (Rohe et al., 2011). To meet the conditions in Theorem 2, we started VIPS with $\hat{p}$ equal to the average degree of A, and $\hat{q} = \hat{p}/r_0$. $\hat{p}$ and $\hat{q}$ were updated alternatingly with u according to Eq. (13) after three meta iterations in Algorithm 1, a setting similar to that of Theorem 2.

In Figure 4-(a), the average expected degree is fixed as the SNR p_0/q_0 increases on the X axis, whereas in Figure 4-(b), the SNR is fixed and we vary the average expected degree on the X axis. The results show that VIPS consistently outperforms MFVI, indicating the advantage of the added dependence structure. Note that we plot BP with the model parameters initialized at true (p_0, q_0) , since it is sensitive to initialization setting, and behaves poorly with mis-specified ones. Despite this, VIPS is largely comparable to BP and Spectral Clustering. For average degree 20 (Figure 4-(b)), BP outperforms all other methods, because of the correct parameter setting. This NMI value becomes 0.4 with high variance, if we provide initial $\hat{p}, \hat{q}$ values to match the average degree but $\hat{p}/\hat{q} = 10$. In contrast, VIPS is much more robust to the initial choice of $\hat{p}, \hat{q}$, which we show in Section C of the Appendix.

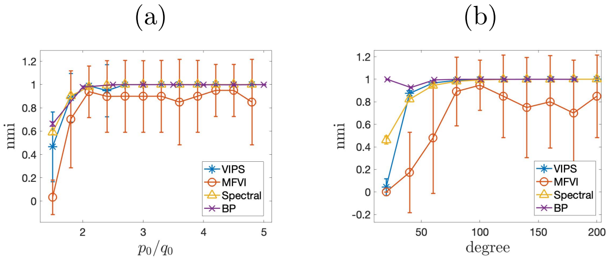


Figure 4: Comparison of NMI under different SNR p_0/q_0 and network degrees. The lines and error bars are means and standard deviations from 20 random trials. (a) Vary p_0/q_0 with degree fixed at 70. (b) Vary the degree with $p_0/q_0 = 2$. In both figures $n = 2000$.

Additional experiments (Appendix, section D) show that VIPS with fixed mis-specified parameters (within reasonable deviation from the truth), fixed true parameters and parameters updated with Eq. (13) converge to the truth when initialized by independent Bernoulli's.

5 Discussion and Generalizations

In this paper, we propose a simple Variational Inference algorithm with Pairwise Structure (VIPS) in a SBM with two equal sized communities. VI has been extensively applied in the latent variable models mainly due to their scalability and flexibility for incorporating changes in model structure. However, theoretical understanding of the convergence properties is limited and mostly restricted to MFVI with fully factorized variational distributions. Theoretically we prove that in a SBM with two equal sized communities, VIPS can converge to the ground truth with probability tending to one for different random initialization schemes and a range of graph densities. In contrast, MFVI only converges for a constant fraction of Bernoulli(1/2) random initializations. We consider settings where the model parameters are known, estimated or appropriately updated as part of the iterative algorithm.

Though our main results are for $K = 2, \pi = 0.5$, we conclude with a discussion on generalizations to unbalanced clusters and SBMs with $K > 2$ equal communities. To apply VIPS for general $K > 2$ clusters, we will have $K^2 - 1$ categorical distribution parameters ψ^{cd} for $c, d \in \{1, 2, \ldots, K\}$ and marginal likelihood $\phi_1, \ldots, \phi_{K-1}, \xi_1, \ldots, \xi_{K-1}$. The updates are similar to Eq. (10) and Eq. (11) and are deferred to the Appendix (section C). Similar to the $K = 2$ case, we update the local and global parameters iteratively. As for the unbalanced case (see Appendix Section C), the updates involve an additional term which is the logit of π. We assume that π is known and fixed.

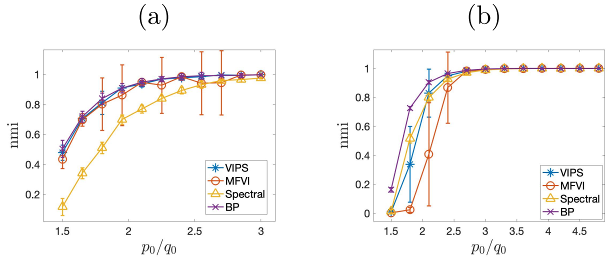


Figure 5: Comparison of VIPS, MFVI, Spectral and BP using error-bars from 20 random trials for $n = 2000$, average degree 50, p_0/q_0 is changed on X axis. (a) $\pi = 0.3$ (b) K =3, $B = (p - q)I + qJ$. For BP, MFVI and VIPS, we use true parameters.

In Figure 5-(a), we show results for unbalanced SBM with $\pi = 0.3$, which is assumed to be known. In Figure 5-(b), similar to the setting in (Mukherjee et al., 2018), we consider a SBM with three equal-sized communities. The parameters are set as $n = 2000$, average degree 50, p_0 and q_0 are changed to get different SNR values and the random initialization is from Dirichlet$(1, 1, 1)$. For a fair comparison of VIPS, MFVI and BP, we use the true p_0, q_0 values in all three algorithms; robustness to parameter specification of VIPS is included in the Appendix C. We see that for the unbalanced setting (Figure 5-(a)) VIPS performs as well as BP and better than Spectral Clustering. For the $K = 3$ setting (Figure 5-(b)) VIPS performs worse than BP and Spectral for very low SNR values, whereas for higher SNR it performs comparably to Spectral and BP, and better than MFVI, which has much higher variance.

Acknowledgement

P.S. was partially funded by NSF grant DMS 1713082.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1): 471–487, 2015.

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of machine learning research*, 9(Sep):1981–2014, 2008.

A. A. Amini, A. Chen, P. J. Bickel, and E. Levina. Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122, 2013.

Arash A Amini, Elizaveta Levina, et al. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

Pranjal Awasthi and Andrej Risteski. On some provably correct cases of variational inference for topic models. In *Advances in Neural Information Processing Systems*, pages 2098–2106, 2015.

Sivaraman Balakrishnan, Martin J Wainwright, Bin Yu, et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

Peter Bickel, David Choi, Xiangyu Chang, Hai Zhang, et al. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943, 2013.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

David M Blei, Michael I Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112 (518):859–877, 2017.

Kamalika Chaudhuri, Fan Chung Graham, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *COLT*, volume 23 of *JMLR Proceedings*, pages 35.1–35.23. JMLR.org, 2012.

Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6): 066106, 2011a.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011b. doi: 10.1103/PhysRevE.84.066106. URL `https://link.aps.org/doi/10.1103/PhysRevE.84.066106`.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

Paul Erdös. On a lemma of littlewood and offord. *Bulletin of the American Mathematical Society*, 51 (12):898–902, 1945.

Behrooz Ghorbani, Hamid Javadi, and Andrea Montanari. An instability in variational inference for topic models. *arXiv preprint arXiv:1802.00568*, 2018.

Ryan J Giordano, Tamara Broderick, and Michael I Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. In *NIPS*, pages 1441–1449, 2015.

Prem K Gopalan and David M Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110 (36):14534–14539, 2013.

Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165 (3-4):1025–1049, 2016.

Matthew D Hoffman and David M Blei. Structured stochastic variational inference. In *Artificial Intelligence and Statistics*, 2015.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1): 1303–1347, 2013.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

T Jaakkola. Tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, page 129, 2001.

Tommi S. Jaakkola and Michael I. Jordon. Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, pages 163–173. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-60032-3. URL `http://dl.acm.org/citation.cfm?id=308574.308663`.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL `https://doi.org/10.1023/A:1007665907178`.

Jeongyeol Kwon and Constantine Caramanis. Global convergence of em algorithm for mixtures of two component linear regression. *arXiv preprint arXiv:1810.05752*, 2018.

Catherine Matias and Vincent Miele. Statistical clustering of temporal networks through a dynamic stochastic block model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4): 1119–1141, 2017.

Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

Elchanan Mossel, Joe Neeman, Allan Sly, et al. Belief propagation, robust reconstruction and optimal recovery of block models. *The Annals of Applied Probability*, 26(4):2211–2256, 2016.

Soumendu Sundar Mukherjee, Purnamrita Sarkar, YX Rachel Wang, and Bowei Yan. Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. In *Advances in Neural Information Processing Systems*, pages 10694–10704, 2018.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational bayes. *arXiv preprint arXiv:1712.08983*, 2017.

Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach.* Cognitive Systems Laboratory, School of Engineering and Applied Science . . . , 1982.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Elsevier, 2014.

Amelia Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. In *2017 International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67. IEEE, 2017.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *ICML*, pages 324–333, 2016.

Zahra S Razaee, Arash A Amini, and Jingyi Jessica Li. Matched bipartite block model with covariates. *Journal of Machine Learning Research*, 20(34):1–44, 2019.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Simone Romano, James Bailey, Vinh Nguyen, and Karin Verspoor. Standardized mutual information for clustering comparisons: one step further in adjustment for chance. In *International Conference on Machine Learning*, pages 1143–1151, 2014.

Purnamrita Sarkar, Y. X. Rachel Wang, and Soumendu Sunder Mukherjee. When random initializations help: a study of variational inference for community detection. *arXiv e-prints*, art. arXiv:1905.06661, May 2019.

T. A. B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14 (1):75–100, 1997.

Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. In *NIPS*, pages 3564–3572, 2015.

Dustin Tran, Rajesh Ranganath, and David Blei. Hierarchical implicit models and likelihood-free variational inference. In *NIPS*, pages 5529–5539, 2017.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Bo Wang, DM Titterington, et al. Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

Ted Westling and Tyler H. McCormick. Beyond prediction: A framework for inference with variational approximations in mixture models. *arXiv preprint arXiv:1510.08151*, 2015.

Mateusz Wilinski, Piero Mazzarisi, Daniele Tantari, and Fabrizio Lillo. Detectability of macroscopic structures in directed asymmetric stochastic block model. *Physical Review E*, 99(4):042310, 2019.

Eric P Xing, Michael I Jordan, and Stuart Russell. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 583–591. Morgan Kaufmann Publishers Inc., 2002.

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient em on multi-component mixture of gaussians. In *Advances in Neural Information Processing Systems*, pages 6956–6966, 2017.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621, 2014.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5660–5669. PMLR, 2018.

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

Supplementary material for “A Theoretical Case Study of Structured Variational Inference for Community Detection”

This supplementary document contains detailed proofs and derivation of theoretical results presented in the main paper “A Theoretical Case Study of Structured Variational Inference for Community Detection”, and additional experimental results. In particular, Section A contains the detailed derivation of updates of the Variational Inference with Pairwise Structure (VIPS) algorithm. Section B contains detailed proofs of the theoretical results presented in the main paper. Section C contains details on how to generalize VIPS to $K = 2$ with unbalanced community sizes and $K = 3$ with equal sized communities, and experimental results on robustness to parameter mis-specification. Section D contains additional experimental results and figures.

