Journal of Machine Learning Research 22 (2021) 1-46 Submitted 7/19; Revised 9/20; Published 2/21

When random initializations help: a study of variational inference for community detection

Purnamrita Sarkar[*†] PURNA.SARKAR@AUSTIN.UTEXAS.EDU
Department of Statistics and Data Science
University of Texas, Austin
Austin, TX 78712, USA

Y. X. Rachel Wang[*†] RACHEL.WANG@SYDNEY.EDU.AU
School of Mathematics and Statistics
University of Sydney
NSW 2006, Australia

Soumendu Sundar Mukherjee[†] SOUMENDU041@GMAIL.COM
Interdisciplinary Statistical Research Unit (ISRU)
Indian Statistical Institute, Kolkata
Kolkata 700108, India

Editor: Bert Huang

Abstract

Variational approximation has been widely used in large-scale Bayesian inference recently, the simplest kind of which involves imposing a mean field assumption to approximate complicated latent structures. Despite the computational scalability of mean field, theoretical studies of its loss function surface and the convergence behavior of iterative updates for optimizing the loss are far from complete. In this paper, we focus on the problem of community detection for a simple two-class Stochastic Blockmodel (SBM) with equal class sizes. Using batch co-ordinate ascent (BCAVI) for updates, we show different convergence behavior with respect to different initializations. When the parameters are known or estimated within a reasonable range and held fixed, we characterize conditions under which an initialization can converge to the ground truth. On the other hand, when the parameters need to be estimated iteratively, a random initialization will converge to an uninformative local optimum.

Keywords: Variational Approximation, Stochastic Blockmodels, Batch Co-ordinate Ascent, Local Optima

1. Introduction

Variational approximation has recently gained a huge momentum in contemporary Bayesian statistics (Jordan et al., 1999; Blei et al., 2003; Jaakkola and Jordon, 1999). Mean field is the simplest type of variational approximation, and is a popular tool in large scale Bayesian inference. It is particularly useful for problems which involve complicated latent structure, so that direct computation with the likelihood is not feasible. The main idea of variational

∗. Equal contribution.
†. All authors contributed equally to the short version of the paper that appeared in NeurIPS 2018 (Mukherjee et al. (2018)).

©2021 Purnamrita Sarkar, Y. X. Rachel Wang and Soumendu S. Mukherjee.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v22/19-630.html.

approximation is to obtain a tractable lower bound on the complete log-likelihood of any model. This is, in fact, akin to the Expectation Maximization algorithm (Dempster et al., 1977), where one obtains a lower bound on the marginal log-likelihood function via the expectation with respect to the conditional distribution of the latent variables under the current estimates of the underlying parameters. In contrast, for mean field variational approximation, the lower bound or ELBO is computed using the expectation with respect to a product distribution over the latent variables. The Kullback-Leibler divergence is used to measure how well the product distribution approximates the true posterior.

While there are many advances in developing new mean field type approximation methods for Bayesian models, the theoretical behavior of these algorithms is not well understood. There is one line of theoretical work that studies the asymptotic consistency of variational inference, most of which focuses on the global optimizer of variational methods under specific models. For example, for Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Gaussian mixture models, it is shown in Pati et al. (2018) that the global optimizer is statistically consistent. Westling and McCormick (2019) connects variational estimators to profile M-estimation, and shows consistency and asymptotic normality of those estimators. For Stochastic Blockmodels (SBM) (Holland et al., 1983; Hofman and Wiggins, 2008), Bickel et al. (2013) shows that the global optimizer of the variational log-likelihood is consistent and asymptotically normal. For more general cases, Wang and Blei (2019) proves a variational Bernstein-von Mises theorem, which states that the variational posterior converges to the Kullback-Leibler minimizer of a normal distribution, centered at the truth.

Recently, a lot more effort is being directed towards understanding the statistical convergence behavior of non-convex algorithms in general. For Gaussian mixture models (GMM) and exponential families with missing data, Wang and Titterington (2004, 2006) prove local convergence to the true parameters. The same authors also show that the covariance matrix from variational Bayesian approximation for the GMM is "too small" compared with that obtained for the maximum likelihood estimator (Wang and Titterington, 2005). Wu et al. (2012) propose a variational Bayes algorithm based on component splitting for fitting GMM and show in simulation that random intializations converge to the ground truth for a simple two component setting. The robustness of variational Bayes estimators is further discussed in Giordano et al. (2018). For LDA, Awasthi and Risteski (2015) shows that, with proper initialization, variational inference algorithms converge to the global optimum.

In this paper, we will focus on the community detection problem in networks under SBM. Here the latent structure involves unknown community memberships and as a result, the data likelihood requires summing over all possible community labels. Optimization of the likelihood involves a combinatorial search, and thus is infeasible for large-scale graphs. The mean field approximation has been used popularly for this task (Blei et al., 2017; Zhang and Zhou, 2017). In Bickel et al. (2013), it is proved that the global optimum of the mean field approximation to the likelihood behaves optimally in the dense degree regime, where the average expected degree of the network grows faster than the logarithm of the number of vertices. In Zhang and Zhou (2017), it is shown that if the initialization of mean field is close enough to the truth then one gets convergence to the truth at the minimax rate. However, in practice, it is usually not possible to initialize like that unless one uses a pilot algorithm. Most initialization techniques like spectral clustering (Rohe et al., 2011; Ng et al., 2002) will

return correct clustering in the dense degree regime, thus rendering the need for mean field updates redundant.

Indeed, in many practical scenarios, without prior knowledge one simply uses multiple random initializations, the efficacy of which is model-dependent. In order to understand the behavior of random initializations, one needs to first better understand the landscape of the mean field loss. There are few such studies for non-convex optimization in the literature; notable examples include (Mei et al., 2018; Ghorbani et al., 2018; Jin et al., 2016; Xu et al., 2016). In (Xu et al., 2016), the authors fully characterize the landscape of the likelihood of the equal proportion Gaussian Mixture Model with two components, where the main message is that most random initializations should indeed converge to the ground truth. In contrast, for topic models, it has been established that, for some parameter regimes, variational inference exhibits instability and returns a posterior mean that is uncorrelated with the truth (Ghorbani et al., 2018). In this respect, for network models, there has not been much work characterizing the behavior of the variational loss surface.

In this article, in the context of a specific SBM, we give a complete characterization of all the critical points and establish the behavior of random initializations for batch coordinate ascent (BCAVI) updates for mean field likelihood (with known and unknown model parameters). Our results thus complement those of Zhang and Zhou (2017). For simplicity, we work with equal-sized two-class stochastic blockmodels. When the parameters are known, we show conditions under which random initializations can converge to the ground truth. In particular, we show that centering random initializations around a half ensures convergence happens a good fraction of time, and this property holds even if we only have access to reasonable estimates of true parameters. We also analyze the setting with unknown model parameters, where they are estimated jointly with the community memberships. In this case, we see that indeed, with high probability, a random initialization never converges to the ground truth, thus showing the critical importance of a good initialization for network models.

2. Setup and preliminaries

The stochastic blockmodel (SBM), proposed by Holland et al. (1983) in social science, is one of the most popular random graph models incorporating community structures. A SBM with parameters (B, Z, π) is a generative model of networks with community structure on n nodes. Its dynamics is as follows: there are K communities $\{1, \ldots, K\}$ and each node belongs to a single community, where this membership is captured by the rows of the $n \times K$ matrix Z, where the ith row of Z, i.e. $Z_{i,\cdot}$, is the community membership vector of the ith node and has a $\texttt{Multinomial}(1; \pi)$ distribution, independently of the other rows. Given the community structure, links between pairs of nodes are determined solely by the block memberships of the nodes in an independent manner. That is, if A denotes the adjacency matrix of the network, then given Z, A_{ij} and A_{kl} are independent for $(i, j) \neq (k, l)$, $i < j$, $k < l$, and

$$\mathbb{P}(A_{ij} = 1 \mid Z) = \mathbb{P}(A_{ij} = 1 \mid Z_{ia} = 1, Z_{jb} = 1) = B_{ab}.$$

$B = ((B_{ab}))$ is called the block (or community) probability matrix. We have the natural restriction that B is symmetric for undirected networks.

The block memberships are hidden variables and one only observes the network in practice. The goal often is to fit an appropriate SBM to learn the community structure, if any, and also estimate the parameters B and π.

The complete likelihood for the SBM is given by

$$\mathbb{P}(A, Z; B, \pi) = \prod_{i<j} \prod_{a,b} (B_{ab}^{A_{ij}} (1 - B_{ab})^{1-A_{ij}})^{Z_{ia} Z_{jb}} \prod_i \prod_a \pi_a^{Z_{ia}}. \tag{1}$$

As Z is not observable, if we integrate out Z, we get the data likelihood

$$\mathbb{P}(A; B, \pi) = \sum_{Z \in \mathcal{Z}} \mathbb{P}(A, Z; B, \pi), \tag{2}$$

where $\mathcal{Z}$ is the space of all $n \times K$ matrices with exactly one 1 in each row.

In principle we can optimize the data likelihood to estimate B and π. However, $\mathbb{P}(A; B, \pi)$ involves a sum over a complicated large finite set (the cardinality of this set is K^n), and hence is not easy to deal with. A well-known alternative approach is to optimize the variational log-likelihood (Bickel et al., 2013), which has a less complicated dependency structure, the simplest of which is mean field log-likelihood (see, e.g., (Wainwright and Jordan, 2008)). We defer a detailed discussion of the mean field principle in the Appendix.

For the SBM, the variational log-likelihood with respect to a distribution ψ is given by

$$\sum_Z \log \left(\frac{\mathbb{P}(A, Z; B, \pi)}{\psi(Z)} \right) \psi(Z) = \mathbb{E}_\psi \left(\sum_{i<j,a,b} Z_{ia} Z_{jb} (\theta_{ab} A_{ij} - f(\theta_{ab})) \right) - \mathtt{KL}(\psi || \pi^{\otimes n}),$$

where $\theta_{ab} = \log \left(\frac{B_{ab}}{1 - B_{ab}} \right)$, $f(\theta) = \log(1 + e^\theta)$ and $\pi^{\otimes n}$ denotes the product measure on $\mathcal{Z}$ with the rows of Z being i.i.d. $\mathtt{Multinomial}(1; \pi)$. A special case of the variational log-likelihood is the mean field log-likelihood (see, e.g., (Wainwright and Jordan, 2008)), where one approximates Ψ by

$$\Psi_{MF} \equiv \{\psi : \psi(z_1, \ldots, z_n) = \prod_{j=1}^n \psi_j(z_j)\}. \tag{3}$$

Define $\ell_{MF}(\psi, \theta, \pi) = \sum_{i<j,a,b} \psi_{ia} \psi_{jb} (\theta_{ab} A_{ij} - f(\theta_{ab})) - \sum_i \mathtt{KL}(\psi_i || \pi)$. For SBM the mean field approximation is equivalent to optimizing $\ell_{MF}(\psi, \theta, \pi)$ as follows:

$$\begin{aligned} &\max_\psi \ell_{MF}(\psi, \theta, \pi) \\ \text{subject to} &\sum_a \psi_{ia} = 1, \text{ for all } 1 \le i \le n \\ &\psi_{ia} \ge 0, \text{ for all } 1 \le i \le n, 1 \le a \le K, \end{aligned}$$

where each ψ_i is a discrete probability distribution over $\{1, \ldots, K\}$.

2.1 Mean field updates for a two-parameter two-block SBM

Consider the stochastic blockmodel with two blocks with prior block probability $\pi, 1 - \pi$ respectively and block probability matrix $B = (p - q)I + qJ$, where $p > q$, I is the identity

matrix, and $J = \mathbf{1}\mathbf{1}^\top$ is the matrix of all 1's. For simplicity, we will denote ψ_{i1} as ψ_i. Then the mean field log-likelihood is

$$\begin{aligned}\ell(\psi, p, q, \pi) = \frac{1}{2}\sum_{i,j:i\neq j}[\psi_i(1-\psi_j)+\psi_j(1-\psi_i)][A_{ij}\log\left(\frac{q}{1-q}\right)+\log(1-q)] \\ + \frac{1}{2}\sum_{i,j:i\neq j}[\psi_i\psi_j+(1-\psi_i)(1-\psi_j)][A_{ij}\log\left(\frac{p}{1-p}\right)+\log(1-p)] \\ - \sum_i[\log\left(\frac{\psi_i}{\pi}\right)\psi_i + \log\left(\frac{1-\psi_i}{1-\pi}\right)(1-\psi_i)]. \qquad (4)\end{aligned}$$

For simplicity of exposition, we will assume that π (which is essentially a prior on the block memberships) is known and equals 1/2. Let $\mathcal{C}_i, i = 1, 2$ be the two communities. Let $\tilde{\pi} = \frac{|\mathcal{C}_1|}{n}$. It is clear that $\tilde{\pi} = \frac{1}{2} + O_P(\frac{1}{\sqrt{n}})$. Assuming $\tilde{\pi} = \frac{1}{2}$ from the start will not change our conclusions but make the algebra a lot nicer, which we do henceforth. Now

$$\begin{aligned}\frac{\partial \ell}{\partial \psi_i} &= \frac{1}{2}\sum_{j:j\neq i} 2[1-2\psi_j][A_{ij}\log\left(\frac{q}{1-q}\right)+\log(1-q)] \\ &\quad + \frac{1}{2}\sum_{j:j\neq i} 2[2\psi_j-1][A_{ij}\log\left(\frac{p}{1-p}\right)+\log(1-p)] - \log\left(\frac{\psi_i}{1-\psi_i}\right) \\ &= 4t\sum_{j:j\neq i}(\psi_j - \frac{1}{2})(A_{ij}-\lambda) - \log\left(\frac{\psi_i}{1-\psi_i}\right),\end{aligned}$$

where $t = \frac{1}{2}\log\left(\frac{p(1-q)}{q(1-p)}\right)$ and $\lambda = \frac{1}{2t}\log\left(\frac{1-q}{1-p}\right)$. Detailed calculations of other first and second order partial derivatives are given in Section B of the Appendix. The co-ordinate ascent (CAVI) updates for ψ are

$$\log\frac{\psi_i^{(new)}}{1-\psi_i^{(new)}} = 4t\sum_{j\neq i}(\psi_j - \frac{1}{2})(A_{ij}-\lambda).$$

Introducing an intermediate variable ξ for the updates, let $f(x) = \log(\frac{x}{1-x})$ and $\xi_i = f(\psi_i)$. Then at iteration s, given the current values of p and q for computing t and λ, the batch version (BCAVI) of this is

$$\xi^{(s)} = 4t(A - \lambda(J-I))(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}),$$

and $\psi^{(s)} = g(\xi^{(s)})$, where g is the sigmoid function $g(x) = 1/(1+e^{-x})$.

We will study these updates in two setttings: i) when the true model parameters p_0, q_0 are known (or estimated and kept fixed), and ii) when the model parameters p_0, q_0 need to be jointly estimated with ψ. The detailed BCAVI updates for each setting will be described in Section 3. A summary of notations used in the model description and BCAVI updates so far is provided in Table 1.

Notation	Definition
A	Adjacency matrix
B	$B = (p-q)I + qJ$ is the block probability matrix
P	$P = ZBZ^T$, where Z is the $n \times 2$ membership matrix
ψ	n-dimensional mean field parameters
ξ	Intermediate variable in the updates, $\xi_i = \log\left(\frac{\psi_i}{1-\psi_i}\right)$
t, λ	$t = \frac{1}{2}\log\left(\frac{p(1-q)}{q(1-p)}\right)$ and $\lambda = \frac{1}{2t}\log\left(\frac{1-q}{1-p}\right)$.
p_0, q_0, t_0, λ_0	True model parameters and their related quantities

Table 1: Notations used in the two-parameter two-block SBM and BCAVI updates.

3. Main results

In this section, we state and discuss our main results. All the proofs appear in the Appendix.

We begin with introducing some notations. In the following, we will see the following vectors repeatedly: $\psi = \frac{1}{2}\mathbf{1}, \mathbf{1}, \mathbf{0}, \mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}$. Among these, $\mathbf{1}$ corresponds to the case where every node is assigned by ψ to $\mathcal{C}_1$, and, similarly, for $\mathbf{0}$, to $\mathcal{C}_2$. On the other hand, $\mathbf{1}_{\mathcal{C}_i}$ are the indicators of the clusters $\mathcal{C}_i$ and hence correspond to the ground truth community assignment. Finally, $\frac{1}{2}\mathbf{1}$ corresponds to the solution where a node belong to each community with equal probability.