A Detailed Derivation of the Updates of VIPS

In the main paper Eq. (5), the Evidence Lower BOund (ELBO) for pairwise structured variational inference is

$$\begin{aligned}\mathcal{L}(Q;\pi,B) =&\frac{1}{2}\mathbb{E}_Q\sum_{i\neq j,a,b} Z_{ia}Z_{jb}(A^{zz}_{ij}\alpha_{ab}+f(\alpha_{ab}))+\frac{1}{2}\mathbb{E}_Q\sum_{i\neq j,a,b} Y_{ia}Y_{jb}(A^{yy}_{ij}\alpha_{ab}+f(\alpha_{ab}))\\ &+\mathbb{E}_Q\sum_{i\neq j,a,b} Z_{ia}Y_{jb}(A^{zy}_{ij}\alpha_{ab}+f(\alpha_{ab}))+\mathbb{E}_Q\sum_{i,a,b} Z_{ia}Y_{ib}(A^{zy}_{ii}\alpha_{ab}+f(\alpha_{ab}))\\ &-\sum_{i=1}^{m}\mathrm{KL}(Q(z_i,y_i)||P(z_i)P(y_i))\end{aligned}$$

where $\alpha_{ab} = \log(B_{ab}/(1 - B_{ab}))$ and $f(\alpha) = -\log(1 + e^{\alpha})$. Denote the first four terms in ELBO as T_1, T_2, T_3, T_4, where T_1, T_2 correspond to the likelihood of the blocks A^{zz} and A^{yy} in the adjacency matrix, T_3 corresponds to the likelihood of $(z_i, y_j), i \neq j$ and T_4 corresponds to (z_i, y_i). Plugging in the marginal density of the independent nodes in T_1, T_2, T_3 and joint density of the dependent nodes in T_4, we have

$$\begin{aligned}T_1 =&\frac{1}{2}\sum_{i\neq j}\Big\{[(1-\phi_i)(1-\phi_j)+\phi_i\phi_j](A^{zz}_{ij}\log\frac{p}{1-p}+\log(1-p))+\\ &[(1-\phi_i)\phi_j+\phi_i(1-\phi_j)](A^{zz}_{ij}\log\frac{q}{1-q}+\log(1-q))\Big\}\end{aligned} \tag{A.1}$$

$$\begin{aligned}T_2 =&\frac{1}{2}\sum_{i\neq j}\Big\{[(1-\xi_i)(1-\xi_j)+\xi_i\xi_j](A^{yy}_{ij}\log\frac{p}{1-p}+\log(1-p))+\\ &[(1-\xi_i)\xi_j+\xi_i(1-\xi_j)](A^{yy}_{ij}\log\frac{q}{1-q}+\log(1-q))\Big\}\end{aligned} \tag{A.2}$$

$$\begin{aligned}T_3 =&\sum_{i\neq j}\Big\{[(1-\phi_i)(1-\xi_j)+\phi_i\xi_j](A^{zy}_{ij}\log\frac{p}{1-p}+\log(1-p))+\\ &[(1-\phi_i)\xi_j+\phi_i(1-\xi_j)](A^{zy}_{ij}\log\frac{q}{1-q}+\log(1-q))\Big\}\end{aligned} \tag{A.3}$$

$$\begin{aligned}T_4 =&\sum_{i}\Big\{(1-\psi^{01}_i-\psi^{10}_i)(A^{zy}_{ii}\log\frac{p}{1-p}+\log(1-p))+\\ &(\psi^{01}_i+\psi^{10}_i)(A^{zy}_{ii}\log\frac{q}{1-q}+\log(1-q))\Big\}\end{aligned} \tag{A.4}$$

The KL regularization term (6) is

$$\mathrm{KL}(Q(z_i,y_i)||P(z_i)P(y_i)) =\psi^{00}_i\log\frac{\psi^{00}_i}{(1-\pi)^2}+\psi^{01}_i\log\frac{\psi^{01}_i}{\pi(1-\pi)}+$$

$$\begin{aligned}&\psi_i^{10}\log\frac{\psi_i^{10}}{\pi(1-\pi)}+\psi_i^{11}\log\frac{\psi_i^{11}}{\pi^2}\\ =&\sum_{0\le c,d\le 1}\psi_i^{cd}\log\frac{\psi_i^{cd}}{\pi^c\pi^d(1-\pi)^{1-c}(1-\pi)^{1-d}}\end{aligned}$$

To take the derivative of $\mathcal{L}(Q;\pi,B)$ with respect to $\psi_i^{cd}, cd\neq 0$, we first have the derivative of the KL term

$$\frac{\partial}{\partial\psi_i^{cd}}\mathrm{KL}(Q(z_i,y_i)||P(z_i)P(y_i)) = \log\frac{\psi_i^{cd}}{\pi^{c+d}(1-\pi)^{2-c-d}}-\log\frac{\psi_i^{00}}{(1-\pi)^2} \tag{A.5}$$

$$= \log\frac{\psi_i^{cd}}{1-\psi_i^{01}-\psi_i^{10}-\psi_i^{11}} \qquad (\pi=\frac{1}{2}) \tag{A.6}$$

Denote the right hand side of Eq. (A.6) as $\theta_i^{cd}\coloneqq\log\frac{\psi_i^{cd}}{1-\psi_i^{01}-\psi_i^{10}-\psi_i^{11}}$. For the reconstruction terms, denoting $T(a,p)\coloneqq a\log(\frac{p}{1-p})+\log(1-p)$ for simplicity, the derivative can be computed as

$$\begin{aligned}\frac{\partial}{\partial\psi_i^{10}}(\sum T_k) = &\sum_{j,j\neq i}\Big[(2\phi_j-1)T(A_{ij}^{zz},p)-(2\phi_j-1)T(A_{ij}^{zz},q)\Big]+\\ &\sum_{j,j\neq i}\Big[(2\xi_j-1)T(A_{ij}^{zy},p)-(2\xi_j-1)T(A_{ij}^{zy},q)\Big]+\\ &\Big[-T(A_{ii}^{zy},p)+T(A_{ii}^{zy},q)\Big]\end{aligned} \tag{A.7}$$

$$\begin{aligned}\frac{\partial}{\partial\psi_i^{01}}(\sum T_k) = &\sum_{j,j\neq i}\Big[(2\xi_j-1)T(A_{ij}^{yy},p)-(2\xi_j-1)T(A_{ij}^{yy},q)\Big]+\\ &\sum_{j,j\neq i}\Big[(2\phi_j-1)T(A_{ji}^{zy},p)-(2\phi_j-1)T(A_{ji}^{zy},q)\Big]+\\ &\Big[-T(A_{ii}^{zy},p)+T(A_{ii}^{zy},q)\Big]\end{aligned} \tag{A.8}$$

$$\begin{aligned}\frac{\partial}{\partial\psi_i^{11}}(\sum T_k) = &\sum_{j,j\neq i}\Big[(2\phi_j-1)T(A_{ij}^{zz},p)-(2\phi_j-1)T(A_{ij}^{zz},q)\Big]+\\ &\sum_{j,j\neq i}\Big[(2\xi_j-1)T(A_{ij}^{yy},p)-(2\xi_j-1)T(A_{ij}^{yy},q)\Big]+\\ &\sum_{j,j\neq i}\Big[(2\xi_j-1)T(A_{ij}^{zy},p)-(2\xi_j-1)T(A_{ij}^{zy},q)\Big]+\\ &\sum_{j,j\neq i}\Big[(2\phi_j-1)T(A_{ji}^{zy},p)-(2\phi_j-1)T(A_{ji}^{zy},q)\Big]\end{aligned} \tag{A.9}$$

Setting the derivatives to 0 we get the update for θ as (10), (9), (11).

B Proofs of Main Results

To prove Theroem 1, we first need a few lemmas. First we have the following lemma for the parameters p, q and λ.

Lemma A.1. *If $p\asymp q\asymp\rho_n, \rho_n\to 0$ and $p-q=\Omega(\rho_n)$, then*

$$\lambda-q=\Omega(\rho_n)>0, \tag{B.1}$$

$$\frac{p+q}{2} - \lambda = \Omega(\rho_n) > 0. \tag{B.2}$$

Proof. The proof follows from Proposition 2 in (Sarkar et al., 2019). □

In the proof, we utilize the spectral property of the population matrix P and generalize it to the finite sample case by bounding the term related to the residual $R = A - P$. We use Berry-Esseen Theorem to bound the residual terms conditioning on u.

Lemma A.2 (Berry-Esseen bound)**.** *Define*

$$r_i = \sum_{j=1}^{n} (A_{ij} - P_{ij})(u(j) - \frac{1}{2}), \tag{B.3}$$

where u and A are independent.

$$\sup_{x \in \mathbb{R}} \left| P\left(r_i / \sigma_u \le x \mid u \right) - \Phi(x) \right| \le \frac{C \rho_u}{\sigma_u^3},$$

where C is a general constant, $\Phi(\cdot)$ is the CDF of standard Gaussian, ρ_u and σ_u depend on u.

Proof. Since r_i is the sum of independent, mean zero random variables, the sum of the conditional variances is

$$\sigma_u^2 = \mathrm{Var}(r_i|u) = p(1-p) \sum_{i \in G_1} (u(i) - \frac{1}{2})^2 + q(1-q) \sum_{i \in G_2} (u(i) - \frac{1}{2})^2,$$

and the sum of the conditional absolute third central moments is

$$\rho_u = p(1-p)(1-2p+2p^2) \sum_{i \in G_1} |u(i) - \frac{1}{2}|^3 + q(1-q)(1-2q+2q^2) \sum_{i \in G_2} |u(i) - \frac{1}{2}|^3.$$

The desired bound follows from the Berry-Esseen Theorem. □

The next lemma shows despite the fact that A introduces some dependency among r_i due to its symmetry, we can still treat r_i as almost iid.

Lemma A.3 (McDiarmid's Inequality)**.** *Let r_i be the noise defined in Lemma A.2 and let $h(r_i)$ be a bounded function with $\|h\|_\infty \le M$. Then*

$$P\left(\left| \frac{2}{n} \sum_{i \in \mathcal{A}} h(r_i) - \mathbb{E}(h(r_i)|u) \right| > w \mid u \right) \le \exp\left(- \frac{c_0 w^2}{nM} \right)$$

for some general constant c_0, provided $|\mathcal{A}| = \Theta_P(n)$.