The next propositions show some useful inequalities for t and λ computed from general p and q.

Proposition 1 *Suppose* $1 > p > q > 0$. *Then*

1. $\frac{(p-q)(1+p-q)}{2(1-q)p} < t < \frac{(p-q)(1-p+q)}{2(1-p)q}$, *and*
2. $q < \lambda < p$.

The next proposition refines the separation between λ and p, q, when $p \asymp q \asymp \rho_n, \rho_n \to 0$.

Proposition 2 *If* $p \asymp q \asymp \rho_n, \rho_n \to 0$ *and* $p - q = \Omega(\rho_n)$, *then*

$$\lambda - q = \Omega(\rho_n) > 0, \tag{5}$$

$$\frac{p+q}{2} - \lambda = \Omega(\rho_n) > 0. \tag{6}$$

3.1 Known p_0, q_0:

In this case, denoting the true model parameters p_0, q_0 ($p_0 > q_0$), we assume these parameters are known and thus need only consider the updates for ψ. We consider the case where the true p_0, q_0 are of the same order, that is, $p_0 \asymp q_0 \asymp \rho_n$ with ρ_n possibly going to 0. The BCAVI updates are:

$$\xi^{(s+1)} = 4t_0(A - \lambda_0(J - I))(\psi^{(s)} - \frac{1}{2}\mathbf{1}), \tag{7}$$

where t_0 and λ_0 are calculated using p_0 and q_0. In what follows, we will also study the population version of this update which replaces A by $\mathbb{E}(A \mid Z) = ZBZ^\top - p_0 I =: P - p_0 I$. Hence for convenience, denote $M := P - p_0 I - \lambda_0(J - I)$. The population BCAVI updates are

$$\xi^{(s+1)} = 4t_0 M(\psi^{(s)} - \frac{1}{2}\mathbf{1}). \tag{8}$$

The eigendecomposition of $P - \lambda_0 J$ will play a crucial role in our analysis. Note that it has rank two and two eigenvalues $n\alpha_\pm$, where $\alpha_+ = \frac{p_0+q_0}{2} - \lambda_0, \alpha_- = \frac{p_0-q_0}{2}$, with eigenvectors $\mathbf{1}$ and $\mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}$ respectively. Now it can be easily checked that the eigenvalues of M are $\nu_1 = n\alpha_+ - (p_0 - \lambda_0)$, $\nu_2 = n\alpha_- - (p_0 - \lambda_0)$ and $\nu_j = -(p_0 - \lambda_0)$, $j = 3, \ldots, n$. The eigenvector of M corresponding to ν_1 is $u_1 = \mathbf{1}$, and the one corresponding to ν_2 is $u_2 = \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}$.

We first present a proposition related to the landscape of the objective function. Consider the population mean field log-likelihood, which replaces A by its expectation $\mathbb{E}(A|Z)$ in Eq (4). In the known p_0, q_0 case, $\frac{1}{2}\mathbf{1}$ is a saddle point of the population mean field log-likelihood.

Proposition 3 *$\psi = \frac{1}{2}\mathbf{1}$ is a saddle point of the population mean field log-likelihood when p_0 and q_0 are known, for all n large enough.*

We next give conditions on the initialization which determine their convergence behavior when using the population BCAVI (8). To facilitate our discussion, we will write the BCAVI updates in the eigenvector coordinates of M. To this end, define $\zeta_i^{(s)} = \langle \psi^{(s)}, u_i \rangle / \|u_i\|^2 = \langle \psi^{(s)}, u_i \rangle / n$, for $i = 1, 2$. We can then write

$$\psi^{(s)} = \langle \psi^{(s)}, u_1/\|u_1\| \rangle u_1/\|u_1\| + \langle \psi^{(s)}, u_2/\|u_2\| \rangle u_2/\|u_2\| + v^{(s)} = \zeta_1^{(s)} u_1 + \zeta_2^{(s)} u_2 + v^{(s)}. \tag{9}$$

So, using (8) in conjunction with the above decomposition, coordinate-wise we have:

$$\begin{aligned} \xi_i^{(s+1)} &= 4t_0 n \left((\zeta_1^{(s)} - \frac{1}{2})\alpha_+ + \sigma_i \zeta_2^{(s)} \alpha_- \right) + 4t_0 \nu_3 \left((\zeta_1^{(s)} - \frac{1}{2}) + \sigma_i \zeta_2^{(s)} + v_i^{(s)} \right) \\ &=: n a_{\sigma_i}^{(s)} + b_i^{(s)}, \end{aligned} \tag{10}$$

where $\sigma_i = 1$, if i is in $\mathcal{C}_1$, and -1 otherwise. Note that the mean-field parameters are obtained by passing $\xi_i^{(s+1)}$ elementwise through a sigmoid. So, in order to converge to the ground truth, say $\mathbf{1}_{\mathcal{C}_1}$, we hope that $\xi_i^{(s+1)}$ goes to positive infinity for nodes in $\mathcal{C}_1$, and $\xi_i^{(s+1)}$ goes to negative infinity for nodes in $\mathcal{C}_2$. In Eq (10), in the first iteration, $\xi_i^{(1)}$ is dominated by $n a_{\sigma_i}^{(0)}$. In other words, if $|n a_{\sigma_i}^{(0)}| \to \infty$, and $a_{+1}^{(0)}$ and $a_{-1}^{(0)}$ are of opposite signs, we expect $\psi^{(1)}$ to converge to the ground truth. If they are of the same sign, then $\psi^{(1)}$ should converge to $\mathbf{1}$ or $\mathbf{0}$. The next theorem gives a more rigorous statement of this. In Table 3 we enumerate the different limits for different signs of $a_{\sigma_i}^{(0)}$.

A summary of main notations used in our analysis can be found in Table 2.

Theorem 4 (Population behavior) *The limit behavior of the population BCAVI updates* (8) *is characterized by the signs of $a_{\pm 1}^{(0)}$, where $a_{\pm 1}^{(s)}$ for iteration s is defined in* (10). *Assume that $|n a_{\pm 1}^{(0)}| \to \infty$. Define $\ell(\psi^{(0)}) = \mathbb{1}(a_{+1}^{(0)} > 0)\mathbf{1}_{\mathcal{C}_1} + \mathbb{1}(a_{-1}^{(0)} > 0)\mathbf{1}_{\mathcal{C}_2}$, where $\mathbb{1}(\cdot)$*

Notation	Definition
ρ_n	Average density of the network, $p_0 \asymp q_0 \asymp \rho_n$
$\alpha_\pm$	$\alpha_+ = \frac{p_0+q_0}{2} - \lambda_0, \alpha_- = \frac{p_0-q_0}{2}$
$\mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}$	$\mathbf{1}_{\mathcal{C}_i}$ are the indicators of the cluster $\mathcal{C}_i$, $i = 1, 2$
M	$M = P - p_0 I - \lambda_0(J - I)$
$\nu_1, \ldots, \nu_n$	Eigenvalues of M, $\nu_1 = n\alpha_+ - (p_0 - \lambda_0)$, $\nu_2 = n\alpha_- - (p_0 - \lambda_0)$, $\nu_j = -(p_0 - \lambda_0)$, $j = 3, \ldots, n$
u_1, u_2, v	u_1, u_2 are eigenvectors of M, $u_1 = \mathbf{1}$, $u_2 = \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}$ v is orthogonal to u_1, u_2, defined in Eq (9).
ζ_1, ζ_2	$\zeta_i = \langle \psi, u_i \rangle / n$, $i = 1, 2$
$a_{\pm 1}^{(s)}$	Defined in Eq (10).

Table 2: Notations used in the analysis.

Signs of $a_{+1}^{(0)}, a_{-1}^{(0)}$	Stationary point $\ell(\psi^{(0)})$
$a_{+1}^{(0)} > 0, a_{-1}^{(0)} > 0$	$\mathbf{1}$
$a_{+1}^{(0)} < 0, a_{-1}^{(0)} < 0$	$\mathbf{0}$
$a_{+1}^{(0)} > 0, a_{-1}^{(0)} < 0$	$\mathbf{1}_{\mathcal{C}_1}$
$a_{+1}^{(0)} < 0, a_{-1}^{(0)} > 0$	$\mathbf{1}_{\mathcal{C}_2}$

Table 3: $\ell(\psi^{(0)})$ describes four stationary points depending on the sign of $a_{\pm 1}^{(0)}$.

denotes an indicator function of the event (see Table 3 for all cases of $\ell(\psi^{(0)})$). Then, under the same assumption on p_0, q_0 in Proposition 2, we have

$$\frac{\|\psi^{(1)} - \ell(\psi^{(0)})\|^2}{n} = O(\exp(-\Theta(n \min\{|a_{+1}^{(0)}|, |a_{-1}^{(0)}|\}))) = o(1).$$

We also have for any $s \geq 2$

$$\frac{\|\psi^{(s)} - \ell(\psi^{(0)})\|^2}{n} = \begin{cases} O(\exp(-\Theta(nt_0\alpha_-))), & \text{if } a_{+1}^{(0)} a_{-1}^{(0)} < 0, \\ O(\exp(-\Theta(nt_0\alpha_+)), & \text{if } a_{+1}^{(0)} a_{-1}^{(0)} > 0. \end{cases}$$

Remark 5 *1. Note that $\ell(\psi^{(0)})$ describes 4 stationary points characterized by the signs of $a_{\pm 1}^{(0)}$, which are calculated from $\psi^{(0)}$: $\mathbf{1}, \mathbf{0}, \mathbf{1}_{\mathcal{C}_1}$, and $\mathbf{1}_{\mathcal{C}_2}$. We enumerate these in Table 3. The theorem explains which stationary point the population BCAVI converges to is determined by the signs of $a_{\pm 1}^{(0)}$.*

2. Since the proof of Theorem 4 shows BCAVI can only converge to one of the four points $\{\mathbf{1}, \mathbf{0}, \mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}\}$ starting from any given $\psi^{(0)}$ satisfying the condition in the theorem, there are only five stationary points of the mean field log-likelihood, namely $\mathbf{1}, \mathbf{0}, \mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}$, and the saddle point $\frac{1}{2}\mathbf{1}$ in Proposition 3.

3. *We see from Theorem 4 that, essentially, we have exponential convergence within two iterations.*

4. *Since $p_0 \asymp q_0 \asymp \alpha_+ \asymp \rho_n$, as long as one of the projections $\zeta_1^{(0)} - 1/2$, $\zeta_2^{(0)}$ of the initialization $\psi^{(0)}$ is non-vanishing (of order $\Theta(1)$), the condition $|na_{\pm 1}^{(0)}| \to \infty$ requires $n\rho_n \to \infty$.*

From Theorem 4, we can calculate lower bounds on the volumes of the basins of attractions of the limit points of the population BCAVI updates. We have the following corollary.

Corollary 6 *Define the set of initialization points converging to a stationary point $\mathbf{c}$ as*

$$\mathcal{S}_{\mathbf{c}} := \{v \mid \limsup_{s\to\infty} n^{-1}\|\psi^{(s)} - \mathbf{c}\|^2 = O(\exp(-\Theta(nt_0 \min\{|\alpha_+|, \alpha_-\}))), \text{ when } \psi^{(0)} = v\}.$$

Let $\mathfrak{M}$ be some measure on $[0,1]^n$, absolutely continuous with respect to the Lebesgue measure. Consider the stationary point $\mathbf{1}$, then

$$\mathfrak{M}(\mathcal{S}_{\mathbf{1}}) \geq \lim_{\gamma\uparrow 1} \mathfrak{M}(H_+^\gamma \cap H_-^\gamma \cap [0,1]^n),$$

where the half-spaces $H_\pm^\gamma$ are given as

$$H_\pm^\gamma = \big\{x \mid \langle x, \alpha_+ u_1 \pm \alpha_- u_2\rangle > \frac{n\alpha_+}{2} + \frac{n^{1-\gamma}}{4t}\big\}.$$

Similar formulas can be obtained for the other stationary points.

For specific measures $\mathfrak{M}$, one can obtain explicit formulas for these volumes. In practice, these are quite easy to calculate by Monte Carlo simulations.

Now we turn to the sample behavior of the updates in (7).

Theorem 7 (Sample behavior) *For all $s \geq 1$, the same conclusion as Theorem 4 holds for the sample BCAVI updates in (7) with probability at least $1 - n^{-r}, r > 0$, as long as $n|a_{\pm 1}^{(0)}| \gg \max\{\sqrt{n\rho_n \log n}\|\psi^{(0)} - \frac{1}{2}\|_\infty, 1\}$, $n\rho_n = \Omega(\log n)$ and $\psi^{(0)}$ is independent of A.*

Remark 8 *Since $\|\psi^{(0)} - \frac{1}{2}\|_\infty = O(1)$, we can check that the lower bound required on $n|a_\pm^{(0)}|$ by Theorem 7 always holds when we use initializations of the form $\psi_i^{(0)} \overset{iid}{\sim} f_\mu$, where f_μ is some distribution with support $[0,1]$, mean μ and $\mu \neq \frac{1}{2}$. Here $n|a_{\pm 1}^{(0)}| = \Theta_P(n\rho_n)$ and we already have $n\rho_n = \Omega(\log n)$. When $\mu = \frac{1}{2}$, $n|a_\pm^{(0)}| = O_P(\sqrt{n}\rho_n)$ which does not satisfy the lower bound. In this case, we have the following theorem showing convergence can happen for a good fraction of the random initializations.*

Theorem 9 (Convergence for random initializations) *When p_0 and q_0 are known and $\rho_n \to 0$ at a rate such that $\rho_n\sqrt{n}/\log n \to \infty$, initializing with $\psi_i^{(0)} \sim$ iid Bernoulli$(\frac{1}{2})$ and using the sample BCAVI updates (7), with probability at least $1 - \frac{\arctan(c_\ell) - \arctan(c_\ell^{-1})}{\pi} - o(1)$,*

$$\|\psi^{(s)} - z_0\|_1 \leq n\exp(-C_1 t_0 (p_0 - q_0)n) + \frac{C_2\rho_n}{(p_0 - q_0)^2 n}\|\psi^{(s-1)} - z_0\|_1$$

for $s \geq 3$, some general constants C_1, C_2 (independent of n and model parameters), and $z_0 = \mathbf{1}_{\mathcal{C}_1}$ or $\mathbf{1}_{\mathcal{C}_2}$. Here

$$c_\ell = \frac{(p_0 - \lambda_0) + c(\lambda_0 - q_0)}{c(p_0 - \lambda_0) + (\lambda_0 - q_0)}, \qquad c = \frac{(\lambda_0 - q_0)(1 - \epsilon_n)}{(p_0 - \lambda_0)(1 + \epsilon_n)} - \eta,$$

$\epsilon_n \to 0$ slowly such that $\rho_n \sqrt{n} \epsilon_n / |\log \rho_n| \to \infty$ and $\eta > 0$ is some arbitrarily small constant.

Remark 10 *1. Note that the convergence probability can also be written as $\frac{1}{2} + \frac{2 \arctan(c_\ell^{-1})}{\pi} - o(1)$, which is larger than $1/2$. The distance between this lower bound probability and 1 decreases as $|c - 1|$ decreases.*

2. Thus Theorem 9 shows that random initializations lead to convergence to the global optima (i.e. the ground truth, $\mathbf{1}_{\mathcal{C}_1}$ and $\mathbf{1}_{\mathcal{C}_2}$) of the variational objective function with probability strictly greater than half. This means that one can do N independent random initializations, and with probability greater than $1 - (1/2)^N$, at least one of the initializations will converge to the ground truth. To see this statement is valid, we note that even though each initialization uses the same data matrix A to obtain the estimates, the bounds on $\psi^{(0)}$ and A used in our proof are completely separable.

3. With multiple random initializations, the best clustering can be picked by finding one with the largest ELBO, since it is a well-known fact that the ground truth maximizes the ELBO (e.g., Bickel et al. (2013) Eq. (3) and Lemma 3). This justifies the common practice of using multiple random starts and picking the result with the largest ELBO. Hence it is important to note the key here is that the success probability of a random initialization is lower bounded by a constant.

4. We can also obtain mis-clustering rate directly from the L_1 norm bound. For every iteration s, let $\hat{z}_i^{(s)} = \mathbb{1}(\psi_i^{(s)} > 1/2)$ be the estimated labels. Then the mis-clustering rate is given by

$$\begin{aligned} \frac{\|\hat{z}^{(s)} - z_0\|_0}{n} &\leq \frac{2\|\psi^{(s)} - z_0\|_1}{n} \\ &\leq 2\,(1 + o(1)) \exp(-C_1 t_0 (p_0 - q_0) n) + 2 \left(\frac{C_2 \rho_n}{(p_0 - q_0)^2 n} \right)^{s-2}. \end{aligned}$$

The next corollary shows that even if we do not know p_0 and q_0 and only have their estimates, the above convergence still holds as long as the estimates are reasonably close to p_0 and q_0.

Corollary 11 (Using parameter estimates) *The same conclusion as in Theorem 9 holds if we replace p_0, q_0 with some $\hat{p}, \hat{q} \asymp \rho_n$, $|\hat{p} - \hat{q}| = \Omega(\rho_n)$, satisfying*

1. $\frac{p_0 + q_0}{2} > \hat{\lambda}$,
2. $\hat{\lambda} - q_0 = \Omega(\rho_n) > 0$,

where $\hat{\lambda}$ is computed using $\hat{p}$ and $\hat{q}$.

Remark 12 *1. In practice, $\hat{p}$, $\hat{q}$ can be estimates depending on A, then the statement in Corollary 11 still holds.*

2. *When $\hat{p}, \hat{q} \asymp \rho_n$, $\hat{p} - \hat{q} = \Omega(\rho_n) > 0$, $\hat{\lambda}$ lies between $(\hat{p}+\hat{q})/2$ and $\hat{q}$ as suggested by Proposition 2. The conditions in Corollary 11 imply an upper bound on $\hat{p}$ and a lower bound on $\hat{q}$. Similar constraints hold if $\hat{q} - \hat{p} = \Omega(\rho_n) > 0$. An example of the estimate regime is shown in Figure 1, where $p_0 = 0.3$, $q_0 = 0.1$, and the yellow area contains $\hat{p}$, $\hat{q}$ such that $\frac{p_0+q_0}{2} > \hat{\lambda} > q_0$.*

3. *Such estimates can be obtained by applying any strongly consistent SDP method to a smaller subgraph, which makes it computationally efficient as well. Consider randomly sampling $\sqrt{n}$ nodes from the original graph. This subgraph has average degree $\sqrt{n}\rho_n \to \infty$ under the setting of Theorem 9, and class size of the order $\Theta(\sqrt{n})$. Then applying a strongly consistent SDP method, such as Li et al. (2018), one can achieve exact recovery of community labels on this subgraph with high probability. $\hat{p}, \hat{q}$ are obtained by simply averaging the edge counts, and using Bernstein's inequality $|\hat{p} - p_0|, |\hat{q} - q_0| = O(\sqrt{\frac{\rho_n}{n}\log n}) \ll \rho_n$ with high probability.*

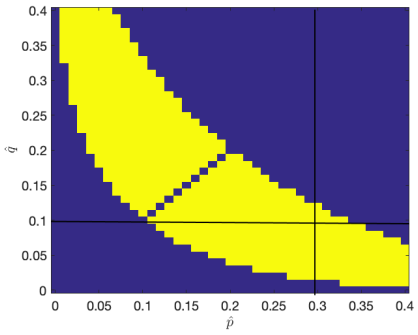


Figure 1: For $p_0 = 0.3$, $q_0 = 0.1$, the yellow area shows where $\frac{p_0+q_0}{2} > \hat{\lambda} > q_0$ is satisfied.

3.2 Unknown p_0, q_0:

In this case, the model parameters p and q are updated jointly with ψ. The full BCAVI updates are

$$p^{(s)} = \frac{(\psi^{(s-1)})^\top A \psi^{(s-1)} + (\mathbf{1} - \psi^{(s-1)})^\top A (\mathbf{1} - \psi^{(s-1)})}{(\psi^{(s-1)})^\top (J - I)\psi^{(s-1)} + (\mathbf{1} - \psi^{(s-1)})^\top (J - I)(\mathbf{1} - \psi^{(s-1)})}, \tag{11}$$

$$q^{(s)} = \frac{(\psi^{(s-1)})^\top A (\mathbf{1} - \psi^{(s-1)})}{(\psi^{(s-1)})^\top (J - I)(\mathbf{1} - \psi^{(s-1)})},$$

$$t^{(s)} = \frac{1}{2}\log\left(\frac{p^{(s)}(1-q^{(s)})}{q^{(s)}(1-p^{(s)})}\right), \ \lambda^{(s)} = \frac{1}{2t^{(s)}}\log\left(\frac{1-q^{(s)}}{1-p^{(s)}}\right),$$

$$\xi^{(s)} = 4t^{(s)}(A - \lambda^{(s)}(J - I))(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}).$$

Similar to before, $p_0 \asymp q_0 \asymp \rho_n$ with ρ_n possibly going to 0. In the population version, we would replace A with $\mathbb{E}(A \mid Z) = P - pI$.

In this case with unknown p_0, q_0, our next result shows that $\frac{1}{2}\mathbf{1}$ changes from a saddle point (Proposition 3) to a local maximum.

Proposition 13 *Let $n \geq 2$. Then $(\psi, p, q) = (\frac{1}{2}\mathbf{1}, \frac{\mathbf{1}^\top A \mathbf{1}}{n(n-1)}, \frac{\mathbf{1}^\top A \mathbf{1}}{n(n-1)})$ is a strict local maximum of the mean field log-likelihood.*

Since p_0, q_0 and ψ are unknown and need to be estimated iteratively, we have the following updates for $p^{(1)}$ and $q^{(1)}$ given the initialization $\psi^{(0)}$ and show that they can be written in terms of the projection of the initialization in the principal eigenspace of P.

Lemma 14 *Let $x = (\psi^{(0)})^T\psi^{(0)} + (\mathbf{1} - \psi^{(0)})^T(\mathbf{1} - \psi^{(0)})$ and $y = 2(\psi^{(0)})^T(\mathbf{1} - \psi^{(0)}) = n - x$. Projecting $\psi^{(0)}$ onto u_1 and u_2 and writing $\psi^{(0)} = \zeta_1 u_1 + \zeta_2 u_2 + w$, where $w \in \text{span}\{u_1, u_2\}^\perp$, then*

$$\begin{aligned} p^{(1)} &= \frac{p_0 + q_0}{2} + \frac{(p_0 - q_0)(\zeta_2^2 - x/2n^2)}{\zeta_1^2 + (1 - \zeta_1)^2 - x/n^2} + O_P(\sqrt{\rho_n}/n), \\ q^{(1)} &= \frac{p_0 + q_0}{2} - \frac{(p_0 - q_0)(\zeta_2^2 + y/2n^2)}{2\zeta_1(1 - \zeta_1) - y/n^2} + O_P(\sqrt{\rho_n}/n). \end{aligned} \tag{12}$$

Since $(\psi^{(0)})^T(\mathbf{1} - \psi^{(0)}) > 0$, we have $\zeta_1(1 - \zeta_1) \geq \zeta_2^2$. This gives:

$$p^{(1)} \in \left(\frac{p_0 + q_0}{2} + O_P(\sqrt{\rho_n}/n), p_0\right], \qquad q^{(1)} \in \left[q_0, \frac{p_0 + q_0}{2} + O_P(\sqrt{\rho_n}/n)\right). \tag{13}$$

It is interesting to note that $p^{(1)}$ is always smaller than $q^{(1)}$ except when it is $O(\sqrt{\rho_n}/n)$ close to $(p_0 + q_0)/2$. In that regime, one needs to worry about the sign of t and λ. In all other regimes, t, λ are positive.

Using the update forms in Lemma 14, the following result shows that the stationary points of the population mean field log-likelihood lie in the principle eigenspace $\text{span}\{u_1, u_2\}$ of P in a limiting sense.

Proposition 15 *Consider the case with unknown p_0, q_0 and $\rho_n \to 0$, $n\rho_n \to \infty$. Let $(\psi, \tilde{p}, \tilde{q})$ be a stationary point of the population mean field log-likelihood. If $\psi = \psi_u + \psi_{u^\perp}$, where $\psi_u \in span\{u_1, u_2\}$ and $\psi_{u^\perp} \perp span\{u_1, u_2\}$, then $\|\psi_{u^\perp}\| = o(\sqrt{n})$ as $n \to \infty$.*

We next present the two main results of this section, which analyze the convergence of the full BCAVI (Eq (11)) updates with respect to different types of initializations. We first consider a simple random initialization, where the entries of $\psi^{(0)}$ are i.i.d with mean μ. In this case, ζ_2 is vanishing, which is unsurprising since ζ_2 measures correlation with the second eigenvector of P, u_2 which is the $\mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}$ vector. Then by Lemma 14, $p^{(1)}$ and $q^{(1)}$ concentrates around the average of the conditional expectation matrix, i.e. $(p_0 + q_0)/2$. In this case, the update converges to $\frac{1}{2}\mathbf{1}$ with small deviations within one update as stated in the next theorem. This result shows the futility of random initialization when p_0, q_0 are unknown, in contrast to the results in Section 3.1.

Theorem 16 *Consider the initial distribution $\psi_i^{(0)} \overset{iid}{\sim} f_\mu$ where f is a distribution supported on $(0,1)$ with mean μ. If μ is bounded away from 0 and 1 and $n\rho_n = \Omega(\log n)$, using the updates in (11), then $\|\psi^{(1)} - \frac{1}{2}\mathbf{1}\|_2 = O_P(1)$,*

$$\|\psi^{(s)} - \frac{1}{2}\mathbf{1}\|_2 \leq O_P(1/\sqrt{n})\|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 + O_P(\rho_n^{3/2})$$

for $s \geq 2$.

As another type of initialization, it is also instructive to analyze the case where the initialization is in fact correlated with the truth. To this end, we will consider a initialization scheme, $\psi_i^{(0)} = \mu_i + \epsilon_i^{(0)}$, where the expectation of $\psi_i^{(0)}$ is μ_i, $\epsilon_i^{(0)}$ are independent zero mean noise such that the support of $\psi_i^{(0)}$ is $[0,1]$. In this case, provided there is sufficient separation between the cluster means, defined as

$$\Delta\mu = \frac{\sum_{i\in\mathbf{1}_{\mathcal{C}_1}} \mu_i}{n} - \frac{\sum_{i\in\mathbf{1}_{\mathcal{C}_2}} \mu_i}{n}, \tag{14}$$

we have convergence to ground truth within one iteration.

Theorem 17 *Consider the initialization $\psi^{(0)} = \mu_i + \epsilon_i^{(0)}$ such that $E[\psi_i^{(0)}] = \mu_i$, $\epsilon_i^{(0)}$ are independent and $\max_i Var(\psi_i^{(0)}) < \infty$. Assume $\frac{1}{n}\sum_{i=1}^n \mu_i = 1/2$ (WLOG) and $n\rho_n = \Omega(\log n)$. Then provided*

$$|\Delta\mu| = \Omega\left(\frac{\rho_n^{3/2}\sqrt{\log n}}{(p_0-q_0)^2\sqrt{n}}\right)^{1/3}, \tag{15}$$

we have for large enough n, with probability at least $1 - \exp(-\Theta(\log n))$, $\psi^{(1)} = \mathbf{1}_{\mathcal{C}_1} + O(\exp(-\Omega(\sqrt{n\rho_n \log n})))$ or $\mathbf{1}_{\mathcal{C}_2} + O(\exp(-\Omega(\sqrt{n\rho_n \log n})))$, where the error term is uniform for all the coordinates. For $s \geq 2$,

$$\|\psi^{(s)} - z_0\|_1 \leq n\exp(-c_1 t_0(p_0-q_0)n) + \frac{c_2\rho_n}{n(p_0-q_0)^2}\|\psi^{(s-1)} - z_0\|_1$$

for some general constants c_1, c_2 (independent of n and model parameters), with probability at least $1 - n^{-r}$, $r > 0$, uniformly for all s.

Remark 18 *1. The lemma states that provided the separation between p_0 and q_0 does not vanish too fast, and there is enough separation between the two cluster means, we have converge to the truth within one iteration. In the case of $p_0 - q_0 \asymp \rho_n$, the theorem requires $|\Delta\mu| = \Omega\left(\left(\frac{\log n}{n\rho_n}\right)^{1/6}\right)$. The lower bound can approach 0 when $n\rho_n/\log n \to \infty$. Thus in this special case, our constraint on the initialization is weaker than Zhang and Zhou (2017), which requires $\|\psi^{(0)} - z_0\|_1 \leq c_{init} n/2$ for some sufficiently small constant c_{init} under a balanced two-block model.*

2. *Consider randomly sampling a subset of nodes $\mathcal{S}$ from the graph, with $|\mathcal{S}| = \Theta\left(n \cdot \left(\frac{\log n}{n\rho_n}\right)^{1/6}\right)$.*

   If we initialize $\psi_i^{(0)}$ with the correct labels for $i \in \mathcal{S}$ (which can be done by applying a strongly consistent SDP algorithm like Li et al. (2018) to $\mathcal{S}$), and $\psi_i^{(0)}$ iid randomly for $i \notin \mathcal{S}$, the separation condition on $\Delta\mu$ would be satisfied. We show this also with a simulation in Section 4, Figure 4 (a).

4. Numerical results

In Figure 2-(a), we have generated a network from an SBM with parameters $p_0 = 0.4, q_0 = 0.025$, and two equal sized blocks of 100 nodes each. We generate 5000 initializations $\psi^{(0)}$ from $\text{Beta}(\alpha, \beta)^{\otimes n}$ (for four sets of α and β) and map them to $a_{\pm 1}^{(0)}$. We perform sample BCAVI updates on $\psi^{(0)}$ with known p_0, q_0 and color the points in the $a_{\pm 1}^{(0)}$ co-ordinates according the limit points they have converged to. In this case, $\alpha_+ > 0$, hence based on Theorems 4 and 7, we expect points with $a_{+1}^{(0)} a_{-1}^{(0)} < 0$ to converge to the ground truth (colored green or magenta) and those with $a_{+1}^{(0)} a_{-1}^{(0)} > 0$ to converge to $\mathbf{0}$ or $\mathbf{1}$. As expected, points falling in the center of the first and third quadrants have converged to $\mathbf{0}$ or $\mathbf{1}$. The points converging to the ground truth lie more toward the boundaries but mostly remain in the same quadrants, suggesting possible perturbations arising from the sample noise and small network size. We see that this issue is alleviated when we increase n.

The notable thing is, in Figure 2-(a) and (d), the Beta distribution has mean 0.16 and 0.71 respectively. So the initialization is more skewed towards values that are closer to zero or closer to one. In these cases most of the random runs converge to the all zeros or all ones, with very few converging to the ground truth. However, for Figure 2-(b) and (d), the mean of the Beta is 0.3 and 0.7, and we see considerably more convergences to the ground truth. Also, (b) and (d) are, in some sense, mirror images of each other, i.e. in one, the majority converges to $\mathbf{0}$; whereas in the other, the majority converges to $\mathbf{1}$.

In Figure 3, we examine whether convergence can hold even when the exact values of p_0, q_0 are unknown using the initiliazation scheme in Theorem 9 and Corollary 11. In each heatmap, the dashed lines indicate the true parameter values used to generate an adjacency matrix A. The heatmap contains pairs of $\hat{p}, \hat{q}$ that we use in the sample BCAVI updates (7) for fixed parameters initialized with $\psi_i^{(0)} \sim$ iid Bernoulli($\frac{1}{2}$). For each pair of parameters, we use 50 such random initializations and compute the average clustering accuracy. In both cases, we can see that as long as the parameter estimates fall into a reasonable range around the true values, convergence to the ground truth happens for a high fraction of the random initializations. The plots are symmetric in terms of $\hat{p}$ and $\hat{q}$, suggesting the estimates do not have to respect the relationship $\hat{p} > \hat{q}$ as discussed in Remark 12.