Proof. The proof follows from Lemma 20 in (Sarkar et al., 2019). □

Lemma A.4. *Let r_i be defined as in Lemma A.2 and assume A and u are independent, we have $\sup_{i \in \mathcal{A}} |r_i| = O_P(\sqrt{n\rho_n \log n})$ if the index set $|\mathcal{A}| = \Theta_P(n)$.*

Proof. Since r_i is the sum of independent bounded random variables, for all i, $r_i = O_P(\sqrt{n\rho_n})$. By Hoeffding inequility, we know for all $t > 0$

$$P(|r_i| > t) \le \exp(-\frac{t^2}{2n\rho_n})$$

and by the union bound

$$P(\sup_i |r_i| > t) \le \exp(C \log n - \frac{t^2}{2n\rho_n})$$

For $\forall \epsilon > 0$, let $t = C_\epsilon \sqrt{n\rho_n \log n}$ with $n^{\frac{C_\epsilon^2}{2} - 1} > 1/\epsilon$, then by definition $\sup_i |r_i| = O_P(\sqrt{n\rho_n \log n})$ □

Next we have a lemma ensuring the signal in the first iteration is not too small.

Lemma A.5 (Littlewood-Offord). *Let $s_1 = (p-\lambda)\sum_{i\in G_1}(u^{(0)}(i)-1/2) + (q-\lambda)\sum_{i\in G_2}(u^{(0)}(i)-1/2)$, $s_2 = (q-\lambda)\sum_{i\in G_1}(u^{(0)}(i)-1/2) + (p-\lambda)\sum_{i\in G_2}(u^{(0)}(i)-1/2)$. Then*

$$P\left(|s_1| \le c\right) \le B \cdot \frac{c}{\rho_n\sqrt{n}}$$

for $c > 0$ and B as constant. The same bound holds for $|s_2|, |s_1 - s_2|$.

Proof. Noting that $2u^{(0)}(i) - 1 \in \{-1, 1\}$ each with probability 1/2, and Lemma A.1, this is a direct consequence of the Littlewood-Offord bound in (Erdös, 1945). □

Finally, we have the following upper and lower bound for some general update ϕ_i.

Lemma A.6. *Assume ϕ_i has the update form $\phi_i = (a + e^{4t(s+r_i)})/(b + e^{4t(s+r_i)})$ for $i \in [m]$, $b > a > 0$ and $b - a$, $(b-a)/b$ are of constant order. r_i is defined as in Lemma A.2. Let set $\mathcal{A} \subset [m]$, with $\Delta > 0$, we have*

$$\sum_{i\in\mathcal{A}} \phi_i \ge |\mathcal{A}| - \frac{b-a}{b}|\mathcal{A}|\Phi(\frac{-s+\Delta}{\sigma_u}) - C'|\mathcal{A}|\frac{\rho_u}{\sigma_u^3} - C''|\mathcal{A}|e^{-4t\Delta} - O_P(\sqrt{|\mathcal{A}|}),$$
$$\sum_{i\in\mathcal{A}} \phi_i \le |\mathcal{A}| - \frac{b-a}{b}|\mathcal{A}|\Phi(\frac{-s-\Delta}{\sigma_u}) + C'|\mathcal{A}|\frac{\rho_u}{\sigma_u^3} + |\mathcal{A}|e^{-4t\Delta} + O_P(\sqrt{|\mathcal{A}|}).$$

Proof. Define the set $J^+ = \{i : r_i > -s + \Delta\}$, $\Delta \ge 0$. For $i \in \mathcal{A} \cap J^+$

$$\phi_i = \frac{a + e^{4t(s+r_i)}}{b + e^{4t(s+r_i)}} \ge \frac{a + e^{4t\Delta}}{b + e^{4t\Delta}} \ge 1 - (b-a)e^{-4t\Delta}$$

For $i \in (\mathcal{A} \cap J^+)^c$, $\phi_i \ge a/b$, therefore

$$\begin{aligned}\sum_{i\in\mathcal{A}} \phi_i \ge& |\mathcal{A} \cap J^+|(1 - (b-a)e^{-4t\Delta}) + \frac{a}{b}(|\mathcal{A}| - |\mathcal{A} \cap J^+|)\\ =& |\mathcal{A} \cap J^+|(\frac{b-a}{b} - (b-a)e^{-4t\Delta}) + \frac{a}{b}|\mathcal{A}|\end{aligned}$$

By Lemmas A.2 and A.3, we have

$$\begin{aligned}|\mathcal{A} \cap J^+| =& \sum_{i\in\mathcal{A}} \mathbf{1}[r_i > -s + \Delta]\\ =& |\mathcal{A}| \cdot P(r_i > -s + \Delta) + O_P(\sqrt{|\mathcal{A}|})\\ \ge& |\mathcal{A}| \cdot (1 - \Phi(\frac{-s+\Delta}{\sigma_u}) - C_0\frac{\rho_u}{\sigma_u^3}) + O_P(\sqrt{|\mathcal{A}|}).\end{aligned}$$

Combining the above,

$$\sum_{i\in\mathcal{A}} \phi_i \ge |\mathcal{A}| - \frac{b-a}{b}|\mathcal{A}|\Phi(\frac{-s+\Delta}{\sigma_u}) - C'|\mathcal{A}|\frac{\rho_u}{\sigma_u^3} - C''|\mathcal{A}|e^{-4t\Delta} - O_P(\sqrt{|\mathcal{A}|})$$

Similarly, define the set $J^- = \{i : r_i < -s - \Delta\}$, $\Delta \ge 0$. For $i \in \mathcal{A} \cap J^-$,

$$\phi_i = \frac{a + e^{4t(s+r_i)}}{b + e^{4t(s+r_i)}} \le \frac{a + e^{-4t\Delta}}{b + e^{-4t\Delta}} \le \frac{a}{b} + e^{-4t\Delta}$$

For $i \in (\mathcal{A} \cap J^-)^c$, $\phi_i \le 1$, so

$$\sum_{i\in\mathcal{A}} \phi_i \le |\mathcal{A} \cap J^-|(\frac{a}{b} + e^{-4t\Delta}) + (|\mathcal{A}| - |\mathcal{A} \cap J^-|)$$

$$=|\mathcal{A}| - |\mathcal{A} \cap J^-|(1 - \frac{a}{b} - e^{-4t\Delta}) + O_P(\sqrt{|\mathcal{A}|})$$

By Lemmas A.2 and A.3,

$$|\mathcal{A} \cap J^-| \geq |\mathcal{A}| \cdot (\Phi(\frac{-s-\Delta}{\sigma_u}) - C_0\frac{\rho_u}{\sigma_u^3}) - O_P(\sqrt{|\mathcal{A}|})$$

so

$$\sum_{i\in\mathcal{A}} \phi_i \leq |\mathcal{A}| - \frac{b-a}{b}|\mathcal{A}|\Phi(\frac{-s-\Delta}{\sigma_u}) + C'|\mathcal{A}|\frac{\rho_u}{\sigma_u^3} + |\mathcal{A}|e^{-4t\Delta} + O_P(\sqrt{|\mathcal{A}|})$$

□

Proof of Theorem 1. Throughout the proof, we assume A has self-loops for convenience, which does not affect the asymptotic results.

Analysis of the first iteration in the first meta iteration:

For random initialized $u^{(0)}$, the initial signal $|\langle u^{(0)}, v_2\rangle| = O_P(\sqrt{n})$. Using the graph split $A^{(0)}$, we write the update of θ^{10} as

$$\begin{aligned}\theta^{10} =& 4t([6(A^{(0)})^{zz}, 6(A^{(0)})^{zy}] - \lambda J)(u^{(0)} - \frac{1}{2}\mathbf{1}_n)\\ =& \underbrace{4t([P^{zz}, P^{zy}] - \lambda J)(u^{(0)} - \frac{1}{2}\mathbf{1}_n)}_{\text{signal}} + \underbrace{4t[6(A^{(0)})^{zz} - P^{zz}, 6(A^{(0)})^{zy} - P^{zy}](u^{(0)} - \frac{1}{2}\mathbf{1}_n)}_{\text{noise}},\end{aligned} \tag{B.4}$$

where P is the population matrix of A. Denote $R^{(0)} = 6A^{(0)} - P$ and $r^{(0)} = [(R^{(0)})^{zz}, (R^{(0)})^{zy}](u^{(0)} - \frac{1}{2}\mathbf{1})$ Since P has singular value decomposition as $P = \frac{p+q}{2}\mathbf{1}_n\mathbf{1}_n^T + \frac{p-q}{2}v_2v_2^T$, the signal part is blockwise constant and we can write

$$\theta^{10} = 4t(s_1\mathbf{1}_{C_1} + s_2\mathbf{1}_{C_2} + r^{(0)}), \tag{B.5}$$

where

$$\begin{aligned} s_1 =& (\frac{p+q}{2} - \lambda)(\langle u^{(0)}, \mathbf{1}_n\rangle - m) + (\frac{p-q}{2})\langle u^{(0)}, v_2\rangle\\ s_2 =& (\frac{p+q}{2} - \lambda)(\langle u^{(0)}, \mathbf{1}_n\rangle - m) - (\frac{p-q}{2})\langle u^{(0)}, v_2\rangle \end{aligned} \tag{B.6}$$

By (12), since we initialize with $\theta^{01}, \theta^{11} = 0$, the marginal probabilities are updated as

$$\phi_1^{(1)} = \frac{1+e^{\theta^{10}}}{3+e^{\theta^{10}}}, \quad \xi_1^{(1)} = \frac{2}{3+e^{\theta^{10}}} \tag{B.7}$$

Next we show the signal $|\langle u, v_2\rangle|$ increases from $O_P(\sqrt{n})$ to $\Omega_P(n\sqrt{\rho_n})$. (We omit the superscript on logits s,x and y now for simplicity.) Since

$$\langle u_1^{(1)}, v_2\rangle = \langle \phi_{1i}^{(1)}, v_{21}\rangle + \langle \xi_{1i}^{(1)}, v_{22}\rangle = \sum_{i\in C_1}\phi_i^{(1)} - \sum_{i\in C_2}\phi_i^{(1)} + \langle \xi^{(1)}, v_{22}\rangle$$

we use Lemma A.6 to bound $\sum_{i\in C_1}\phi_i^{(1)}$ and $\sum_{i\in C_2}\phi_i^{(1)}$. Since s_1 and s_2 depends on $u^{(0)}$, we consider two cases conditioning on $u^{(0)}$.