In Figure 4 (a), we examine initializations of the type described in Theorem 17 and the resulting estimation error. In particular, we provide correct labels for a set of nodes $\mathcal{S}$ and set $\psi_i^{(0)}, i \in \mathcal{S}$ at those correct labels, and then initialize the rest of the nodes at random. We show our results for three settings of p_0, q_0, with $n = 500$. On the Y axis we plot the classification accuracy over 50 random runs across the $|\mathcal{S}|/n$ on the X axis. We see the surprising result that for the highest separation only 10% of labeled nodes can result in

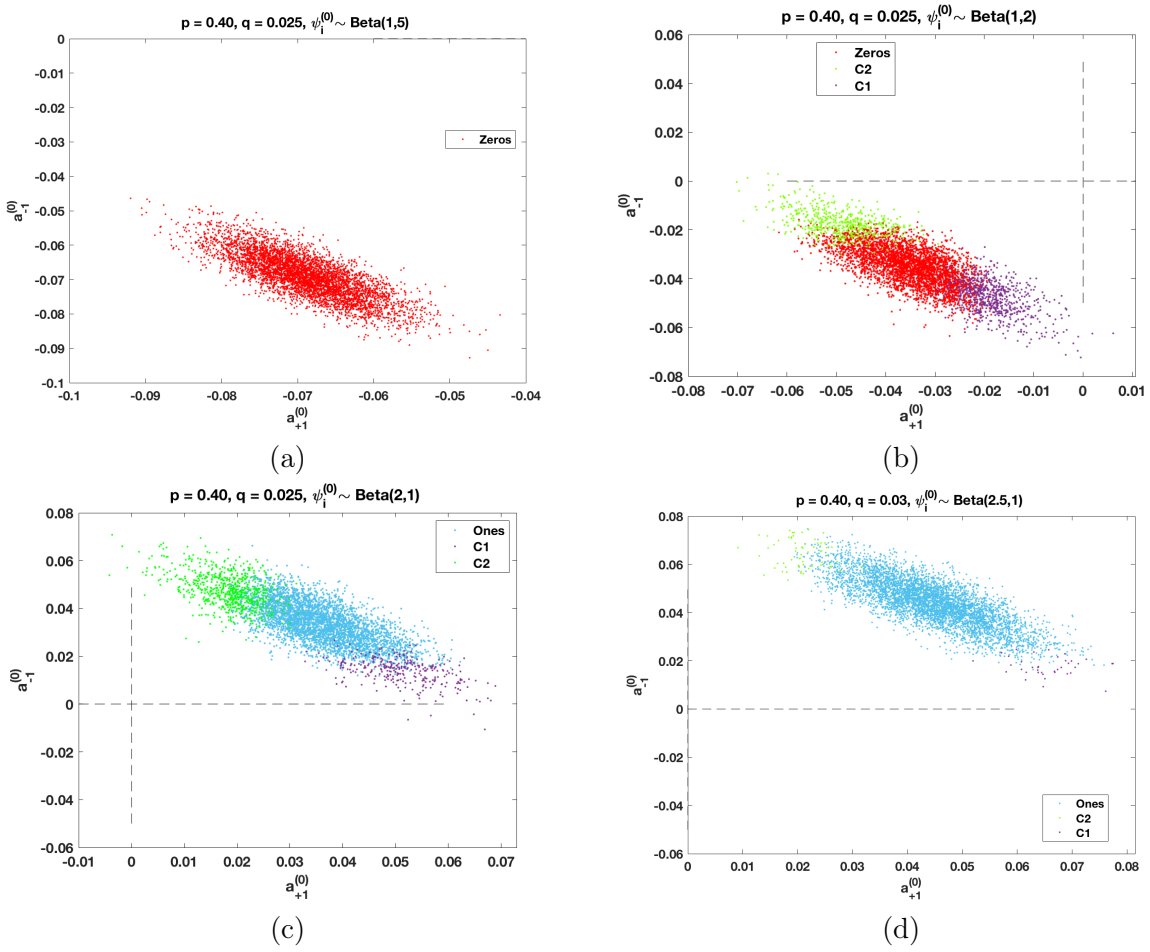
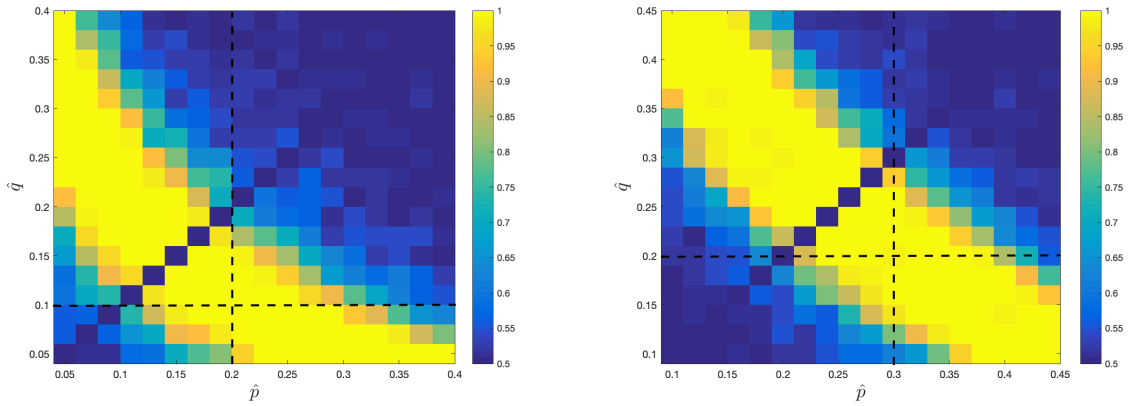


Figure 2: $n = 200$ and 5000, $\psi^{(0)} \sim \text{Beta}(\alpha, \beta)^{\otimes n}$ for various values of α and β. These $\psi^{(0)}$ are mapped to $(a_{+1}^{(0)}, a_{-1}^{(0)})$ (see (10)) and plotted. C_1 (magenta) and C_2 (green) correspond to the limit points $\mathbf{1}_{\mathcal{C}_1}$ and $\mathbf{1}_{\mathcal{C}_2}$. Other limit points are 'Ones', i.e. $\mathbf{1}$ (blue) and 'Zeros', i.e. $\mathbf{0}$ (red).

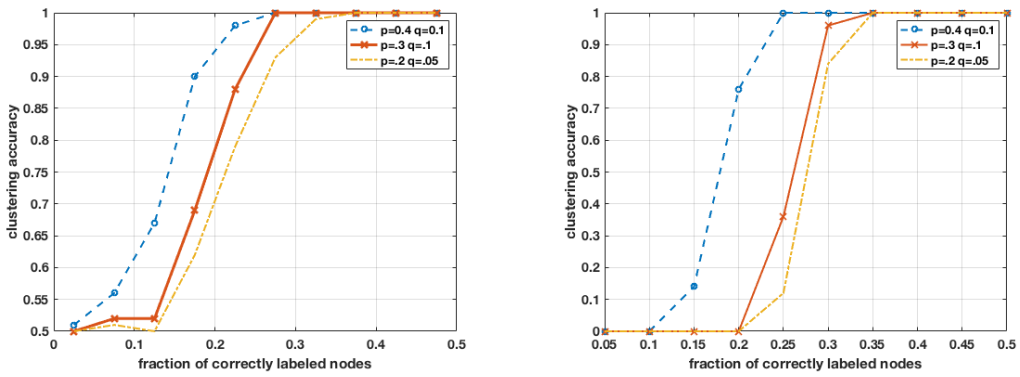
better than random classification, whereas for about 20% correctly labeled nodes, the average accuracy is better than 90%.

In addition, we compare the performance of the random initialization scheme in Theorem 9 with other more informative initializations obtained from running spectral clustering (Rohe et al., 2011) and semi-definite programming (SDP, Li et al. (2018)). As expected, spectral clustering and SDP given better initializations than random and lead to higher accuracy, specially on sparse graphs. Nonetheless, overall random initializations yield very reasonable results over a range of p_0/q_0 values and for moderately sparse graphs. Details of the experiments and results can be found in Appendix Section D.



(a) (b)

Figure 3: Average clustering accuracy using 50 random initializations $\psi_i^{(0)} \sim$ iid Bernoulli($\frac{1}{2}$) and different $\hat{p}, \hat{q}$ values in the BCAVI updates with fixed parameters. The dashed lines show the true parameter values, (a) $p_0 = 0.2$, $q_0 = 0.1$, (b) $p_0 = 0.3$, $q_0 = 0.2$.



(a) (b)

Figure 4: Random initializations with small subset labeled correctly for a graph of size $n = 400$ with (a) 2 equal sized blocks and (b) 3 equal sized blocks. X axis is fraction of correctly clustered nodes, and Y axis is average accuracy.

5. Discussion

In this paper, we work with the BCAVI mean field variational algorithm for a simple two class stochastic blockmodel with equal sized classes. Mean field methods are used widely for their scalability. However, existing theoretical works typically analyze the behavior of the global optima, or the local convergence behavior when initialized near the ground truth. In the simple setting considered, we show two interesting results. First, we show that, when the model parameters are known, random initializations centered around half converge to

the ground truth a good fraction of time. The same convergence holds if some reasonable estimates of the model parameters are known and held fixed throughout the updates. In contrast, when the parameters are not known and estimated iteratively with the mean field parameters, we show that a random initialization converges, with high probability, to a meaningless local optimum. This shows the futility of using multiple random initializations when no prior knowledge is available.

In view of recent works on the optimization landscape for Gaussian mixtures (Jin et al., 2016; Xu et al., 2016), we would like to comment that, despite falling into the category of latent variable models, the SBM has fundamental differences from Gaussian mixtures which require different analysis techniques. The posterior probabilities of the latent labels in the latter model can be easily estimated when the parameters are known, whereas this is not the case for SBM since the posterior probability $\mathbb{P}(Z_i|A)$ depends on the entire network. The significance of the results in Section 3.1 lies in characterizing the convergence of label estimates given the correct parameters for general initializations, which is different from the type of parameter convergence shown in (Jin et al., 2016; Xu et al., 2016). Furthermore, as most of the existing literature for the SBM focuses on estimating the labels first, our results provide an important complementary direction by suggesting that one could start with parameter estimation instead.

While we only show results for two classes, we expect that our main theoretical results generalize well to $K > 2$ and will leave the analysis for future work. As an illustration, consider a setting similar to that of Figure 2 but for $n = 450$ with $K = 3$ equal sized classes. $p_0 = 0.5$, $q_0 = 0.01$ are known and $\psi^{(0)}$ is initialized with a Dirichlet$(0.1, 0.1, 0.1)$ distribution.

We examine the convergence behavior of BCAVI for 1000 random initializations of $\psi^{(0)}$. In Figure 5, each row represents the cluster membership vector a random initialization converges to. We represent the node memberships with three different colors in the columns. The rows have been permuted to group together initializations that converge to the same stationary point. We can see that all 1000 random initializations converge to stationary points lying in the span of $\{\mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}, \mathbf{1}_{\mathcal{C}_3}\}$, which are the membership vectors for each class. There are $1+\binom{3}{2} = 4$ different types of stationary points, not counting class label permutations. Another stationary point (the all ones vector that puts everyone in the same class) can be obtained with other initialization schemes, e.g., when the rows of $\psi^{(0)}$ are identical. For a general K- blockmodel, we conjecture that the number of stationary points grows exponentially with K. Similar to Figure 2, a significant fraction of the random initializations converge to the ground truth when p_0, q_0 are known. On the other hand, when p_0, q_0 are unknown, random initializations always converge to the uninformative stationary point $(1/3, 1/3, 1/3)$, analogous to Theorem 16.

We believe that for a more general SBM, the separation condition in Theorem 17 will be some suitably defined distance from the ground truth, which will be a matrix for three or more blocks ($K \geq 3$). Considering a special case of Theorem 17, if one can obtain correct labels of a n^α, $\alpha \in [0, 1]$, size subset $\mathcal{S}$ of nodes (including all K labels), then an initialization with the nodes in $\mathcal{S}$ fixed at the correct labels, and the rest initialized at random with probability $1/2$, then under suitable conditions on α, we expect BCAVI to converge to the ground truth. We show that this intuition is indeed correct in Fig 4 (b) for $K = 3$. Here we change the size of the random subset of nodes which are initialized at the correct label, and plot average accuracy over fifty random runs on the Y axis. The three lines correspond to

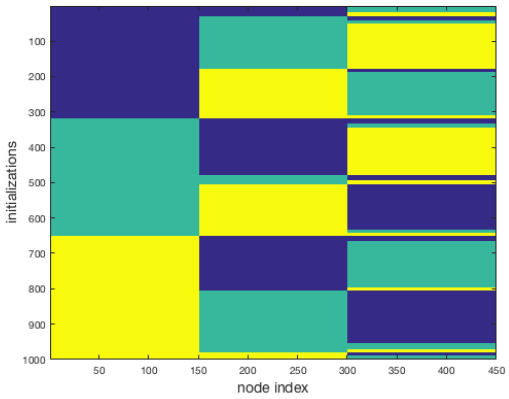


Figure 5: Convergence to stationary points for known p_0, q_0, $K = 3$. Rows permuted for clarity.

different p/q ratio. We see that the same trend holds for both $K = 2$ and 3. But for $K = 3$, we need a larger set of correctly labeled samples to reach the same accuracy. In particular, for the largest p/q ratio, with 20% of correct labels, for $K = 2$, average accuracy is 90% whereas for $K = 3$, the average accuracy is about 75%.

For models beyond SBM, if the model can be expressed with a low rank plus noise decomposition, then we believe that in the high signal to noise setting, the separation condition will be reduced to the amount of correlation of the initialization with principal eigenvectors in the population matrix.

Acknowledgments

SSM thanks Professor Peter J. Bickel for helpful discussions. PS is partially funded by NSF grant DMS1713082. YXRW is supported by the ARC DECRA Fellowship.

Appendix A.

This appendix provides derivation of stationarity equations for the mean field log-likelihood, the proofs of our main results, and some additional simulation results.

Appendix A. The Variational principle and mean field

We start with the following simple observation:

$$\log P(A;B,\pi) = \log \sum_Z P(A,Z;B,\pi) = \log\left(\sum_Z \frac{P(A,Z;B,\pi)}{\psi(Z)}\psi(Z)\right)$$
$$\overset{\text{(Jensen)}}{\geq} \sum_Z \log\left(\frac{P(A,Z;B,\pi)}{\psi(Z)}\right)\psi(Z) \qquad \forall \psi \text{ prob. on } \mathcal{Z}.$$

In fact, equality holds for $\psi^*(Z) = P(Z|A;B,\pi)$. Therefore, if Ψ denotes the set of all probability measures on $\mathcal{Z}$, then

$$\log P(A;B,\pi) = \max_{\psi\in\Psi} \sum_Z \log\left(\frac{P(A,Z;B,\pi)}{\psi(Z)}\right)\psi(Z). \tag{16}$$

The crucial idea from variational inference is to replace the set Ψ above by some easy-to-deal-with subclass Ψ_0 to get a lower bound on the log-likelihood.

$$\log P(A;B,\pi) \geq \max_{\psi\in\Psi_0\subset\Psi} \sum_Z \log\left(\frac{P(A,Z;B,\pi)}{\psi(Z)}\right)\psi(Z). \tag{17}$$

Also the optimal $\psi_\star \in \Psi_0$ is a potential candidate for an estimate of $P(Z|A;B,\pi)$. Estimating $P(Z|A;B,\pi)$ is profitable since then we can obtain an estimate of the community membership matrix by setting $Z_{ia} = 1$ for the ith agent where

$$a = \arg\max_b P(Z_{ib} = 1|A;B,\pi). \tag{18}$$

The goal now has become optimizing the lower bound in (17).