Case 1: $s_1 > s_2$. By Lemma A.6, let $\Delta = \frac{1}{4}(s_1 - s_2)$ with $\mathcal{A} = C_1, C_2$, $(a, b) = (1, 3)$, conditioning on $u^{(0)}$,

$$\sum_{i\in C_1}\phi_{1i}^{(1)} \geq \frac{n}{6}(1 - \Phi(-\frac{s_1 - \frac{1}{4}(s_1 - s_2)}{\sigma_u})) + \frac{n}{12} - C'n\frac{\rho_u}{\sigma_u^3} - C''ne^{-t(s_1-s_2)} - O_P(\sqrt{n}),$$

$$\sum_{i\in C_2}\phi_{1i}^{(1)} \leq \frac{n}{4} - \frac{n}{6}\Phi(-\frac{s_2+\frac{1}{4}(s_1-s_2)}{\sigma_u}) + C'n\frac{\rho_u}{\sigma_u^3} + C''ne^{-t(s_1-s_2)} + O_P(\sqrt{n}),$$

where the $O_P(\sqrt{n})$ term can be made uniform in $u^{(0)}$. So we have

$$\begin{aligned}\langle\phi_1^{(1)}, v_{21}\rangle \geq& \frac{n}{6}(\Phi(-\frac{s_2+\frac{1}{4}(s_1-s_2)}{\sigma_u}) - \Phi(-\frac{s_1-\frac{1}{4}(s_1-s_2)}{\sigma_u}))\\ &- C'n\frac{\rho_u}{\sigma_u^3} - C''ne^{-t(s_1-s_2)} - O_P(\sqrt{n})\\ \geq& \frac{n}{6\sqrt{2\pi}}(\frac{s_1-s_2}{2\sigma_u})\exp\big(-\frac{s_1^2\vee s_2^2}{2\sigma_u^2}\big)\\ &- C'n\frac{\rho_u}{\sigma_u^3} - C''ne^{-t(s_1-s_2)} - O_P(\sqrt{n}).\end{aligned} \tag{B.8}$$

Here to approximate the CDF Φ, we have used

$$\begin{aligned}|\Phi(x)-1/2| &= \frac{1}{\sqrt{2\pi}}\int_0^{|x|} e^{-u^2/2}du\\ &\geq \frac{|x|}{\sqrt{2\pi}}e^{-x^2/2}.\end{aligned} \tag{B.9}$$

Case 2: $s_1 < s_2$. The same analysis applies with s_1 and s_2 interchanged.

Combining *Case 1* and *Case 2*, for any given $u^{(0)}$,

$$\begin{aligned}|\langle\phi_1^{(1)}, v_{21}\rangle| \geq& \frac{n}{6\sqrt{2\pi}}(\frac{|s_1-s_2|}{2\sigma_u})\exp\big(-\frac{s_1^2\vee s_2^2}{2\sigma_u^2}\big)\\ &- C'n\frac{\rho_u}{\sigma_u^3} - C''ne^{-t|s_1-s_2|} - O_P(\sqrt{n}).\end{aligned} \tag{B.10}$$

We note that $|s_1|$, $|s_2|$, $|s_1-s_2|$ are of order $\Omega_P(\sqrt{n}\rho_n)$ by Lemma A.5. Also $\sigma_u^2, \rho_u \asymp n\rho_n$, $e^{-4t|s_1-s_2|} = \exp(-\Omega(\rho_n\sqrt{n}))$. We can conclude that $|\langle\phi_1^{(1)}, v_{21}\rangle| = \Omega_P(n\sqrt{\rho_n})$.

For $\langle\xi_1^{(1)}, v_{22}\rangle$ we have

$$\begin{aligned}|\langle\xi_1^{(1)}, v_{22}\rangle| =& \Big|\sum_{i\in C_1'}\xi_i^{(1)} - \sum_{i\in C_2'}\xi_i^{(1)}\Big| = \Big|\sum_{i\in C_2'}\phi_i^{(1)} - \sum_{i\in C_1'}\phi_i^{(1)} + |C_1'| - |C_2'|\Big|\\ =& O_P(\sqrt{n})\end{aligned}$$

Therefore we have $|\langle u_1^{(1)}, v_2\rangle| = \Omega_P(n\sqrt{\rho_n})$. By (B.7), $\langle u_1^{(1)}, \mathbf{1}\rangle - m = 0$.

Due to the symmetry in s_1 and s_2, WLOG in the following analysis, we assume $\langle u_1^{(1)}, v_2\rangle > 0$ (equivalently $s_1 > s_2$).

Analysis of the second iteration in the first meta iteration:

Similar to (B.4), we can write

$$\begin{aligned}\theta^{01} =& 4t([6(A^{(1)})^{yz}, 6(A^{(1)})^{yy}] - \lambda J)(u_1^{(1)} - \frac{1}{2}\mathbf{1}_n)\\ =& \underbrace{4t([P^{yz}, P^{yy}] - \lambda J)(u_1^{(1)} - \frac{1}{2}\mathbf{1}_n)}_{\text{signal}} + \underbrace{4t(R^{(1)})^{yz}(\phi_1^{(1)} - \frac{1}{2}\mathbf{1}_m) + 4t(R^{(1)})^{yy}(\xi_1^{(1)} - \frac{1}{2}\mathbf{1}_m)}_{\text{noise} := 4tr_i^{(1)}}.\end{aligned}$$

Noting the signal part is blockwise constant, we have

$$\theta^{01} = 4t(x_1\mathbf{1}_{C_1'} + x_2\mathbf{1}_{C_2'} + r^{(1)}),$$

where

$$x_1 =(\frac{p+q}{2}-\lambda)(\langle u_1^{(1)}, \mathbf{1}_n\rangle - m) + (\frac{p-q}{2})\langle u_1^{(1)}, v_2\rangle$$
$$x_2 =(\frac{p+q}{2}-\lambda)(\langle u_1^{(1)}, \mathbf{1}_n\rangle - m) - (\frac{p-q}{2})\langle u_1^{(1)}, v_2\rangle$$

By (B.7), $\langle u_1^{(1)}, \mathbf{1}_n\rangle - m = 0$ and we have

$$x_1 =(\frac{p-q}{2})\langle u_1^{(1)}, v_2\rangle,$$
$$x_2 =-x_1.$$

It follows then from the first iteration that $x_1, -x_2 = \Omega_P(n\rho_n^{3/2})$. The update for $u_2^{(1)}$ is

$$\phi_2^{(1)} = \frac{1+e^{\theta^{10}}}{2+e^{\theta^{10}}+e^{\theta^{01}}}, \quad \xi_2^{(1)} = \frac{1+e^{\theta^{01}}}{2+e^{\theta^{10}}+e^{\theta^{01}}} \tag{B.11}$$

Since the signal part of θ^{10} and θ^{01} are blockwise constant on C_1, C_2 and C_1', C_2' respectively, $\langle u_2^{(1)}, v_2\rangle$ can be calculated as

$$\begin{aligned}\langle \phi_2^{(1)}, v_{21}\rangle =& \sum_{i\in C_{11}} \frac{1+e^{4t(s_1+r_i^{(0)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}} + \sum_{i\in C_{12}} \frac{1+e^{4t(s_1+r_i^{(0)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}}\\ &- \sum_{i\in C_{21}} \frac{1+e^{4t(s_2+r_i^{(0)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}} - \sum_{i\in C_{22}} \frac{1+e^{4t(s_2+r_i^{(0)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}}\end{aligned}$$

$$\begin{aligned}\langle \xi_2^{(1)}, v_{22}\rangle =& \sum_{i\in C_{11}} \frac{1+e^{4t(x_1+r_i^{(1)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}} + \sum_{i\in C_{21}} \frac{1+e^{4t(x_1+r_i^{(1)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}}\\ &- \sum_{i\in C_{12}} \frac{1+e^{4t(x_2+r_i^{(1)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}} - \sum_{i\in C_{22}} \frac{1+e^{4t(x_2+r_i^{(1)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}}\end{aligned}$$

In the case of $\langle u_1^{(1)}, v_2\rangle > 0$, we know , $s_1 > s_2$ and $x_1 > 0 > x_2$. We first show that $\langle \phi_2^{(1)}, v_{21}\rangle$ is positive by finding a lower bound for the summations over C_{12}, C_{21}, C_{22} (since the sum over C_{11} is always positive).

For the summation over C_{12}, note that $|x_2|$ dominates both s_1 and $r_i^{(0)}$, $r_i^{(1)}$ by Lemma A.4, we have

$$\sum_{i\in C_{12}} \frac{1+e^{4t(s_1+r_i^{(0)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}} = \sum_{i\in C_{12}} \frac{1+e^{4t(s_1+r_i^{(0)})}}{2+e^{4t(s_1+r_i^{(0)})}} + n\exp(-\Omega_P(n\rho_n^{3/2})).$$

To lower bound the first term, we use Lemma A.6 by first conditioning on $u^{(0)}$,

$$\begin{aligned}&\sum_{i\in C_{12}} \frac{1+e^{4t(s_1+r_i^{(0)})}}{2+e^{4t(s_1+r_i^{(0)})}}\\ \geq& \frac{n}{8}\left(1-\frac{1}{2}\Phi(\frac{-s_1+\Delta}{\sigma_u})\right) - C'n\frac{\rho_u}{\sigma_u^3} - C''ne^{-4t\Delta} - O_P(\sqrt{n})\end{aligned} \tag{B.12}$$

For the summation over C_{22},

$$\sum_{i\in C_{22}} \frac{1+e^{4t(s_2+r_i^{(0)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}} \leq \sum_{i\in C_{22}} \frac{1+e^{4t(s_2+r_i^{(0)})}}{2+e^{4t(s_2+r_i^{(0)})}}$$

$$\leq \frac{n}{8}\left(1-\frac{1}{2}\Phi(\frac{-s_2-\Delta}{\sigma_u})\right)+C'n\frac{\rho_u}{\sigma_u^3}+C''ne^{-4t\Delta}+O_P(\sqrt{n}) \tag{B.13}$$

For the summation over C_{21}, x_1 dominates s_2 and $r_i^{(0)}, r_i^{(1)}$ by Lemma A.4,

$$\sum_{i\in C_{21}} \frac{1+e^{4t(s_2+r_i^{(0)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}} = n\exp(-\Omega_P(n\rho_n^{3/2})). \tag{B.14}$$

Combining (B.12) - (B.14), setting $\Delta = \frac{1}{4}(s_1-s_2)$, we have

$$\begin{aligned}\langle \phi_2^{(1)}, v_{21}\rangle \geq & \frac{n}{8}\big[\frac{1}{2}\Phi(\frac{-s_2-\Delta}{\sigma_u})-\frac{1}{2}\Phi(\frac{-s_1+\Delta}{\sigma_u})\big]-C'n\frac{\rho_u}{\sigma_u^3}-C''ne^{-4t\Delta}-O_P(\sqrt{n})\\ \geq & \frac{n}{16}\frac{1}{\sqrt{2\pi}}(\frac{s_1-s_2}{\sigma_u})\exp\big(-\frac{s_1^2\vee s_2^2}{2\sigma_u^2}\big)-C'n\frac{\rho_u}{\sigma_u^3}-C''ne^{-t(s_1-s_2)}-O_P(\sqrt{n})\end{aligned}$$

by the same argument as (B.8). As before, we can see that

$$\langle \phi_2^{(1)}, v_{21}\rangle = \Omega_P(n\sqrt{\rho_n})$$

For $\langle \xi_2^{(1)}, v_{22}\rangle$, since $(1+e^x)/(2+e^x)\leq 1/2+e^x$, we have

$$\begin{aligned}&\sum_{i\in C_{12}} \frac{1+e^{4t(x_2+r_i^{(1)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}} + \sum_{i\in C_{22}} \frac{1+e^{4t(x_2+r_i^{(1)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_2+r_i^{(1)})}}\\ \leq & \frac{n}{8}+\sum_{i\in C_2'} e^{4t(x_2+r_i^{(1)})}+O_P(\sqrt{n})\\ \leq & \frac{n}{8}+O_P(\sqrt{n}).\end{aligned} \tag{B.15}$$

For the other two sums, we have

$$\begin{aligned}\sum_{i\in C_{11}} \frac{1+e^{4t(x_1+r_i^{(1)})}}{2+e^{4t(s_1+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}} &\geq \frac{n}{8}-O_P(\sqrt{n})-n\exp(-\Omega_P(n\rho_n^{3/2})),\\ &\geq \frac{n}{8}-O_P(\sqrt{n})\end{aligned} \tag{B.16}$$

and

$$\sum_{i\in C_{21}} \frac{1+e^{4t(x_1+r_i^{(1)})}}{2+e^{4t(s_2+r_i^{(0)})}+e^{4t(x_1+r_i^{(1)})}} \geq \frac{n}{8}-O_P(\sqrt{n}) \tag{B.17}$$

Equations (B.15) - (B.17) imply

$$\langle \xi_2^{(1)}, v_{22}\rangle \geq \frac{n}{8}-O_P(\sqrt{n}).$$

Therefore $\langle u_2^{(1)}, v_2\rangle \geq n/8 - O_P(\sqrt{n})$. Since by (B.11), $\phi_2^{(1)} = \mathbf{1}_m - \xi_2^{(1)}$, the inner product $\langle u_2^{(1)}, \mathbf{1}\rangle - m = 0$.

Analysis of the third iteration in the first meta iteration:

Similar to the previous two iterations, we can write

$$\theta^{11} = 4t(y_1\mathbf{1}_{C_1}+y_2\mathbf{1}_{C_2}+y_1\mathbf{1}_{C_1'}+y_2\mathbf{1}_{C_2'}+r^{(2)}),$$

where

$$y_1 = (\frac{p-q}{2})\langle u_2^{(1)}, v_2\rangle, \quad y_2 = -y_1$$

$$r^{(2)} = [(R^{(2)})^{zz}, (R^{(2)})^{zy}](u_2 - \frac{1}{2}\mathbf{1}_n) + [(R^{(2)})^{yz}, (R^{(2)})^{yy}](u_2^{(1)} - \frac{1}{2}\mathbf{1}_n).$$

It follows from the second iteration that $y_1, -y_2 = \Omega_P(n\rho_n)$.

The update for $u_3^{(1)}$ is

$$\phi_3^{(1)} = \frac{e^{\theta^{11}} + e^{\theta^{10}}}{1 + e^{\theta^{10}} + e^{\theta^{01}} + e^{\theta^{11}}}, \quad \xi_3^{(1)} = \frac{e^{\theta^{11}} + e^{\theta^{01}}}{1 + e^{\theta^{10}} + e^{\theta^{01}} + e^{\theta^{11}}} \tag{B.18}$$

The $\langle u_3^{(1)}, v_2 \rangle$ can be calculated as

$$\begin{aligned}
&\langle u_3^{(1)}, v_2 \rangle = \\
&\sum_{i \in C_{11}} \frac{2e^{8t(y_1 + r_i^{(2)})} + e^{4t(x_1 + r_i^{(1)})} + e^{4t(s_1 + r_i^{(0)})}}{1 + e^{4t(s_1 + r_i^{(0)})} + e^{4t(x_1 + r_i^{(1)})} + e^{8t(y_1 + r_i^{(2)})}} + \sum_{i \in C_{12}} \frac{e^{4t(s_1 + r_i^{(0)})} - e^{4t(x_2 + r_i^{(1)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})}} \\
&+ \sum_{i \in C_{21}} \frac{e^{4t(x_1 + r_i^{(1)})} - e^{4t(s_2 + r_i^{(0)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_2 + r_i^{(0)})} + e^{4t(x_1 + r_i^{(1)})}} - \sum_{i \in C_{22}} \frac{2e^{8t(y_2 + r_i^{(2)})} + e^{4t(s_2 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})}}{1 + e^{4t(s_2 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})} + e^{8t(y_2 + r_i^{(2)})}}
\end{aligned} \tag{B.19}$$

Using the order of the x terms and y terms and Lemma A.4, we can lower bound $\langle u_3^{(1)}, v_2 \rangle$ by

$$\begin{aligned}
\langle u_3^{(1)}, v_2 \rangle &\geq \frac{n}{4} + \sum_{i \in C_{12}} \frac{e^{4t(s_1 + r_i^{(0)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})}} + \frac{n}{8} - \sum_{i \in C_{22}} \frac{e^{4t(s_2 + r_i^{(0)})}}{1 + e^{4t(s_2 + r_i^{(0)})}} - O_P(\sqrt{n}) \\
&\geq \frac{n}{4} - O_P(\sqrt{n}).
\end{aligned} \tag{B.20}$$

Next we bound $\langle u_3^{(1)}, \mathbf{1}_n \rangle - m$.

$$\begin{aligned}
&\langle u_3^{(1)}, \mathbf{1}_n \rangle = \\
&\sum_{i \in C_{11}} \frac{2e^{8t(y_1 + r_i^{(2)})} + e^{4t(x_1 + r_i^{(1)})} + e^{4t(s_1 + r_i^{(0)})}}{1 + e^{4t(s_1 + r_i^{(0)})} + e^{4t(x_1 + r_i^{(1)})} + e^{8t(y_1 + r_i^{(2)})}} + \sum_{i \in C_{12}} \frac{2e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})}} \\
&+ \sum_{i \in C_{21}} \frac{2e^{4tr_i^{(2)}} + e^{4t(x_1 + r_i^{(1)})} + e^{4t(s_2 + r_i^{(0)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_2 + r_i^{(0)})} + e^{4t(x_1 + r_i^{(1)})}} + \sum_{i \in C_{22}} \frac{2e^{8t(y_2 + r_i^{(2)})} + e^{4t(s_2 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})}}{1 + e^{4t(s_2 + r_i^{(0)})} + e^{4t(x_2 + r_i^{(1)})} + e^{8t(y_2 + r_i^{(2)})}},
\end{aligned} \tag{B.21}$$

Then

$$\langle u_3^{(1)}, \mathbf{1}_n \rangle = \frac{3n}{8} + \sum_{i \in C_{12}} \frac{2e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})}} + \sum_{i \in C_{22}} \frac{e^{4t(s_2 + r_i^{(0)})}}{1 + e^{4t(s_2 + r_i^{(0)})}} + O_P(\sqrt{n}),$$

$$\langle u_3^{(1)}, \mathbf{1}_n \rangle \geq \frac{3n}{8} - O_P(\sqrt{n}),$$

and

$$\begin{aligned}
\langle u_3^{(1)}, \mathbf{1}_n \rangle &\leq \frac{3n}{8} + \sum_{i \in C_{12}} \left(\frac{e^{4tr_i^{(2)}}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})}} + \frac{e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})}}{1 + e^{4tr_i^{(2)}} + e^{4t(s_1 + r_i^{(0)})}} \right) \\
&\quad + \sum_{i \in C_{22}} \frac{e^{4t(s_2 + r_i^{(0)})}}{1 + e^{4t(s_2 + r_i^{(0)})}} + O_P(\sqrt{n}) \\
&\leq \frac{3n}{4} + O_P(\sqrt{n})
\end{aligned}$$

It follows then

$$-n/8 - O_P(\sqrt{n}) \le \langle u_3^{(1)}, \mathbf{1}_n \rangle - m \le n/4 + O_P(\sqrt{n}). \tag{B.22}$$

Analysis of the second meta iteration:

We first show that from the previous iteration, the signal $\langle u_3, v_2 \rangle$ will always dominate $|\langle u_3, \mathbf{1}_n \rangle - m|$ which gives desired sign and magnitude of the logits. Then we show the algorithm converges to the true labels after the second meta iteration.

Using the same decomposition as (B.5),

$$\begin{aligned}
s_1^{(2)} =& (\frac{p+q}{2} - \lambda)(\langle u_3^{(1)}, \mathbf{1}_n \rangle - m) + \frac{p-q}{2} \langle u_3^{(1)}, v_2 \rangle && \text{(B.23)} \\
\ge& -\frac{n}{8}(\frac{p+q}{2} - \lambda) + \frac{n}{4} \cdot \frac{p-q}{2} - o_P(n\rho_n) \\
\ge& \frac{n}{8}(\lambda - q) - o_P(n\rho_n) \\
s_2^{(2)} =& (\frac{p+q}{2} - \lambda)(\langle u_3^{(1)}, \mathbf{1}_n \rangle - m) - \frac{p-q}{2} \langle u_3^{(1)}, v_2 \rangle && \text{(B.24)} \\
\le& \frac{n}{4}(\frac{p+q}{2} - \lambda) - \frac{n}{4} \cdot \frac{p-q}{2} + o_P(n\rho_n) \\
=& -\frac{n}{4}(\lambda - q) + o_P(n\rho_n), && \text{(B.25)}
\end{aligned}$$

where we have used Lemma A.1.

After the first meta iteration, the logits satisfy

$$\begin{aligned}
&s_1^{(2)}, -s_2^{(2)} = \Omega_P(n\rho_n), \qquad x_1^{(1)}, -x_2^{(1)} = \Omega_P(n\rho_n^{\frac{3}{2}}), \\
&y_1^{(1)}, -y_2^{(1)} = \Omega_P(n\rho_n).
\end{aligned}$$

Here we have added the superscripts for the first meta iteration for clarity.