Appendix B. Derivation of stationarity equations

$$\frac{\partial \ell}{\partial \psi_i} = 4t\sum_{j:j\neq i}(\psi_j - \frac{1}{2})(A_{ij}-\lambda) - \log\left(\frac{\psi_i}{1-\psi_i}\right),$$
$$\frac{\partial \ell}{\partial p} = \frac{1}{2}\sum_{i,j:i\neq j}(\psi_i\psi_j + (1-\psi_i)(1-\psi_j))\left(A_{ij}\left(\frac{1}{p}+\frac{1}{1-p}\right) - \frac{1}{1-p}\right),$$
$$\frac{\partial \ell}{\partial q} = \frac{1}{2}\sum_{i,j:i\neq j}(\psi_i(1-\psi_j) + (1-\psi_i)\psi_j)\left(A_{ij}\left(\frac{1}{q}+\frac{1}{1-q}\right) - \frac{1}{1-q}\right). \tag{19}$$

Therefore

$$\frac{\partial^2 \ell}{\partial \psi_j \partial \psi_i} = 4t(A_{ij}-\lambda)(1-\delta_{ij}) - \frac{1}{\psi_i(1-\psi_i)}\delta_{ij},$$

$$\frac{\partial^2 \ell}{\partial \psi_i \partial p} = \frac{1}{2} \sum_{j:j\neq i} \left(\frac{1}{2} - \psi_j\right)\left(A_{ij}\left(\frac{1}{p} + \frac{1}{1-p}\right) - \frac{1}{1-p}\right),$$
$$\frac{\partial^2 \ell}{\partial \psi_i \partial q} = \frac{1}{2} \sum_{j:j\neq i} \left(\psi_i - \frac{1}{2}\right)\left(A_{ij}\left(\frac{1}{q} + \frac{1}{1-q}\right) - \frac{1}{1-q}\right),$$
$$\frac{\partial^2 \ell}{\partial p^2} = \frac{1}{2} \sum_{i,j:i\neq j} (\psi_i\psi_j + (1-\psi_i)(1-\psi_j))\left(A_{ij}\left(-\frac{1}{p^2} + \frac{1}{(1-p)^2}\right) - \frac{1}{(1-p)^2}\right),$$
$$\frac{\partial^2 \ell}{\partial q^2} = \frac{1}{2} \sum_{i,j:i\neq j} (\psi_i(1-\psi_j) + (1-\psi_i)\psi_j)\left(A_{ij}\left(-\frac{1}{q^2} + \frac{1}{(1-q)^2}\right) - \frac{1}{(1-q)^2}\right),$$
$$\frac{\partial^2 \ell}{\partial q \partial p} = 0. \tag{20}$$

Appendix C. Proofs of main results

Proof [Proof of Proposition 1] For any $a > b > 0$, we have

$$\frac{a-b}{a} < \log\left(\frac{a}{b}\right) < \frac{a-b}{b},$$

which can be proved using the inequality $\log(1+x) < x$ for $x > -1, x \neq 0$. Therefore

$$\frac{p-q}{p} < \log\left(\frac{p}{q}\right) < \frac{p-q}{q}, \text{ and } \frac{p-q}{1-q} < \log\left(\frac{1-q}{1-p}\right) < \frac{p-q}{1-p}.$$

So

$$\frac{(p-q)(1+p-q)}{2(1-q)p} < t = \frac{1}{2}\left(\log\left(\frac{p}{q}\right) + \log\left(\frac{1-q}{1-p}\right)\right) < \frac{(p-q)(1-p+q)}{2(1-p)q},$$

and

$$q = \frac{\frac{p-q}{1-q}}{\frac{p-q}{q} + \frac{p-q}{1-q}} < \lambda = \frac{\log(\frac{1-q}{1-p})}{\log(\frac{p}{q}) + \log(\frac{1-q}{1-p})} < \frac{\frac{p-q}{1-p}}{\frac{p-q}{p} + \frac{p-q}{1-p}} = p.$$

■

Proof [Proof of Proposition 2] Let $y = (p-q)/(1-p) > 0$. We will use the well known inequalities (Topsøe, 2004):

$$\log(1+y) \geq \frac{2y}{2+y} \geq \frac{y}{1+y}, \tag{21}$$
$$\log(1+y) \leq y - \frac{y^2}{2(1+y)} \tag{22}$$

Using Eq (22),

$$\lambda = \frac{\log\frac{1-q}{1-p}}{\log\frac{p}{q} + \log\frac{1-q}{1-p}} \geq \frac{y}{(1+y)\log\frac{p}{q} + y} \geq \frac{(p-q)}{\log\frac{p}{q} + (p-q)}$$

Using Eq (21) we get:

$$\begin{aligned}\lambda - q &\geq \frac{(p-q) - q\log(p/q) - O(\rho_n^2)}{\log\frac{p}{q} + (p-q)} \\ &\geq \frac{(p-q) - q\left(\frac{p-q}{q} - \frac{(p-q)^2}{2pq}\right) - O(\rho_n^2)}{\log\frac{p}{q} + (p-q)} \\ &\geq \frac{\frac{(p-q)^2}{2p} - O(\rho_n^2)}{\log\frac{p}{q} + O(\rho_n)} = \Omega(\rho_n)\end{aligned}$$

The last step is true since $p - q = \Omega(\rho_n)$.

Now we prove Eq (6). Let $x := p/q - 1 = \Omega(1)$, since $p - q = \Omega(\rho_n)$.

$$\begin{aligned}\lambda &\leq \frac{p-q}{(1-p)\log\frac{p}{q} + (p-q)} \\ \frac{p+q}{2} - \lambda &\geq \frac{\frac{p+q}{2}\log(p/q) - (p-q) - O(\rho_n^2)}{(1-p)\log\frac{p}{q} + (p-q)} \\ &= q\frac{(1+x/2)\log(1+x) - x - O(\rho_n)}{\log(p/q) + O(\rho_n)} \qquad (23)\end{aligned}$$

Consider the function $h(x)$ defined below, where $x = p/q - 1 = \Omega(1)$.

$$\begin{aligned}h(x) &= (2+x)\log(1+x) - 2x \\ h'(x) &= \log(1+x) + \frac{2+x}{1+x} - 2 = \log(1+x) - \frac{x}{1+x} \\ &\geq \frac{2x}{2+x} - \frac{x}{1+x} = \frac{x^2}{(2+x)(1+x)} = \Omega(1)\end{aligned}$$

Plugging into Eq (23) we get:

$$\frac{p+q}{2} - \lambda \geq q\frac{h(x) - O(\rho_n)}{2\log(p/q) + O(\rho_n)} = \Omega(\rho_n)$$

■

C.1 Proofs of results in Section 3.1

Proof [Proof of Proposition 3] That $\psi = \frac{1}{2}\mathbf{1}$ is a stationary point is obvious from the stationarity equations (19). The eigenvalues of $-4I + 4t_0 M$, the Hessian at $\frac{1}{2}\mathbf{1}$, are $h_i = -4 + 4t_0\nu_i$. We have $\nu_1 = n\alpha_+ - (p_0 - \lambda_0) = \Theta(n)$, and hence so is h_1. Also, $p_0 - \lambda_0 > 0$, so that $\nu_3 < 0$, and hence $h_3 < 0$. Thus we have two eigenvalues of the opposite sign. ■

Proof [Proof of Theorem 4] From (10), we have

$$\psi_i^{(s+1)} = g(na_{\sigma_i}^{(s)} + b_i^{(s)}) = g(na_{\sigma_i}^{(s)}) + \delta_i^{(s)},$$

where $|\delta_i^{(s)}| = O(\exp(-n|a_{\sigma_i}^{(s)}|))$, where we have used the fact that

$$g(nx+y) - g(nx) = g(nx)g(nx+y)(e^y - 1)\exp(-(nx+y)).$$

Writing as a vector, we have

$$\psi^{(s+1)} = g(na_{+1}^{(s)})\mathbf{1}_{\mathcal{C}_1} + g(na_{-1}^{(s)})\mathbf{1}_{\mathcal{C}_2} + \delta^{(s)}, \tag{24}$$

where $\|\delta^{(s)}\|_\infty = \max_i |\delta_i^{(s)}| = O(\exp(-n\min\{|a_{+1}^{(s)}|, |a_{-1}^{(s)}|\}))$. Note that by our assumption, $\|\delta^{(0)}\|_\infty = O(\exp(-n\min\{|a_{+1}^{(s)}|, |a_{-1}^{(s)}|\})) = o(1)$. Now

$$\zeta_1^{(s+1)} = \frac{\langle \psi^{(s+1)}, u_1\rangle}{n} = \frac{g(na_{+1}^{(s)}) + g(na_{-1}^{(s)})}{2} + O(\|\delta^{(s)}\|_\infty),$$

and

$$\zeta_2^{(s+1)} = \frac{\langle \psi^{(s+1)}, u_2\rangle}{n} = \frac{g(na_{+1}^{(s)}) - g(na_{-1}^{(s)})}{2} + O(\|\delta^{(s)}\|_\infty).$$

Note that $g(na_{\pm 1}^{(s)}) = \mathbf{1}_{\{a_{\pm 1}^{(s)} > 0\}} + O(\|\delta^{(s)}\|_\infty)$. Now, using (24),we have

$$\begin{aligned}
&\frac{\|\psi^{(s+1)} - \ell(\psi^{(0)})\|_2^2}{n} \\
&\quad= \frac{\|(g(na_{+1}^{(s)}) - \mathbf{1}_{\{a_{+1}^{(0)}>0\}})\mathbf{1}_{\mathcal{C}_1} + (g(na_{-1}^{(s)}) - \mathbf{1}_{\{a_{-1}^{(0)}>0\}})\mathbf{1}_{\mathcal{C}_2} + \delta^{(s)}\|^2}{n} \\
&\quad\le \frac{2(\|(g(na_{+1}^{(s)}) - \mathbf{1}_{\{a_{+1}^{(0)}>0\}})\mathbf{1}_{\mathcal{C}_1}\|_2^2 + \|(g(na_{-1}^{(s)}) - \mathbf{1}_{\{a_{-1}^{(0)}>0\}})\mathbf{1}_{\mathcal{C}_2}\|_2^2 + \|\delta^{(s)}\|^2)}{n} \\
&\quad\le |g(na_{+1}^{(s)}) - \mathbf{1}_{\{a_{+1}^{(0)}>0\}}|^2 + |g(na_{-1}^{(s)}) - \mathbf{1}_{\{a_{-1}^{(0)}>0\}}|^2 + 2\|\delta^{(s)}\|_\infty^2 \\
&\quad= |\mathbf{1}_{\{a_{+1}^{(s)}>0\}} - \mathbf{1}_{\{a_{+1}^{(0)}>0\}}|^2 + |\mathbf{1}_{\{a_{+1}^{(s)}>0\}} - \mathbf{1}_{\{a_{-1}^{(0)}>0\}}|^2 + O(\|\delta^{(s)}\|_\infty^2).
\end{aligned} \tag{25}$$

From the above representation and our assumption on $n|a_{\pm 1}^{(0)}|$, the bound for $s = 1$ follows. We will now consider the four different cases of different signs of $a_{\pm 1}^{(s)}$.

Case 1: $a_{+1}^{(s)} > 0, a_{-1}^{(s)} > 0$. In this case $g(na_1^{(s)}) = g(na_{-1}^{(s)}) = 1 + O(\|\delta^{(s)}\|_\infty)$, so that

$$(\zeta_1^{(s+1)}, \zeta_2^{(s+1)}) = (1, 0) + O(\|\delta^{(s)}\|_\infty).$$

This implies that

$$a_{\pm 1}^{(s+1)} = 2t_0\alpha_+ + O(\|\delta^{(s)}\|_\infty).$$

Since $\alpha_+ > 0$ by Proposition 2, $a_{\pm 1}^{(s+1)}$ have the same sign as $a_{\pm 1}^{(s)}$. Note that, here and in the subsequent cases, we are using that fact that $\|\delta^{(s)}\|_\infty = o(1)$, for $s = 0$, by our assumption and

it stays the same for $s \geq 1$ because of relations like the above (that is $a_{\pm 1}^{(1)} = -2t_0\alpha_+ + o(1)$, so that $\|\delta^{(1)}\|_\infty = \exp(-n\min\{|a_{+1}^{(1)}|, |a_{-1}^{(1)}|\}) = O(\exp(-Cnt_0\alpha_+)) = o(1)$, and so on).

Case 2: $a_{+1}^{(s)} < 0, a_{-1}^{(s)} < 0$. In this case $1 - g(na_1^{(s)}) = 1 - g(na_{-1}^{(s)}) = 1 + O(\|\delta^{(s)}\|_\infty)$, so that

$$(\zeta_1^{(s+1)}, \zeta_2^{(s+1)}) = (0,0) + O(\|\delta^{(s)}\|_\infty).$$

This implies that

$$a_{\pm 1}^{(s+1)} = -2t_0\alpha_+ + O(\|\delta^{(s)}\|_\infty).$$

Since $\alpha_+ > 0$ by Proposition 2, $a_{\pm 1}^{(s+1)}$ have the same sign as $a_{\pm 1}^{(s)}$.

Case 3: $a_{+1}^{(s)} > 0, a_{-1}^{(s)} < 0$. In this case $g(na_1^{(s)}) = 1 - g(na_{-1}^{(s)}) = 1 + O(\|\delta^{(s)}\|_\infty)$, so that

$$(\zeta_1^{(s+1)}, \zeta_2^{(s+1)}) = (\frac{1}{2}, \frac{1}{2}) + O(\|\delta^{(s)}\|_\infty).$$

This implies that

$$a_{\pm 1}^{(s+1)} = \pm 2t_0\alpha_- + O(\|\delta^{(s)}\|_\infty).$$

Since $\alpha_- > 0$, $a_{\pm 1}^{(s+1)}$ have the same sign as $a_{\pm 1}^{(s)}$.

Case 4: $a_{+1}^{(s)} < 0, a_{-1}^{(s)} > 0$. In this case $1 - g(na_1^{(s)}) = g(na_{-1}^{(s)}) = 1 + O(\|\delta^{(s)}\|_\infty)$, so that

$$(\zeta_1^{(s+1)}, \zeta_2^{(s+1)}) = (\frac{1}{2}, -\frac{1}{2}) + O(\|\delta^{(s)}\|_\infty).$$

This implies that

$$a_{\pm 1}^{(s+1)} = \mp 2t_0\alpha_- + O(\|\delta^{(s)}\|_\infty).$$

Since $\alpha_- > 0$, $a_{\pm 1}^{(s+1)}$ have the same sign as $a_{\pm 1}^{(s)}$.

We conclude that, if $\alpha_+ > 0$, then we stay in the same case where we began.Now the desired conclusion follows from the bound (25).

In the proof above, we can allow sparser graphs, with $p_0, q_0 \gg \frac{1}{n}$. More explicitly, let $p_0 = \rho_n a, q_0 = \rho_n b$, with $a > b > 0$ and $\rho_n \gg \frac{1}{n}$. Then, $t_0 = \Omega(1)$, $nt_0|\alpha_\pm| = \Omega(n\rho_n) \to \infty$. ■

Proof [Proof of Corollary 6]

From Theorem 4, it follows that, when $\alpha_+ > 0$,

$$\begin{aligned}
\mathfrak{M}(\mathcal{S}_1) &\geq \mathfrak{M}(\{\psi^{(0)} \mid a_{+1}^{(0)} > 0, a_{-1}^{(0)} > 0, na_{\pm 1}^{(0)} \gg 1\} \\
&= \mathfrak{M}(\{\psi^{(0)} \mid a_{+1}^{(0)} \gg \frac{1}{n}, a_{-1}^{(0)} \gg \frac{1}{n}\}) \\
&\geq \mathfrak{M}(\{\psi^{(0)} \mid a_{+1}^{(0)} > \frac{1}{n^\gamma}, a_{-1}^{(0)} > \frac{1}{n^\gamma}\}),
\end{aligned}$$

for any $0 < \gamma < 1$ and so on for the other other limit points.

More explicitly,

$$\{\psi^{(0)} \mid a_{+1}^{(0)} > \frac{1}{n^\gamma}, a_{-1}^{(0)} > \frac{1}{n^\gamma}\} = \{\psi^{(0)} \mid (\zeta_1^{(0)} - \frac{1}{2})\alpha_+ + \zeta_2^{(0)}\alpha_- > \frac{1}{4tn^\gamma},$$

$$(\zeta_1^{(0)} - \frac{1}{2})\alpha_+ - \zeta_2^{(0)}\alpha_- > \frac{1}{4tn^\gamma}\}$$
$$= H_+^\gamma \cap H_-^\gamma \cap [0,1]^n,$$

All in all, we have

$$\mathfrak{M}(\mathcal{S}_\mathbf{1}) \geq \lim_{\gamma \uparrow 1} \mathfrak{M}(H_+^\gamma \cap H_-^\gamma \cap [0,1]^n).$$

This completes the proof. ■

Proof [Proof of Theorem 7] We begin by noting that $A - \lambda_0(J-I) - M = A - \mathbb{E}(A|Z) := A - \tilde{P}$. For the first iteration, we rewrite the sample iterations (7) as

$$\xi^{(1)} = 4t_0 M\left(\psi^{(0)} - \frac{1}{2}\mathbf{1}\right) + 4t_0 \underbrace{(A-\tilde{P})\left(\psi^{(0)} - \frac{1}{2}\mathbf{1}\right)}_{=:r^{(0)}}.$$

Therefore, similar to the population case, we have

$$\psi_i^{(1)} = g(na_{\sigma_i}^{(0)} + b_i^{(0)} + 4t_0 r_i^{(0)}).$$

Note that

$$r_i^{(0)} = \sum_{j \neq i} (A_{ij} - \tilde{P}_{ij})(\psi_j^{(0)} - \frac{1}{2}). \tag{26}$$

Since our probability statements will be with respect to the randomness in A and $\psi^{(0)}$ is independent of A, we may assume that $\psi^{(0)}$ is fixed. Let $Y_{ij} = (A_{ij} - \tilde{P}_{ij})(\psi_j^{(0)} - \frac{1}{2})$. Then the Y_{ij} are independent random variables for $j \neq i$, and $\mathbb{E}(Y_{ij}) = 0$. Also, $|Y_{ij}| \leq |\psi_j^{(0)} - \frac{1}{2}| \leq \|\psi^{(0)} - \frac{1}{2}\|_\infty = \Delta$, say, and $\mathbb{E}Y_{ij}^2 = (\psi_j^{(0)} - \frac{1}{2})^2 \mathrm{Var}(A_{ij}) = O(\rho_n(\psi_j^{(0)} - \frac{1}{2})^2)$. So, by Bernstein's inequality,

$$\begin{aligned}
\mathbb{P}(\frac{1}{n}\sum_{j\neq i} Y_{ij} > \epsilon) &\leq \exp\left(\frac{-\frac{1}{2}n^2\epsilon^2}{\sum_{j\neq i}\mathbb{E}Y_{ij}^2 + \frac{1}{3}\Delta n\epsilon}\right) \\
&\leq \exp\left(\frac{-\frac{1}{2}n^2\epsilon^2}{C\rho_n\|\psi^{(0)} - \frac{1}{2}\|_2^2 + \frac{1}{3}\Delta n\epsilon}\right) \\
&\leq \exp\left(\frac{-\frac{1}{2}n^2\epsilon^2}{Cn\rho_n\Delta^2 + \frac{1}{3}\Delta n\epsilon}\right).
\end{aligned} \tag{27}$$

By taking $\epsilon = C'\Delta\sqrt{\frac{\rho_n}{n}\log n}$ for some large C', it follows from the union bound and $n\rho_n = \Omega(\log n)$ that the event $\mathcal{A}_1 = \{\max_i |r_i^{(0)}| = O(\sqrt{n\rho_n \log n}\Delta\}$ has probability at least $1 - \exp(-\Theta(\log n))$.

Now, from our assumption $n|a_{\pm 1}^{(0)}| \gg \max\{\sqrt{n\rho_n \log n}\|\psi^{(0)} - \frac{1}{2}\|_\infty, 1\}$, it follows that $na_{\sigma_i}^{(0)} \gg 4t_0 r_i^{(0)} + b_i^{(0)}$ under event $\mathcal{A}_1$, simultaneously for all i. Thus, similar to the population case, we can write

$$\psi^{(1)} = g(na_{+1}^{(0)})\mathbf{1}_{\mathcal{C}_1} + g(na_{-1}^{(0)})\mathbf{1}_{\mathcal{C}_2} + \hat{\delta}^{(0)},$$

where $\|\hat{\delta}^{(0)}\|_\infty = O(\exp(-n\min\{|a_{+1}^{(0)}|, |a_{-1}^{(0)}|\})) = o(1)$, with probability at least $1-\exp(-\Theta(\log n))$. After this the proof proceeds like the the proof of Theorem 4, and so we omit it.

Let us consider the case with $s = 2$ and we will show $r_i^{(1)}$ can be bounded in a general way. Now

$$\begin{aligned}\xi^{(2)} &= 4t_0 M(\psi^{(1)} - \frac{1}{2}\mathbf{1}) + 4t_0 r^{(1)} \\ &= 4t_0 M(\psi^{(1)} - \frac{1}{2}\mathbf{1}) + \underbrace{4t_0(A-\tilde{P})(\psi^{(1)} - \ell(\psi^{(0)}))}_{R_1} + \underbrace{4t_0(A-\tilde{P})(\ell(\psi^{(0)}) - \frac{1}{2}\mathbf{1})}_{R_2}.\end{aligned}$$

Now the analysis of the first term follows from Theorem 4. Define event $\mathcal{A}_2 = \{\max_i |R_{2,i}| = O(\sqrt{n\rho_n \log n})\}$. Since $\ell(\psi^{(0)}) \in \{\mathbf{1}_{\mathcal{C}_1}, \mathbf{1}_{\mathcal{C}_2}, \mathbf{1}, \mathbf{0}, \frac{1}{2}\mathbf{1}\}$ which is a finite set, by the same argument as Eq (27), $\mathcal{A}_2$ has probability at least $1 - \exp(-\Theta(\log n))$. For R_1, define $\mathcal{A}_3 = \{\|A - \tilde{P}\|_{op} = O(\sqrt{n\rho_n})\}$. $\mathcal{A}_3$ has probability at least $1 - n^{-r}$, $r > 0$ (Theorem 5.2 in Lei et al. (2015)). Under $\mathcal{A}_3$,

$$\begin{aligned}\max_i |R_{1,i}| \le \|R_1\|_2 &\le C\|A - \tilde{P}\|_{op}\|\psi^{(1)} - \ell(\psi^{(0)})\|_2 \\ &= O(\sqrt{n\rho_n})\sqrt{n} \cdot O(\exp(-\Theta(n\min\{|a_{+1}^{(0)}|, |a_{-1}^{(0)}|\}))) = o(1),\end{aligned}$$

using the assumption that $n|a_{\pm 1}^{(0)}| \gg \max\{\sqrt{n\rho_n \log n}\|\psi^{(0)} - \frac{1}{2}\|_\infty, 1\}$. Hence $\max_i |r_i^{(1)}| = O(\sqrt{n\rho_n \log n})$ and $na_{\sigma_i}^{(1)} \gg 4t_0 r_i^{(1)} + b_i^{(1)}$ simultaneously for all i, under $\mathcal{A}_2 \cap \mathcal{A}_3$. The same analysis as in the $s = 1$ case follows.

The case for general s can be proved by induction using the same decomposition of $r^{(s)}$, which can be bounded uniformly for all s under $\mathcal{A}_2 \cap \mathcal{A}_3$. ■

The main proof of Theorem 9 relies on a few lemmas, which we defer to the end of the proof.

Proof [Proof of Theorem 9]

For convenience, we assume A has self loops, which has no effect on the conclusion. Similar to the notation used in the proof of Theorem 7, we decompose ξ_i as the population update plus noise,

$$\xi_i^{(s+1)} = 4t_0 \underbrace{M_{i,\cdot}(\psi^{(s)} - \frac{1}{2}\mathbf{1})}_{\text{signal}} + 4t_0 \underbrace{(A - \mathbb{E}(A|Z))_{i,\cdot}(\psi^{(s)} - \frac{1}{2}\mathbf{1})}_{r_i^{(s)}}. \tag{28}$$

Note that the signal part is constant for $i \in \mathbf{1}_{\mathcal{C}_1}$ and $i \in \mathbf{1}_{\mathcal{C}_2}$. For convenience denote

$$\begin{aligned}s_1 &= M_{i,\cdot}(\psi^{(0)} - \frac{1}{2}\mathbf{1}), \qquad i \in \mathbf{1}_{\mathcal{C}_1} \\ s_2 &= M_{i,\cdot}(\psi^{(0)} - \frac{1}{2}\mathbf{1}), \qquad i \in \mathbf{1}_{\mathcal{C}_2}.\end{aligned} \tag{29}$$

Similarly, define $s_1^{(1)}$ and $s_2^{(1)}$ in terms of $\psi^{(1)}$. By Lemma 23, since $p_0 > \lambda_0 > q_0$, for $\Delta_1, \Delta_2 > 0$,

$$\begin{aligned} s_1^{(1)} &= (p_0 - \lambda_0) \sum_{i \in \mathcal{C}_1} (\psi_i^{(1)} - \frac{1}{2}) + (q_0 - \lambda_0) \sum_{i \in \mathcal{C}_2} (\psi_i^{(1)} - \frac{1}{2}) \\ &\geq (p_0 - \lambda_0) \frac{n}{2} \left(\frac{1}{2} - \Phi\left(- \frac{s_1 - \Delta_1}{\sigma_\psi} \right) \right) + (q_0 - \lambda_0) \frac{n}{2} \left(\frac{1}{2} - \Phi\left(- \frac{s_2 + \Delta_2}{\sigma_\psi} \right) \right) \\ &\quad \underbrace{- O(n\rho_n)(e^{-4t_0\Delta_1} + e^{-4t_0\Delta_2}) - O(n\rho_n)\frac{\rho_\psi}{\sigma_\psi^3} - O_P(\sqrt{n}\rho_n)}_{R_\psi}. \end{aligned} \tag{30}$$

Similarly,

$$\begin{aligned} s_2^{(1)} &= (q_0 - \lambda_0) \sum_{i \in \mathcal{C}_1} (\psi_i^{(1)} - \frac{1}{2}) + (p_0 - \lambda_0) \sum_{i \in \mathcal{C}_2} (\psi_i^{(1)} - \frac{1}{2}) \\ &\leq (q_0 - \lambda_0) \frac{n}{2} \left(\frac{1}{2} - \Phi\left(- \frac{s_1 - \Delta_1}{\sigma_\psi} \right) \right) + (p_0 - \lambda_0) \frac{n}{2} \left(\frac{1}{2} - \Phi\left(- \frac{s_2 + \Delta_2}{\sigma_\psi} \right) \right) + R_\psi \end{aligned} \tag{31}$$

We consider bounding $s_1^{(1)}$ and $s_2^{(1)}$ based on the signs of s_1 and s_2, which only depend on $\psi^{(0)}$. Therefore in each case, we first consider the conditional distribution given $\psi^{(0)}$.

Case 1: $s_1 > 0$, $s_2 < 0$.

Let $\Delta_1 = \epsilon s_1$, $\Delta_2 = -\epsilon s_2$ for some small $\epsilon > 0$. We have

$$\begin{aligned} \frac{1}{2} - \Phi\left(- \frac{(1-\epsilon)s_1}{\sigma_\psi} \right) &\geq \frac{(1-\epsilon)s_1}{\sigma_\psi \sqrt{2\pi}} \exp\left(- \frac{(1-\epsilon)^2 s_1^2}{2\sigma_\psi^2} \right), \\ \Phi\left(- \frac{(1-\epsilon)s_2}{\sigma_\psi} \right) - \frac{1}{2} &\geq - \frac{(1-\epsilon)s_2}{\sigma_\psi \sqrt{2\pi}} \exp\left(- \frac{(1-\epsilon)^2 s_2^2}{2\sigma_\psi^2} \right), \end{aligned}$$

where we have used

$$\begin{aligned} |\Phi(x) - 1/2| &= \frac{1}{\sqrt{2\pi}} \int_0^{|x|} e^{-u^2/2} du \\ &\geq \frac{|x|}{\sqrt{2\pi}} e^{-x^2/2}. \end{aligned} \tag{32}$$

Applying the above to (30),

$$s_1^{(1)} \geq \frac{n(1-\epsilon)}{2\sqrt{2\pi}\sigma_\psi} ((p_0 - \lambda_0)|s_1| + (\lambda_0 - q_0)|s_2|) \exp\left(- \frac{(1-\epsilon)^2 s_2^2 \vee s_1^2}{2\sigma_\psi^2} \right) - R_\psi. \tag{33}$$

Similar arguments show

$$s_2^{(1)} \leq - \frac{n(1-\epsilon)}{2\sqrt{2\pi}\sigma_\psi} ((\lambda_0 - q_0)|s_1| + (p_0 - \lambda_0)|s_2|) \exp\left(- \frac{(1-\epsilon)^2 s_2^2 \vee s_1^2}{2\sigma_\psi^2} \right) + R_\psi \tag{34}$$

Case 2: $s_1 < 0$, $s_2 > 0$.

The same analysis applies with the role of $\mathcal{C}_1$ and $\mathcal{C}_2$ interchanged.

Case 3: $s_1 > 0$, $s_2 > 0$.

WLOG assume $s_1 > s_2 > 0$. Taking $\Delta_1 = \Delta_2 = \epsilon(s_1 - s_2)$, (30) becomes

$$s_1^{(1)} \geq \frac{n}{2\sqrt{2\pi}\sigma_\psi}[(p_0 - \lambda_0)(s_1 - \epsilon(s_1 - s_2)) - (\lambda_0 - q_0)(s_2 + \epsilon(s_1 - s_2))] \exp\left(-\frac{(1-\epsilon)^2 s_1^2}{2\sigma_\psi^2}\right) - R_\psi$$
$$\geq \frac{n}{2\sqrt{2\pi}\sigma_\psi}[(\lambda_0 - q_0) - \epsilon(p_0 - q_0)]|s_1 - s_2| \exp\left(-\frac{(1-\epsilon)^2 s_1^2}{2\sigma_\psi^2}\right) - R_\psi \tag{35}$$

using $p_0 - \lambda_0 > \lambda_0 - q_0$ (Proposition 2). Since $\lambda_0 - q_0 = \Omega(\rho_n)$ also by Proposition 2, choose a ϵ small enough so that $(\lambda_0 - q_0) - \epsilon(p_0 - q_0) \geq \Omega(\rho_n)$.

Similarly, taking $\Delta_1 = \epsilon_n s_1$, $\Delta_2 = \epsilon_n s_2$,

$$s_2^{(1)} \leq -\frac{n}{2\sqrt{2\pi}\sigma_\psi}[(\lambda_0 - q_0)(1-\epsilon_n)s_1 - (p_0 - \lambda_0)(1+\epsilon_n)s_2] \exp\left(-\frac{(1+\epsilon_n)^2 s_1^2}{2\sigma_\psi^2}\right) + R_\psi, \tag{36}$$

Letting $\epsilon_n \to 0$ slowly and denote $c = \frac{(\lambda_0 - q_0)(1-\epsilon_n)}{(p_0 - \lambda_0)(1+\epsilon_n)} - \eta$, for some small $\eta > 0$. When $s_2 \leq cs_1$,

$$s_2^{(1)} \leq -\frac{n}{2\sqrt{2\pi}\sigma_\psi}\eta(p_0 - \lambda_0)|s_1| + R_\psi. \tag{37}$$

By Lemma 22, $s_2 \leq cs_1$ happens with probability

$$P(0 < s_2 \leq cs_1) = \frac{\arctan(c_u)}{2\pi} - \frac{\arctan(c_\ell)}{2\pi} + O(n^{-1/2}),$$

where $c_u = \frac{p_0 - \lambda_0}{\lambda_0 - q_0}$, $c_\ell = \frac{(p_0 - \lambda_0) + c(\lambda_0 - q_0)}{c(p_0 - \lambda_0) + (\lambda_0 - q_0)}$.

When $s_2 > s_1 > 0$, the analysis is the same by symmetry. We have the same bounds for $s_1^{(1)}$ and $s_2^{(1)}$ with s_1 and s_2 interchanged. By a similar calculation, we need

$$P(0 < s_1 \leq cs_2) = \frac{\arctan(c_\ell^{-1})}{2\pi} - \frac{\arctan(c_u^{-1})}{2\pi} + O(n^{-1/2}),$$

Case 4: $s_1 < 0$, $s_2 < 0$. By symmetry, $g(4t_0(s_1 + r_i^{(0)})) - \frac{1}{2} = \frac{1}{2} - g(-4t_0(s_1 + r_i^{(0)}))$ (similarly for $g(4t_0(s_2 + r_i^{(0)}))$). It suffices to apply the same analysis in Case 3 to $-s_1, -s_2$ and $-r_i^{(0)}$. For example, when $s_1 < s_2 < 0$, $-s_1^{(1)}$ is lower bounded by (35), $-s_2^{(1)}$ is upper bounded by (36) when $0 < -s_1 < -cs_2$.