In the first iteration of the second meta iteration, $\langle u_1^{(2)}, v_2 \rangle$ is computed as (B.19) with s_1 and s_2 replaced with $s_1^{(2)}$ and $s_2^{(2)}$ and the noise replaced accordingly. It is easy to see that

$$\langle u_1^{(2)}, v_2 \rangle \ge \frac{3n}{8} - o_P(n). \tag{B.26}$$

Similarly from (B.21),

$$-\frac{n}{8} - o_P(n) \le \langle u_1^{(2)}, \mathbf{1}_n \rangle - m \le o_P(n). \tag{B.27}$$

The logits are updated as $(\frac{p+q}{2} - \lambda)(\langle u_1^{(2)}, \mathbf{1}_n \rangle - m) \pm \frac{p-q}{2} \langle u_1^{(2)}, v_2 \rangle$, so

$$x_1^{(2)}, -x_2^{(2)} = \Omega_P(n\rho_n), \tag{B.28}$$

The same analysis and results hold for $u_2^{(2)}$ and $(y_1^{(2)}, y_2^{(2)})$. We now show after the second meta iteration, in addition to the condition (B.28), we further have

$$2y_1^{(2)} - s_1^{(2)} = \Omega_P(n\rho_n), \quad 2y_1^{(2)} - x_1^{(2)} = \Omega_P(n\rho_n) \tag{B.29}$$

To simplify notation, let

$$\alpha_i(s_1, x_1, y_1) \coloneqq \frac{2e^{8t(y_1 + r_i^{(y)})} + e^{4t(x_1 + r_i^{(x)})} + e^{4t(s_1 + r_i^{(s)})}}{1 + e^{4t(s_1 + r_i^{(s)})} + e^{4t(x_1 + r_i^{(x)})} + e^{8t(y_1 + r_i^{(y)})}}$$

where r's are the noise associated with each signal and we have Lemma A.4 bounding their order uniformly.

We first provide an upper bound on $\langle u_3^{(1)}, v_2\rangle$. In (B.19),

$$\begin{aligned}\langle u_3^{(1)}, v_2\rangle &\le \frac{n}{4} + \sum_{i\in C_{12}} \frac{e^{4t(s_1^{(1)}+r_i^{(0)})}}{1+e^{4t(s_1^{(1)}+r_i^{(0)})}} + \frac{n}{8} - \sum_{i\in C_{22}} \frac{e^{4t(s_2^{(1)}+r_i^{(0)})}}{1+e^{4t(s_2^{(1)}+r_i^{(0)})}} + O_P(\sqrt{n}) \\ &\le \frac{3n}{8} + \frac{n}{8}\left(\Phi(\frac{-s_2^{(1)}+\Delta}{\sigma_u}) - \Phi(\frac{-s_1^{(1)}-\Delta}{\sigma_u})\right) + C'n\frac{\rho_u}{\sigma_u^3} + C''ne^{-4t\Delta} + O_P(\sqrt{n}) \\ &\le \frac{3n}{8} + o_P(n). \end{aligned} \tag{B.30}$$

by Lemma A.6.

For $u_1^{(2)}$, based on (B.19) and (B.21),

$$\begin{aligned}\langle u_1^{(2)}, v_2\rangle &= \sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(1)}, y_1^{(1)}) + \frac{n}{4} - o_P(n), \\ \langle u_1^{(2)}, \mathbf{1}_n\rangle - m &= \sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(1)}, y_1^{(1)}) - \frac{n}{4} - o_P(n).\end{aligned}$$

Similarly,

$$\begin{aligned}\langle u_2^{(2)}, v_2\rangle &= \sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) + \frac{n}{4} - o_P(n), \\ \langle u_2^{(2)}, \mathbf{1}_n\rangle - m &= \sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) - \frac{n}{4} - o_P(n).\end{aligned}$$

For convenience denote $a = \frac{p+q}{2} - \lambda$ and $b = \frac{p-q}{2}$, then we have

$$\begin{aligned}2y_1^{(2)} - s_1^{(2)} &= a(2\langle u_2^{(2)}, \mathbf{1}_n\rangle - \langle u_3^{(1)}, \mathbf{1}_n\rangle - m) + b(2\langle u_2^{(2)}, v_2\rangle - \langle u_3^{(1)}, v_2\rangle) \\ &\ge a\left(2\sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) - \frac{n}{4} - m\right) \\ &\quad + b\left(2\sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) + \frac{n}{2} - \frac{3n}{8}\right) - o_P(n\rho_n) \\ &= 2(a+b)\sum_{i\in C_{11}} \alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) - \frac{3an}{8} + \frac{bn}{8} - o_P(n\rho_n)\end{aligned}$$

by (B.30) and (B.22). Since $\alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) \ge 1 + o_P(1)$, we can conclude

$$2y_1^{(2)} - s_1^{(2)} \ge \frac{3bn}{8} - \frac{an}{8} - o_P(n\rho_n) = \Omega(n\rho_n).$$

Similarly, we can check that

$$\begin{aligned}2y_1^{(2)} - x_1^{(2)} &= a(2\langle u_2^{(2)}, \mathbf{1}_n\rangle - \langle u_1^{(2)}, \mathbf{1}_n\rangle - m) + b(2\langle u_2^{(2)}, v_2\rangle - \langle u_1^{(2)}, v_2\rangle) \\ &= (a+b)\sum_{i\in C_{11}} [2\alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) - \alpha_i(s_1^{(2)}, x_1^{(1)}, y_1^{(1)})] - \frac{(a-b)n}{4} + o_P(n\rho_n) \\ &\ge \frac{(b-a)n}{4} - o_P(n\rho_n) = \Omega(n\rho_n)\end{aligned} \tag{B.31}$$

as $\alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) > \alpha_i(s_1^{(2)}, x_1^{(1)}, y_1^{(1)})$. Thus condition (B.29) holds.

Now we need to analyze the third iteration in this meta iteration. Since $\alpha_i(s_1^{(2)}, x_1^{(2)}, y_1^{(1)}) \leq 2$,

$$y_1^{(2)} + y_2^{(2)} = 2a(\langle u_2^{(2)}, \mathbf{1}_n\rangle - m) = o_P(n\rho_n),$$

then by (B.25)

$$s_1^{(2)} - (y_1^{(2)} + y_2^{(2)}) = \Omega_P(n\rho_n), \quad x_1^{(2)} - (y_1^{(2)} + y_2^{(2)}) = \Omega_P(n\rho_n). \tag{B.32}$$

Now using the update for $u_3^{(2)}$, and defining the noise in the same way as in the first meta iteration,

$$\begin{aligned}
\langle u_3^{(2)}, v_2\rangle = & \sum_{i\in C_{11}} \frac{2e^{8t(y_1^{(2)}+r_i^{(5)})} + e^{4t(x_1^{(2)}+r_i^{(4)})} + e^{4t(s_1^{(2)}+r_i^{(3)})}}{1 + e^{4t(s_1^{(2)}+r_i^{(3)})} + e^{4t(x_1^{(2)}+r_i^{(4)})} + e^{8t(y_1^{(2)}+r_i^{(5)})}} \\
& + \sum_{i\in C_{12}} \frac{e^{4t(s_1^{(2)}+r_i^{(3)})} - e^{4t(x_2^{(2)}+r_i^{(4)})}}{1 + e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} + e^{4t(s_1^{(2)}+r_i^{(3)})} + e^{4t(x_2^{(2)}+r_i^{(4)})}} \\
& + \sum_{i\in C_{21}} \frac{e^{4t(x_1^{(2)}+r_i^{(4)})} - e^{4t(s_2^{(2)}+r_i^{(3)})}}{1 + e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} + e^{4t(s_2^{(2)}+r_i^{(3)})} + e^{4t(x_1^{(2)}+r_i^{(4)})}} \\
& - \sum_{i\in C_{22}} \frac{2e^{8t(y_2^{(2)}+r_i^{(5)})} + e^{4t(s_2^{(2)}+r_i^{(3)})} + e^{4t(x_2^{(2)}+r_i^{(4)})}}{1 + e^{4t(s_2^{(2)}+r_i^{(3)})} + e^{4t(x_2^{(2)}+r_i^{(4)})} + e^{8t(y_2^{(2)}+r_i^{(5)})}} \\
\geq & \sum_{i\in C_{11}} \frac{2e^{8t(y_1^{(2)}+r_i^{(5)})}}{1 + e^{4t(s_1^{(2)}+r_i^{(3)})} + e^{4t(x_1^{(2)}+r_i^{(4)})} + e^{8t(y_1^{(2)}+r_i^{(5)})}} \\
& + \sum_{i\in C_{12}} \frac{e^{4t(s_1^{(2)}+r_i^{(3)})}}{1 + e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} + e^{4t(s_1^{(2)}+r_i^{(3)})} + e^{4t(x_2^{(2)}+r_i^{(4)})}} \\
& + \sum_{i\in C_{21}} \frac{e^{4t(x_1^{(2)}+r_i^{(4)})}}{1 + e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} + e^{4t(s_2^{(2)}+r_i^{(3)})} + e^{4t(x_1^{(2)}+r_i^{(4)})}} \\
& - n\exp(-\Omega_P(n\rho_n)) \\
\geq & \frac{n}{2} - n\exp(-\Omega_P(n\rho_n)),
\end{aligned}$$

using the conditions (B.28) (B.29) (B.32) and Lemma A.4. Since $\|u - z^*\|_1 = m - |\langle u, v_2\rangle|$, $||u_3^{(2)} - z^*||_1 = n\exp(-\Omega_P(n\rho_n))$ after the second meta iteration.

Finally we show the later iterations conserve strong consistency. Since

$$\begin{aligned}
\langle u_3^{(2)}, \mathbf{1}\rangle - m = & \sum_{i\in C_{11}} \frac{e^{8t(y_1^{(2)}+r_i^{(5)})} - 1}{1 + e^{4t(s_1^{(2)}+r_i^{(3)})} + e^{4t(x_1^{(2)}+r_i^{(4)})} + e^{8t(y_1^{(2)}+r_i^{(5)})}} \\
& + \sum_{i\in C_{12}} \frac{e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} - 1}{1 + e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} + e^{4t(s_1^{(2)}+r_i^{(3)})} + e^{4t(x_2^{(2)}+r_i^{(4)})}} \\
& + \sum_{i\in C_{21}} \frac{e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} - 1}{1 + e^{4t(y_1^{(2)}+y_2^{(2)}+r_i^{(5)})} + e^{4t(s_2^{(2)}+r_i^{(3)})} + e^{4t(x_1^{(2)}+r_i^{(4)})}} \\
& + \sum_{i\in C_{22}} \frac{e^{8t(y_2^{(2)}+r_i^{(5)})} - 1}{1 + e^{4t(s_2^{(2)}+r_i^{(3)})} + e^{4t(x_2^{(2)}+r_i^{(4)})} + e^{8t(y_2^{(2)}+r_i^{(5)})}} \\
= & n\exp(-\Omega_P(n\rho_n))
\end{aligned}$$

by (B.28) (B.29) (B.32) and Lemma A.4, we have

$$s_1^{(3)} = a(\langle u_3^{(2)}, \mathbf{1}\rangle - m) + b\langle u_3^{(2)}, v_2\rangle = \frac{p-q}{4}n + n\rho_n\exp(-\Omega_P(n\rho_n)),$$

$$s_2^{(3)} = a(\langle u_3^{(2)}, \mathbf{1}\rangle - m) - b\langle u_3^{(2)}, v_2\rangle = -\frac{p-q}{4}n + n\rho_n \exp(-\Omega_P(n\rho_n)).$$

Next we note the noise in this iteration now arises from the whole graph A, and can be bounded by

$$\begin{aligned} r_i^{(7)} &= [R^{zz}, R^{zy}]_{i,\cdot}(u_3^{(2)} - \frac{1}{2}\mathbf{1}_n) \\ &= [R^{zz}, R^{zy}]_{i,\cdot}(u_3^{(2)} - z^*) + [R^{zz}, R^{zy}]_{i,\cdot}(z^* - \frac{1}{2}\mathbf{1}_n), \end{aligned}$$

where the second term is $O_P(\sqrt{n\rho_n \log n})$ uniformly for all i, applying Lemma A.4. To bound the first term, note that

$$\begin{aligned} \max_i |[R^{zz}, R^{zy}]_{i,\cdot}(u_3^{(2)} - z^*)| &\leq \|[R^{zz}, R^{zy}](u_3^{(2)} - z^*)\|_2 \\ &\leq O_P(\sqrt{n\rho_n})\|u_3^{(2)} - z^*\|_1 = o_P(1). \end{aligned}$$

Therefore $r_i^{(7)}$ is uniformly $O_P(\sqrt{n\rho_n \log n})$ for all i. By a similar calculation to (B.31), we can check that condition (B.29) holds for $y_1^{(2)}$ and $s_1^{(3)}$, since when $s_1, x_1, y_1 = \Omega(n\rho_n)$ condition (B.29) and $1 - o_P(1) \leq \alpha_i(s_1, x_1, y_1) \leq 2 + o_P(1)$ guarantees each other and condition (B.29) is true in the previous iteration. We can check that condition (B.32) also holds. The rest of the argument can be applied to show $\|u_1^{(3)} - z^*\|_1 = n\exp(-\Omega_P(n\rho_n))$. At this point, all the arguments can be repeated for later iterations.

□

Proof of Corollary 1. We first consider $\mu > 0.5$. By (B.6), $s_1 = \Omega_P(n\rho_n)$, $s_2 = \Omega_P(n\rho_n)$. Since $r_i^{(0)} = O_P(\sqrt{n\rho_n \log n})$ uniformly for all i by Lemma A.4, we have

$$\phi_i^{(1)} = \frac{1 + e^{4t(s_1 + r_i^{(0)})}}{3 + e^{4t(s_1 + r_i^{(0)})}} = 1 - \exp(-\Omega_P(n\rho_n))$$

for $i \in C_1$. Similarly for $i \in C_2$, and $\xi_i^{(1)} = \exp(-\Omega_P(n\rho_n))$. Define $u_i' = \mathbf{1}_{[i\in P_1]} + \mathbf{1}_{[i \in P_2]}$. Since the partition into P_1 and P_2 is random, $u_i' \sim$ iid Bernoulli(1/2), and $\|u_1 - u'\|_2 = \sqrt{n}\exp(-\Omega_P(n\rho_n))$.

In the second iteration, we can write

$$\begin{aligned} \theta^{01} &= 4t([A^{yz}, A^{yy}] - \lambda J)(u_1 - \tfrac{1}{2}\mathbf{1}) \\ &= 4t([A^{yz}, A^{yy}] - \lambda J)(u_1 - u') + 4t([A^{yz}, A^{yy}] - \lambda J)(u' - \tfrac{1}{2}\mathbf{1}) \\ &= O_P(n\sqrt{\rho}\exp(-\Omega_P(n\rho_n))) + 4t([A^{yz}, A^{yy}] - \lambda J)(u' - \tfrac{1}{2}\mathbf{1}). \end{aligned}$$

The signal part of the second term is $4t(x_1\mathbf{1}_{C_1'} + x_2\mathbf{1}_{C_2'})$ with x_1 and x_2 having the form of (B.6), with $u^{(0)}$ replaced by u'. Since $x_1, x_2 = \Omega_P(\sqrt{n}\rho_n)$, the rest of the analysis proceeds like that of Theorem 1 restarting from the first iteration.

If $\mu < 0.5$, $s_1 = -\Omega_P(n\rho_n)$, $s_2 = -\Omega_P(n\rho_n)$. We have $\phi_i^{(1)} = \frac{1}{3} + \exp(-\Omega_P(n\rho_n))$, $\xi_i^{(1)} = \frac{2}{3} - \exp(-\Omega_P(n\rho_n))$. This time let $u' = \frac{1}{3}\mathbf{1}_{[i\in P_1]} + \frac{2}{3}\mathbf{1}_{[i\in P_2]}$, then θ^{01} can be written as

$$\theta^{01} = O_P(n\sqrt{\rho}\exp(-\Omega_P(n\rho_n))) + \frac{4t}{3}([A^{yz}, A^{yy}] - \lambda J)(3u' - \frac{3}{2}\mathbf{1}).$$

Noting that $3u_i' - 1 \sim$ iid Bernoulli(1/2), the same argument applies. □

Proof of Proposition 1. (i) We show each point is a stationary point by checking the vector update form of (10), (9), (11). Similar to Theorem 1, we have

$$\theta^{10} = 4t(s_1\mathbf{1}_{C_1} + s_2\mathbf{1}_{C_2} + r_i^{(0)})$$

where $r_i^{(0)} = O_P(\sqrt{n\rho_n \log n})$. Plugging $u^{(0)} = \mathbf{1}_n$ in (9), $s_1 = s_2 = 0.5(\frac{p+q}{2} - \lambda)n$. Similarly

$$\begin{aligned}\theta^{01} =& 4t(x_1\mathbf{1}_{C_1} + x_2\mathbf{1}_{C_2} + r_i^{(1)})\\ \theta^{11} =& 4t(y_1\mathbf{1}_{C_1} + y_2\mathbf{1}_{C_2} + r_i^{(1)})\end{aligned}$$

where $x_1 = x_2 = 0.5(\frac{p+q}{2} - \lambda)n$, $y_1 = y_2 = (\frac{p+q}{2} - \lambda)n$. Plugging in (12) with $\frac{p+q}{2} - \lambda = \Omega_P(\rho_n)$ by Lemma A.1, we have

$$\phi_i^{(1)} = 1 - \exp(-\Omega_P(n\rho_n)), \quad \xi_i^{(1)} = 1 - \exp(-\Omega_P(n\rho_n))$$

for all $i \in [m]$. Hence for sufficiently large n, $u^{(0)} = \mathbf{1}_n$ is the stationary point. For $u^{(0)} = \mathbf{0}_n$, similarly we have

$$\phi_i^{(1)} = \exp(-\Omega_P(n\rho_n)), \quad \xi_i^{(1)} = \exp(-\Omega_P(n\rho_n))$$

so $u^{(0)} = \mathbf{0}_n$ is also a stationary point for large n.

(ii) The statement for $u^{(0)} = \mathbf{0}_n$ and $u^{(0)} = \mathbf{1}_n$ follows from Corollary 1 by $\mu = 0$ and $\mu = 1$.

□

Proof of Proposition 2. Let $\hat{t}, \hat{\lambda}$ be constants defined in terms of $\hat{p}, \hat{q}$. First we observe using $\hat{p}, \hat{q}$ only replaces t, λ with $\hat{t}, \hat{\lambda}$ everywhere in the updates of Algorithm 1. We can check the analysis in Theorem 1 remains unchanged as long as

$$\text{i) } \frac{p+q}{2} > \hat{\lambda}, \quad \text{ii) } \hat{\lambda} - q = \Omega(\rho_n), \quad \text{iii) } \hat{t} = \Omega(1)$$

□

Proof of Theorem 2. Starting with $p^{(0)}$ and $q^{(0)}$ satisfying the conditions in Corollary 2, after two meta iterations of u updates, we have $\|u_3^{(2)} - z^*\|_1 = n\exp(-\Omega(n\rho_n))$. Updating $p^{(1)}, q^{(1)}$ with (14), we first analyze the population version of the numerator of $p^{(1)}$,

$$\begin{aligned}&(\mathbf{1}_n - u)^T P(\mathbf{1}_n - u) + u^T P u + 2(\mathbf{1}_m - \psi^{10} - \psi^{01})^T \mathrm{diag}(P^{zy})\mathbf{1}_m\\ =&(\mathbf{1}_n - z^*)^T P(\mathbf{1}_n - z^*) + (z^*)^T P z^* - 2(u - z^*)^T P(\mathbf{1}_n - z^*) + 2(z^*)^T P(u - z^*)\\ &+ (u - z^*)^T P(u - z^*) + O(n\rho_n).\end{aligned}$$

In the case of $u_3^{(2)}$, the above becomes

$$\frac{n^2}{2}p + O_P(n^{5/2}\rho_n \exp(-\Omega(n\rho_n))) + O(n\rho_n) = \frac{n^2}{2}p + O_P(n\rho_n).$$

Next we can rewrite the noise as

$$\begin{aligned}&(\mathbf{1}_n - u)^T (A - P)(\mathbf{1}_n - u) + u^T (A - P)u\\ =&(\mathbf{1}_n - z^*)^T (A - P)(\mathbf{1}_n - z^*) + (z^*)^T (A - P)z^* - 2(u - z^*)^T (A - P)(\mathbf{1}_n - z^*)\\ &+ 2(z^*)^T (A - P)(u - z^*) + (u - z^*)^T (A - P)(u - z^*).\end{aligned}$$

Similarly in the case of $u_3^{(2)}$, the above is $O_P(\sqrt{n^2\rho_n})$. Therefore the numerator of $p^{(1)}$ is $\frac{n^2}{2}p + O_P(\sqrt{n^2\rho_n})$. To lower bound the denominator, note that

$$\begin{aligned}&u^T(J - I)u + (\mathbf{1} - u)^T (J - I)(\mathbf{1} - u)\\ =&\left(\sum_i u_i\right)^2 + \left(n - \sum_i u_i\right)^2 - u^T u - (\mathbf{1} - u)^T(\mathbf{1} - u)\\ \geq& n^2/2 - 2n,\end{aligned}$$

then we have $p^{(1)} = p + O_P(\sqrt{\rho_n}/n)$. The same analysis shows $q^{(1)} = q + O_P(\sqrt{\rho_n}/n)$.