Now combining all the cases, define event $\mathcal{B}$ as

$$\mathcal{B} = \left\{|s_1^{(1)}|, |s_2^{(1)}| \geq C\eta n \rho_n \sigma_\psi^{-1} \min\{|s_1|, |s_2|, |s_1 - s_2|\} \exp\left(-\frac{(1+\epsilon)^2 s_2^2 \vee s_1^2}{2\sigma_\psi^2}\right) - R_\psi, s_1^{(1)} s_2^{(1)} < 0\right\},$$

where C depends on p_0, q_0. Cases 1–4 imply

$$P(\mathcal{B}) = \sum_{\psi: s_1 s_2 < 0} P(\mathcal{B}|\psi^{(0)} = \psi) P(\psi^{(0)} = \psi) + \sum_{\psi: s_1 s_2 > 0} P(\mathcal{B}|\psi^{(0)} = \psi) P(\psi^{(0)} = \psi)$$

$$\geq P(s_1 s_2 < 0) + 2P(0 < s_2 < cs_1) + 2P(0 < s_1 < cs_2)$$
$$= \frac{1}{2} + \frac{2\arctan(c_u^{-1})}{\pi} + \frac{\arctan(c_u) - \arctan(c_u^{-1})}{\pi} - \frac{\arctan(c_\ell) - \arctan(c_\ell^{-1})}{\pi} + O(n^{-1/2})$$
$$= 1 - \frac{\arctan(c_\ell) - \arctan(c_\ell^{-1})}{\pi} + O(n^{-1/2}), \tag{38}$$

where

$$P(s_1 > 0, s_2 < 0) = P(s_1 < 0, s_2 > 0) = \frac{1}{4} + \frac{\arctan(c_u^{-1})}{\pi} + O(n^{-1/2})$$

using calculations similar to Lemma 22.

Define event $\mathcal{D} = \{|s_1|, |s_2|, |s_1 - s_2| \geq \rho_n\sqrt{n}/c_n\}$ for some $c_n \to \infty$ slowly. Then by Lemma 20, $P(\mathcal{D}) \geq 1 - O(1/c_n)$. Also $\sigma_\psi^2 \asymp n\rho_n$, $\rho_\psi \asymp n\rho_n$, $e^{-4t_0|s_1|}, e^{-4t_0|s_2|} = O_P(\exp(-\rho_n\sqrt{n}))$, and $\epsilon_n \to 0$ slow enough such that $\frac{\rho_n\sqrt{n}\epsilon_n}{|\log \rho_n|} \to \infty$, it follows $R_\psi = o(n\rho_n^{3/2})$ with high probability. Then by (38),

$$P(|s_1^{(1)}|, |s_2^{(1)}| \geq C\eta n\rho_n^{3/2}/c_n, s_1^{(1)} s_2^{(1)} < 0) \geq 1 - \frac{\arctan(c_\ell) - \arctan(c_\ell^{-1})}{\pi} - o(1), \tag{39}$$

for some constant C depending on p_0, q_0.

From now on we will work under the event in (39). In the next iteration, write the true labels as $z_0 = \mathbf{1}_{\mathcal{C}_1}\mathbb{1}\{s_1^{(1)} > 0\} + \mathbf{1}_{\mathcal{C}_2}\mathbb{1}\{s_1^{(1)} < 0\}$. When $s_1^{(1)} > 0$ holds,

$$|\psi_i^{(2)} - z_{0,i}| = \frac{1}{1 + e^{\sigma_i \xi_i^{(2)}}} \leq e^{-x_0} + \mathbb{1}\{\sigma_i \xi_i^{(2)} \leq x_0\} \tag{40}$$

for any $x_0 > 0$. For $i \in \mathcal{C}_1$,

$$\xi_i^{(2)} = 4t_0 s_1^{(1)} + 4t_0 r_i^{(1)}$$
$$= 4t_0 s_1^{(1)} + 4t_0(A - P)_{i,\cdot}(z_0 - \frac{1}{2}\mathbf{1}) + 4t_0(A - P)_{i,\cdot}(\psi^{(1)} - z_0)$$

Taking $x_0 = C\eta t_0 \rho_n^{3/2} n/c_n$, since $s_1^{(1)} \geq C\eta n\rho_n^{3/2}/c_n$, we have $4t_0 s_1^{(1)} - 2x_0 > 2C\eta t_0 n\rho_n^{3/2}/c_n$ for large n. Further, $(A - P)_{i,\cdot}(z_0 - \frac{1}{2}\mathbf{1}) = O(\sqrt{n\rho_n \log n})$ uniformly for all i with high probability by an argument similar to Eq (27), then

$$\mathbb{1}\{\xi_i^{(2)} \leq x_0\} \leq \mathbb{1}\left\{4t_0 s_1^{(1)} - O(\sqrt{n\rho_n \log n}) \leq 2x_0\right\}$$
$$+ \mathbb{1}\left\{4t_0(A - P)_{i,\cdot}(\psi^{(1)} - z_0) \leq -x_0\right\}$$
$$\leq \exp\left(2x_0 - 4t_0 s_1^{(1)} + O(\sqrt{n\rho_n \log n})\right) + \mathbb{1}\left\{4t_0(A - P)_{i,\cdot}(\psi^{(1)} - z_0) \leq -C\eta t_0 \rho_n^{3/2} n/c_n\right\}$$
$$\leq \exp(-C_1'\eta\rho_n^{3/2} n/c_n) + \mathbb{1}\left\{(A - P)_{i,\cdot}(\psi^{(1)} - z_0) \leq -C_2'\eta\rho_n^{3/2} n/c_n\right\} \tag{41}$$

where C_1', C_2' are constants depending on p_0, q_0. Similarly for $i \in \mathcal{C}_2$,

$$\mathbb{1}\{-\xi_i^{(2)} \leq x_0\} \leq \exp(-C_1'\eta\rho_n^{3/2} n/c_n) + \mathbb{1}\left\{(A - P)_{i,\cdot}(\psi^{(1)} - z_0) \geq C_2'\eta\rho_n^{3/2} n/c_n\right\}. \tag{42}$$

Summing (40) using (41) and (42),

$$\begin{aligned}\|\psi^{(2)} - z_0\|_1 &\leq n\exp(-C_1'\eta\rho_n^{3/2}n/c_n) + \sum_i \mathbb{1}\left\{\left|(A-P)_{i,\cdot}(\psi^{(1)} - z_0)\right| \geq C_2'\eta\rho_n^{3/2}n/c_n\right\} \\ &\leq n\exp(-C_1'\eta\rho_n^{3/2}n/c_n) + \frac{c_n^2(\psi^{(1)} - z_0)^T(A-P)^2(\psi^{(1)} - z_0)}{(C_2')^2\eta^2n^2\rho_n^3} \\ &\leq n\exp(-C_1'\eta\rho_n^{3/2}n/c_n) + \frac{c_n^2\|A-P\|_{op}^2\|\psi^{(1)} - z_0\|_2^2}{(C_2')^2\eta^2n^2\rho_n^3} \\ &\leq n\exp(-C_1'\eta\rho_n^{3/2}n/c_n) + \frac{C_2'c_n^2}{\eta^2n\rho_n^2}\|\psi^{(1)} - z_0\|_1, \end{aligned} \tag{43}$$

redefining C_2' in the last line, where we have used the fact that there exist $r > 0$ such that $\|A - P\|_{op} = O(\sqrt{n\rho_n})$ with probability at least $1 - n^{-r}$ (Theorem 5.2 in Lei et al. (2015)). The probability of (43) happening has the same lower bound as in (39).

The case for $s_1^{(1)} < 0$ is similar with $z_0 = \mathbf{1}_{\mathcal{C}_2}$.

For later iterations, note that when $z_0 = \mathbf{1}_{\mathcal{C}_1}$, $\|\psi^{(2)} - z_0\|_1 = n/2 - \langle\psi^{(2)}, u_2\rangle$, then (43) implies

$$\langle\psi^{(2)}, u_2\rangle \geq \frac{n}{2} - \delta_n n$$

for some $\delta_n = o(1)$ slow enough such that $\delta_n > \exp(-C_1\eta\rho_n^{3/2}n/c_n) + \frac{C_2c_n^2}{\eta^2n\rho_n^2}$, and

$$\sum_{i\in\mathcal{C}_1}(\psi_i^{(2)} - 1/2) \geq \frac{n}{4} - \delta_n n.$$

Then since $\lambda_0 - q_0 > 0$, $p_0 + q_0 - 2\lambda_0 > 0$,

$$\begin{aligned} s_1^{(2)} &= (\lambda_0 - q_0)\langle\psi^{(2)}, u_2\rangle + (p_0 + q_0 - 2\lambda_0)\sum_{i\in\mathcal{C}_1}(\psi_i^{(2)} - 1/2) \\ &\geq (\lambda_0 - q_0)(\frac{n}{2} - \delta_n n) + (p_0 + q_0 - 2\lambda_0)(\frac{n}{4} - \delta_n n) \\ &= \frac{1}{4}(p_0 - q_0)n - \delta_n(p_0 - \lambda_0)n \geq C_0(p_0 - q_0)n \end{aligned} \tag{44}$$

for a general constant $C_0 < 1/4$ independent of model parameters, and large n. Similarly,

$$\sum_{i\in\mathcal{C}_2}(\psi_i^{(2)} - 1/2) \leq -\frac{n}{4} + \delta_n n,$$

$$s_2^{(2)} = -(\lambda_0 - q_0)\langle\psi^{(2)}, u_2\rangle + (p_0 + q_0 - 2\lambda_0)\sum_{i\in\mathcal{C}_2}(\psi_i^{(2)} - 1/2) \leq -C_0(p_0 - q_0)n.$$

The rest of the argument in (40)-(43) applies with above bounds for $s_1^{(2)}$ and $s_2^{(2)}$, $x_0 = C_0t_0(p_0 - q_0)n$, giving

$$\|\psi^{(3)} - z_0\|_1 \leq n\exp(-C_1t_0(p_0 - q_0)n) + \frac{C_2\rho_n}{(p_0 - q_0)^2n}\|\psi^{(2)} - z_0\|_1 \tag{45}$$

for some general constants C_1, C_2, independent of model parameters. The probability of (45) happening still has the same lower bound as (39).

The same arguments can be repeated for all the later iterations.

■

Now we state and prove all the lemmas needed in the main proof. First we have a few concentration lemmas.

Lemma 19 (Berry-Esseen bound)

$$\sup_{x\in\mathbb{R}} |P\left(r_i^{(0)}/\sigma_\psi \le x \mid \psi^{(0)}\right) - \Phi(x)| \le C_0 \cdot \frac{\rho_\psi}{\sigma_\psi^3},$$

where C_0 is a general constant, ρ_ψ and σ_ψ depend on $\psi^{(0)}$.

Proof Define

$$\sigma_\psi^2 := p_0(1-p_0)\sum_{i\in\mathcal{C}_1}(\psi_i^{(0)} - 1/2)^2 + q_0(1-q_0)\sum_{i\in\mathcal{C}_2}(\psi_i^{(0)} - 1/2)^2,$$

$$\rho_\psi := p_0(1-p_0)(1-2p_0+2p_0^2)\sum_{i\in\mathcal{C}_1}|\psi_i^{(0)} - 1/2|^3 + q_0(1-q_0)(1-2q_0+2q_0^2)\sum_{i\in\mathcal{C}_2}|\psi_i^{(0)} - 1/2|^3.$$

It follows by the Berry-Esseen bound that

$$\sup_{x\in\mathbb{R}} |P\left(r_i^{(0)}/\sigma_\psi \le x \mid \psi^{(0)}\right) - \Phi(x)| \le C_0 \cdot \frac{\rho_\psi}{\sigma_\psi^3}$$

for some general constant C_0, where Φ is the CDF of standard Gaussian. ■

Lemma 20 (Littlewood-Offord) *Let $s_1 = (p_0-\lambda_0)\sum_{i\in\mathcal{C}_1}(\psi_i^{(0)}-1/2)+(q_0-\lambda_0)\sum_{i\in\mathcal{C}_2}(\psi_i^{(0)}-1/2)$, $s_2 = (q_0-\lambda_0)\sum_{i\in\mathcal{C}_1}(\psi_i^{(0)}-1/2)+(p_0-\lambda_0)\sum_{i\in\mathcal{C}_2}(\psi_i^{(0)}-1/2)$. Then*

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

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

Proof Noting that $2\psi_i^{(0)} - 1 \in \{-1, 1\}$ each with probability $1/2$, and $q_0 < \lambda_0 < p_0$, this is a direct consequence of the Littlewood-Offord bound in Erdös (1945). ■

Lemma 21 (McDiarmid's Inequality) *Recall $r_i^{(0)} = (A - \mathbb{E}(A|Z))(\psi^{(0)} - \frac{1}{2}\mathbf{1})$ and let $h(r_i^{(0)})$ be a bounded function with $\|h\|_\infty \le M$. Then*

$$P\left(\left|\frac{2}{n}\sum_{i\in\mathcal{C}_1} h(r_i^{(0)}) - \mathbb{E}(h(r_i^{(0)})|\psi^{(0)})\right| > w \mid \psi^{(0)}\right) \le \exp\left(-\frac{w^2}{nM}\right).$$

The same bound holds for $i \in \mathcal{C}_2$.

Proof Define $\phi = \frac{2}{n}\sum_{i\in\mathcal{C}_1} h(r_i^{(0)})$, then conditional on $\psi^{(0)}$, ϕ is only a function of $(A_{ij})_{i<j,i\in\mathcal{C}_1}$. Replacing any A_{ij} with $A'_{ij} \in \{0,1\}$,

$$|\phi(A_{12},\dots,A_{ij},\dots) - \phi(A_{12},\dots,A'_{ij},\dots)| \le \frac{8M}{n}.$$

and

$$\sum_{i<j,i\in\mathcal{C}_1} |\phi(A_{12},\dots,A_{ij},\dots) - \phi(A_{12},\dots,A'_{ij},\dots)| \le 2nM$$

The desired bound follows by McDiarmid's inequality. ■

Using the normal approximation, we can also derive the following probability bound for s_1 and s_2.

Lemma 22 *For some constant* $0 < c < 1$,

$$P(0 \le s_2 \le cs_1) = \frac{\arctan(c_u)}{2\pi} - \frac{\arctan(c_\ell)}{2\pi} + O(n^{-1/2}),$$

where $c_\ell = \frac{(p_0-\lambda_0)+c(\lambda_0-q_0)}{c(p_0-\lambda_0)+(\lambda_0-q_0)}$, $c_u = \frac{p_0-\lambda_0}{\lambda_0-q_0}$.