Replacing p and q with $p^{(1)}$ and $q^{(1)}$ in the final analysis after the second meta iteration of Theorem 1 does not change the order of the convergence, and the rest of the arguments can be repeated. □

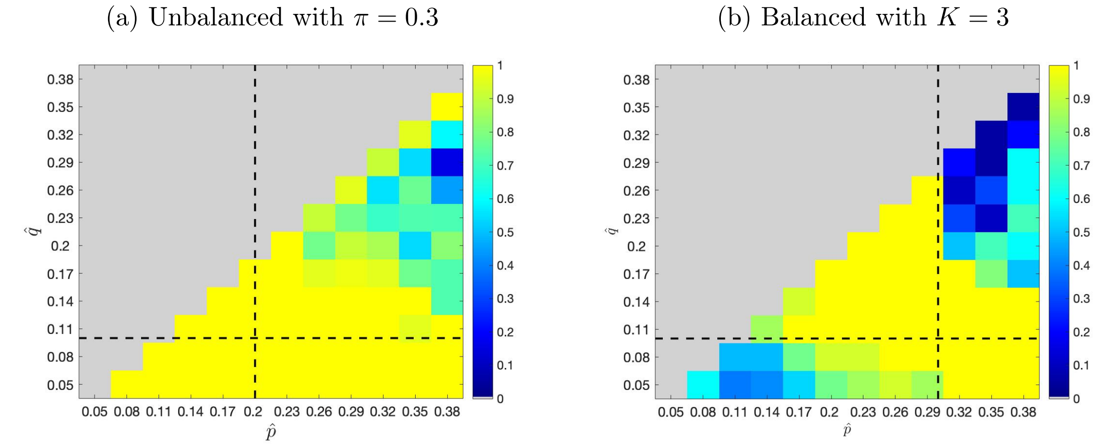
C Generalizations

We present the **update equations for balanced** $K > 2$ models. We will use the notation $a, b \in \{0, \dots, K-1\}$ to be consistent with the two class case. Let $S_{zy} = 2t(\mathrm{diag}(A^{zy}) - \lambda I)\mathbf{1}_m$.

$$\theta^{ab} = \begin{cases} 2t[A^{zz} - \lambda(J-I)](\phi_a - \phi_0) + 2t[A^{zy} - \lambda(J-I) - \mathrm{diag}(A^{zy})](\xi_a - \xi_0) - S_{zy}, & a \neq 0, b = 0 \\ 2t[A^{zz} - \lambda(J-I)](\phi_b - \phi_0) + 2t[A^{zy} - \lambda(J-I) - \mathrm{diag}(A^{zy})](\xi_b - \xi_0) - S_{zy}, & a = 0, b \neq 0 \\ \theta_{a0} + \theta_{b0} + S_{zy} & a \neq 0, b \neq 0 \end{cases} \tag{C.1}$$

The **update equations for unbalanced two class blockmodels** simply adds an additional term of $\log \pi/(1-\pi)$ to the updates of θ_{10} (Eq. (9)), θ_{01} (Eq. (10)) and $2\log \pi/(1-\pi)$ to θ_{11} (Eq. (11)). We assume that the proportions are known.

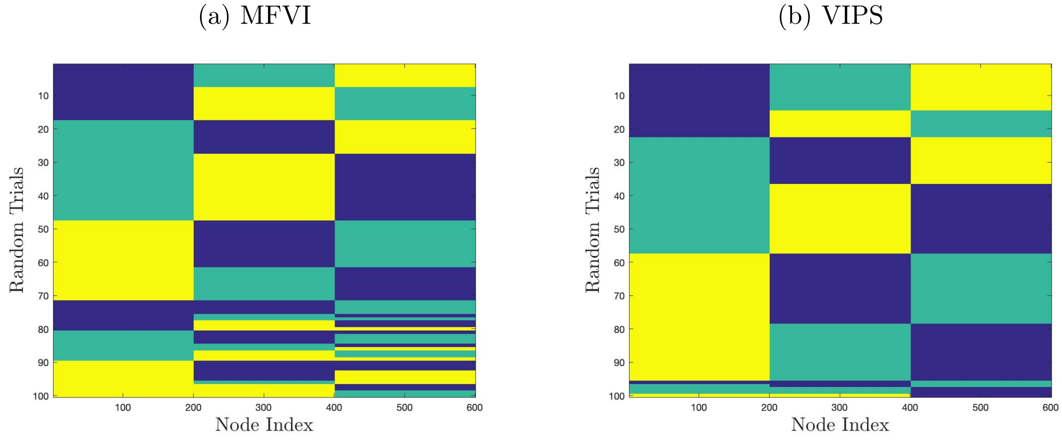
In Figure A.1, we show the heatmap for mis-specified parameters for VIPS on unbalanced SBM ($\pi = .3$) and balanced SBM with $K = 3$. For each starting point of $\hat{p}, \hat{q}$ the average NMI is shown. We see that in both cases the VIPS algorithm converges to the correct labels for a wide range of initial parameter settings.


(a) Unbalanced with $\pi = 0.3$

(b) Balanced with $K = 3$

Figure A.1: NMI with different estimation of $\hat{p}$, $\hat{q}$ with $\hat{p} > \hat{q}$, averaged over 20 random initializations for each $\hat{p}$, $\hat{q}$. The left figure has $\pi = 0.3, K = 2$ and the right figure has balanced clusters with $K = 3$. The true $(p_0, q_0) = (0.2, 0.1)$ and $n = 2000$.

For $K = 3$, we also show Figure A.2, where each row represents the estimated membership of one random trial and both MFVI and VIPS are run with the true p_0, q_0. We show VIPS can recover true membership with higher probability than MFVI.


(a) MFVI

(b) VIPS

Figure A.2: Compare VIPS and MFVI when K=3, equal sized communities, for known p_0, q_0 in 100 random trials. $p_0 = 0.5, q_0 = 0.01$. Rows permuted for visual clarity.

D Additional Experimental Results

In Figure A.3, we compare different update rules in VIPS with (i) parameters p, q fixed at the true values (True), (ii) (p, q) estimated using $(\sum_{i \neq j} A_{ij}/(n(n-1)), \sum_{i \neq j} A_{ij}/(2n(n-1)))$ but fixed (Estimate), and (iii) (p, q) initialized as in (ii) and updated in the algorithm (Update) using Eq. (14). In all settings, VIPS successfully converges to the ground truth, which is consistent with our theoretical results and shows robustness of the parameter setting.

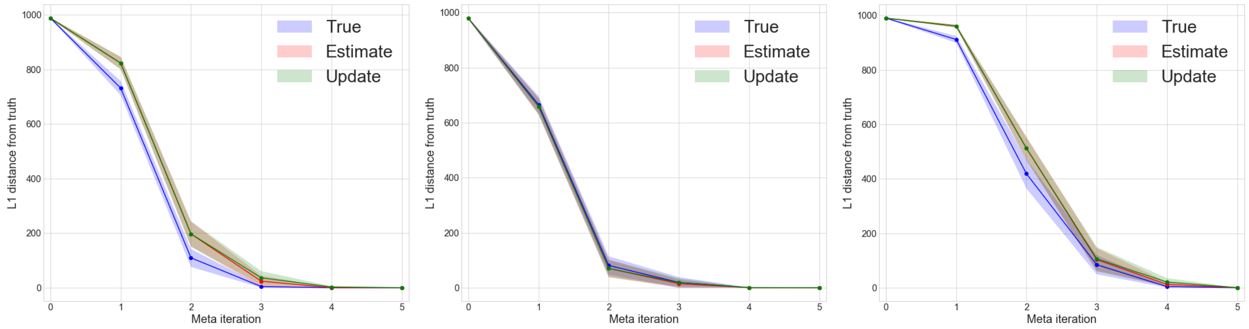


Figure A.3: Values of $\|u - z^*\|_1$ as the number of meta iterations increases. Each line is the mean curve of 50 random trials and the shaded area is the standard deviation. Here $n = 2000$ and $p_0 = 0.1, q_0 = 0.02$. u is initialized by Bernoulli distribution with mean $\mu = 0.1, 0.5, 0.9$ from the left to right.

In Figure A.4, we compare VIPS and MFVI with and without parameter updates. For VIPS, we do parameter updates from 3rd meta iteration onward, and for fairness, we start parameter updates 9 iterations onward for MFVI. In both schemes, the VIPS performs better than MFVI.

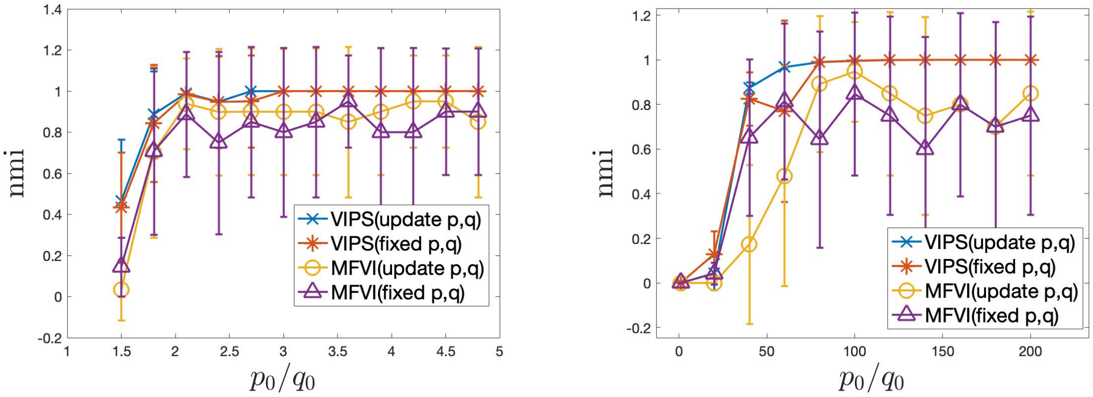


Figure A.4: Two schemes for estimating model parameters for VIPS and MFVI. Both use the initial $\hat{p}$ and $\hat{q}$ as described in Figure 4 in the main paper. The first scheme starts updating $\hat{p}$ and $\hat{q}$ after 3 meta iterations for VIPS and 9 iterations for MFVI. The other scheme has $\hat{p}, \hat{q}$ held fixed.

References

Paul Erdös. On a lemma of littlewood and offord. *Bulletin of the American Mathematical Society*, 51(12):898–902, 1945.

Purnamrita Sarkar, YX Wang, and Soumendu Sunder Mukherjee. When random initializations help: a study of variational inference for community detection. *arXiv preprint arXiv:1905.06661*, 2019.