Proof For convenience, denote $T_1 = \sum_{i\in\mathcal{C}_1}(\psi_i^{(0)} - 1/2)$, $T_2 = \sum_{i\in\mathcal{C}_2}(\psi_i^{(0)} - 1/2)$, then

$$\begin{aligned}\{0 \le s_2 \le cs_1\} &= \left\{\frac{(p_0-\lambda_0)+c(\lambda_0-q_0)}{c(p_0-\lambda_0)+(\lambda_0-q_0)}T_2 \le T_1 \le \frac{p_0-\lambda_0}{\lambda_0-q_0}T_2\right\} \\ &:= \{c_\ell T_2 \le T_1 \le c_u T_2 \text{ and } T_1, T_2 > 0\}\end{aligned}$$

where $1 < c_\ell < c_u$. It is easy to see that $\mathbb{E}(T_1) = \mathbb{E}(T_2) = 0$, $\sigma_T^2 := \mathbb{E}(T_1^2) = \mathbb{E}(T_1^2) \asymp \rho_n^2 n$, $\mathbb{E}|T_1|^3 = \mathbb{E}|T_1|^3 \asymp \rho_n^3 n$. Then

$$P(0 \le s_2 \le cs_1) = P(0 \le T_1 \le c_u T_2) - P(0 \le T_1 < c_\ell T_2). \tag{46}$$

The first part can be calculated as

$$\begin{aligned}P(0 \le T_1 \le c_u T_2) &= \sum_{t\ge 0} P(0 \le T_1 \le c_u t | T_2 = t) P(T_2 = t) \\ &= \sum_{t\ge 0} P(0 \le Z_1 \le c_u T_2 \sigma_T^{-1} | T_2 = t) P(T_2 = t) + O(n^{-1/2}) \\ &= \mathbb{E}\left((\Phi(c_u T_2 \sigma_T^{-1}) - 1/2)\mathbb{1}(T_2 \ge 0)\right) + O(n^{-1/2})\end{aligned}$$

using the Berry-Esseen bound, $Z_1 \sim N(0,1)$. Now note that $(\Phi(c_u T_2 \sigma_T^{-1}) - 1/2)\mathbb{1}(T_2 \ge 0)$ is continuous and monotonic in T_2. For every $t \in (0,1]$, there exists $a(t) > 0$ such that $\Phi(c_u T_2 \sigma_T^{-1}) - 1/2 \ge t \Leftrightarrow T_2 \sigma_T^{-1} \ge a(t)$. We have

$$\mathbb{E}\left((\Phi(c_u T_2 \sigma_T^{-1}) - 1/2)\mathbb{1}(T_2 \ge 0)\right) = \int_0^1 P\left((\Phi(c_u T_2 \sigma_T^{-1}) - 1/2)\mathbb{1}(T_2 \ge 0) \ge t\right) dt$$

$$
\begin{aligned}
&= \int_0^1 P(T_2\sigma_T^{-1} \geq a(t))dt \\
&= \int_0^1 P(Z_2 \geq a(t))dt + O(n^{-1/2}) \\
&= \mathbb{E}\left((\Phi(c_u Z_2) - 1/2)\mathbb{1}(Z_2 \geq 0)\right) + O(n^{-1/2}),
\end{aligned}
$$

$Z_2 \sim N(0,1)$, independent of Z_1. It remains to calculate the expectation, which can be written as

$$
w(x) = \frac{1}{2\pi}\int_0^\infty \int_0^{xz} \exp(-u^2/2)\exp(-z^2/2)dudz
$$

for $x = c_u$. Now

$$
w'(x) = \frac{1}{2\pi}\int_0^\infty z\exp(-(1+x^2)z^2/2)dz = \frac{1}{2\pi(1+x^2)}
$$

Integrating both sides, we get: $w(x) = \frac{\arctan(x)}{2\pi} + C$, where $C = 0$ since $w(0) = 0$. Thus $w(c_u) = \frac{\arctan(c_u)}{2\pi}$. The same calculation can be done for $P(0 \leq T_1 \leq c_\ell T_2)$. Substituting into (46),

$$
P(0 \leq s_2 \leq cs_1) = \frac{\arctan(c_u)}{2\pi} - \frac{\arctan(c_\ell)}{2\pi} + O(n^{-1/2})
$$

■

Finally, we have the following general bounds for $\sum_{i\in\mathcal{C}_1}\psi_i^{(1)}$ and $\sum_{i\in\mathcal{C}_2}\psi_i^{(1)}$.

Lemma 23 *For any $\Delta_1 > 0$,*

$$
\begin{aligned}
\sum_{i\in\mathcal{C}_1}\psi_i^{(1)} &\geq \frac{n}{2}\left(1 - \Phi\left(-\frac{s_1 - \Delta_1}{\sigma_\psi}\right)\right) - \frac{n}{2}e^{-4t_0\Delta_1} - C'n\cdot\frac{\rho_\psi}{\sigma_\psi^3} - O_P(\sqrt{n}), \\
\sum_{i\in\mathcal{C}_1}\psi_i^{(1)} &\leq \frac{n}{2}\left(1 - \Phi\left(-\frac{s_1 + \Delta_1}{\sigma_\psi}\right)\right) + \frac{n}{2}e^{-4t_0\Delta_1} + C'n\cdot\frac{\rho_\psi}{\sigma_\psi^3} + O_P(\sqrt{n}),
\end{aligned} \tag{47}
$$

where Φ is the CDF of standard Gaussian, ρ_ψ and σ_ψ are constants depending on $\psi^{(0)}$ defined in Lemma 19, and the $O_P(\sqrt{n})$ terms are uniform for $\psi^{(0)}$. The same upper and lower bound hold for $i \in \mathcal{C}_2$ and s_2.

Proof Define an index set $J_1^+ = \{i : r_i^{(0)} > -s_1 + \Delta_1\}$, $\Delta_1 > 0$. Then for $i \in \mathcal{C}_1 \cap J_1^+$,

$$
\psi_i^{(1)} = g(4t_0(s_1 + r_i^{(0)})) \geq g(4t_0\Delta_1) \geq 1 - e^{-4t_0\Delta_1}.
$$

It follows then

$$
\sum_{i\in\mathcal{C}_1}\psi_i^{(1)} \geq |\mathcal{C}_1 \cap J_1^+|(1 - e^{-4t_0\Delta_1}) \tag{48}
$$

To calculate the size of the set, note that

$$|\mathcal{C}_1 \cap J_1^+| = \sum_{i \in \mathcal{C}_1} \mathbb{1}(r_i^{(0)} > -s_1 + \Delta_1),$$

By Lemma 21,

$$\begin{aligned}
|\mathcal{C}_1 \cap J_1| &= \frac{n}{2} P(r_i^{(0)} > -s_1 + \Delta_1) \mid \psi^{(0)}) + O_P(\sqrt{n}) \\
&\geq \frac{n}{2} \left(P(r > -s_1 + \Delta_1) - C_0 \cdot \frac{\rho_\psi}{\sigma_\psi^3} \right) - O_P(\sqrt{n}) \\
&= \frac{n}{2} \left(1 - \Phi \left(-\frac{s_1 - \Delta_1}{\sigma_\psi} \right) \right) - C' n \cdot \frac{\rho_\psi}{\sigma_\psi^3} - O_P(\sqrt{n}), \qquad (49)
\end{aligned}$$

where the second line follows from Lemma 19, with Φ as the CDF of standard Gaussian, $r \sim N(0, \sigma_\psi^2)$ and the $O_P(\sqrt{n})$ can be made uniform over $\psi^{(0)}$. (48) and (49) imply

$$\begin{aligned}
\sum_{i \in \mathcal{C}_1} \psi_i^{(1)} &\geq \frac{n}{2} \left(1 - \Phi \left(-\frac{s_1 - \Delta_1}{\sigma_\psi} \right) \right) (1 - e^{-4t_0 \Delta_1}) \\
&\quad - C' n \cdot \frac{\rho_\psi}{\sigma_\psi^3} - O_P(\sqrt{n}) \\
&\geq \frac{n}{2} \left(1 - \Phi \left(-\frac{s_1 - \Delta_1}{\sigma_\psi} \right) \right) - \frac{n}{2} e^{-4t_0 \Delta_1} \\
&\quad - C' n \cdot \frac{\rho_\psi}{\sigma_\psi^3} - O_P(\sqrt{n}). \qquad (50)
\end{aligned}$$

Similarly let $J_1^- = \{i : r_i^{(0)} < -s_1 - \Delta_1\}$, $\Delta_1 > 0$. For $i \in \mathcal{C}_1 \cap J_1^-$,

$$\psi_i^{(1)} = g(4t_0(s_1 + r_i^{(0)})) \leq g(-4t_0 \Delta_1) \leq e^{-4t_0 \Delta_1}.$$

We have

$$\begin{aligned}
\sum_{i \in \mathcal{C}_1} \psi_i^{(1)} &\leq |\mathcal{C}_1 \cap J_1^-| e^{-4t_0 \Delta_1} + \frac{n}{2} - |\mathcal{C}_1 \cap J_1^-| \\
&= \frac{n}{2} - |\mathcal{C}_1 \cap J_1^-| (1 - e^{-4t_0 \Delta_1}), \qquad (51)
\end{aligned}$$

where

$$\begin{aligned}
|\mathcal{C}_1 \cap J_1^-| &= \frac{n}{2} P(r_i^{(0)} < -s_1 - \Delta_1) \mid \psi^{(0)}) + O_P(\sqrt{n}) \\
&\geq \frac{n}{2} \Phi \left(-\frac{s_1 + \Delta_1}{\sigma_\psi} \right) - C' n \cdot \frac{\rho_\psi}{\sigma_\psi^3} - O_P(\sqrt{n}). \qquad (52)
\end{aligned}$$

(51) and (52) give

$$\sum_{i \in \mathcal{C}_1} \psi_i^{(1)} \leq \frac{n}{2} - \frac{n}{2} \Phi \left(-\frac{s_1 + \Delta_1}{\sigma_\psi} \right) (1 - e^{-4t_0 \Delta_1})$$

$$
\begin{aligned}
&+ C'n \cdot \frac{\rho_\psi}{\sigma_\psi^3} + O_P(\sqrt{n}) \\
&\leq \frac{n}{2}\left(1 - \Phi\left(-\frac{s_1 + \Delta_1}{\sigma_\psi}\right)\right) + \frac{n}{2} e^{-4t_0\Delta_1} \\
&+ C'n \cdot \frac{\rho_\psi}{\sigma_\psi^3} + O_P(\sqrt{n}). \qquad (53)
\end{aligned}
$$

■

Proof [Proof of Corollary 11]

Let $\hat{t}, \hat{\lambda}$ be constants defined in the usual way 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 (28). Now

$$
\begin{aligned}
\hat{s}_1 &= (p_0 - \hat{\lambda}) \sum_{i \in \mathcal{C}_1} (\psi_i^{(0)} - 1/2) + (q_0 - \hat{\lambda}) \sum_{i \in \mathcal{C}_2} (\psi_i^{(0)} - 1/2) \\
\hat{s}_2 &= (q_0 - \hat{\lambda}) \sum_{i \in \mathcal{C}_1} (\psi_i^{(0)} - 1/2) + (p_0 - \hat{\lambda}) \sum_{i \in \mathcal{C}_2} (\psi_i^{(0)} - 1/2)
\end{aligned}
$$

We can check the rest of the analysis remains unchanged as long as $\hat{p}, \hat{q} \asymp \rho_n$, $|\hat{p} - \hat{q}| = \Omega(\rho_n)$,

1. $\frac{p_0 + q_0}{2} > \hat{\lambda}$,
2. $\hat{\lambda} - q_0 = \Omega(\rho_n) > 0$.

■

C.2 Proofs of results in Section 3.2

Proof [Proof of Proposition 13] That the described point is a stationary point is easy to verify, because of the presence of the $(\psi_i - \frac{1}{2})$ terms in the stationarity equations (19). Now, from (20), we see that the Hessian matrix at $(\frac{1}{2}\mathbf{1}, \frac{\mathbf{1}^\top A \mathbf{1}}{n(n-1)}, \frac{\mathbf{1}^\top A \mathbf{1}}{n(n-1)}, \frac{1}{2})$ is given by

$$
H = \begin{pmatrix} -4I & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^\top & -\frac{n(n-1)}{4\hat{a}(1-\hat{a})} & 0 \\ \mathbf{0}^\top & 0 & -\frac{n(n-1)}{4\hat{a}(1-\hat{a})} \end{pmatrix},
$$

where $\hat{a} = \frac{\mathbf{1}^\top A \mathbf{1}}{n(n-1)}$. Clearly, H is negative definite. This completes the proof. ■

Proof [Proof of Lemma 14] First note that conditioning on the true labels Z, $\mathbb{E}(A|Z) = \tilde{P}$. For notation simplicity, we omit the superscript of $\psi^{(0)}$. For the update of $p^{(1)}$, we have

$$
\begin{aligned}
p^{(1)} =& \frac{\psi^T \tilde{P} \psi + (\mathbf{1} - \psi)^T \tilde{P} (\mathbf{1} - \psi)}{\psi^T (J - I)\psi + (\mathbf{1} - \psi)^T (J - I)(\mathbf{1} - \psi)} \\
&+ \frac{\psi^T (A - \tilde{P})\psi + (\mathbf{1} - \psi)^T (A - \tilde{P})(\mathbf{1} - \psi)}{\psi^T (J - I)\psi + (\mathbf{1} - \psi)^T (J - I)(\mathbf{1} - \psi)}, \qquad (54)
\end{aligned}
$$

where the first term can be written as

$$\begin{aligned}
&\frac{\psi^T(\frac{p_0+q_0}{2}u_1u_1^T+\frac{p_0-q_0}{2}u_2u_2^T-p_0I)\psi+(\mathbf{1}-\psi)^T(\frac{p_0+q_0}{2}u_1u_1^T+\frac{p_0-q_0}{2}u_2u_2^T-p_0I)(\mathbf{1}-\psi)}{\psi^T(u_1u_1^T-I)\psi+(\mathbf{1}-\psi)^T(u_1u_1^T-I)(\mathbf{1}-\psi)}\\
=&\frac{\frac{p_0+q_0}{2}n^2(\zeta_1^2+(1-\zeta_1)^2)+n^2(p_0-q_0)\zeta_2^2-p_0x}{\zeta_1^2n^2+(1-\zeta_1)^2n^2-x}\\
=&\frac{p_0+q_0}{2}+\frac{(p_0-q_0)(\zeta_2^2-x/2n^2)}{\zeta_1^2+(1-\zeta_1)^2-x/n^2},
\end{aligned}$$

where $x=\psi^T\psi+(\mathbf{1}-\psi)^T(\mathbf{1}-\psi)\geq n/4$. The second term can be bounded by noting $\mathbb{E}(\psi^T(A-\tilde{P})\psi)=0$ and $\mathrm{Var}(\psi^T(A-\tilde{P})\psi)\leq 2n(n-1)p_0$. By Chebyshev's inequality, $\psi^T(A-\tilde{P})\psi=O_P(\sqrt{\rho_n}n)$.

This is because

$$\mathbb{E}_{\psi,A}[\psi^T(A-\tilde{P})\psi]=\mathbb{E}_\psi\mathbb{E}_A[\psi^T(A-\tilde{P})\psi\Big|\psi]=0,$$

and

$$\begin{aligned}
\mathrm{Var}_{\psi,A}[\psi^T(A-\tilde{P})\psi]&=\mathbb{E}\left(\mathrm{Var}(\psi^T(A-\tilde{P})\psi\Big|\psi)\right)+\mathrm{Var}(\mathbb{E}[\psi^T(A-\tilde{P})\psi\Big|\psi])\\
&=\mathbb{E}\left(\mathrm{Var}(\psi^T(A-\tilde{P})\psi\Big|\psi)\right)\\
&=4\mathbb{E}\sum_{i<j}\psi_i\psi_j\mathrm{Var}(A_{ij})\leq 2n(n-1)p_0.
\end{aligned}$$

$(1-\psi)^T(A-\tilde{P})(1-\psi)$ can be handled similarly, and

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

since the first two terms are minimized at $\sum_i\psi_i=n/2$.

The result for $q^{(1)}$ is proved analogously. ■

Proof [Proof of Proposition 15] Let $\psi=\zeta_1u_1+\zeta_2u_2+w$, $w\in\mathrm{span}\{u_1,u_2\}^\perp$, be a stationary point. We will consider the population version of all the updates and replace A with $\mathbb{E}(A|Z):=\tilde{P}$ and $\rho_n\to 0$. By Lemma 14,

$$\begin{aligned}
\tilde{p}&=\frac{p_0+q_0}{2}+\underbrace{\frac{(p_0-q_0)(\zeta_2^2-x/2n^2)}{\zeta_1^2+(1-\zeta_1)^2-x/n^2}}_{\epsilon_1'},\\
\tilde{q}&=\frac{p_0+q_0}{2}-\underbrace{\frac{(p_0-q_0)(\zeta_2^2+y/2n^2)}{2\zeta_1(1-\zeta_1)-y/n^2}}_{\epsilon_2'}.
\end{aligned}\tag{55}$$

In this case, the update equation (8) becomes

$$\begin{aligned}\xi &= 4\tilde{t}(\tilde{P} - \tilde{\lambda}(J-I))(\psi^{(s)} - \frac{1}{2}\mathbf{1}) \\ &= 4\tilde{t}n\left(\left(\zeta_1 - \frac{1}{2}\right)\left(\frac{p_0+q_0}{2} - \tilde{\lambda}\right)u_1 + \frac{p_0-q_0}{2}\zeta_2 u_2\right) + 4\tilde{t}(\tilde{\lambda} - p_0)\left(\psi - \frac{1}{2}\mathbf{1}\right) \\ &:= n\tilde{a} + \tilde{b} \end{aligned} \tag{56}$$

where $\tilde{\lambda}$ and $\tilde{t}$ are defined in terms of $\tilde{p}$ and $\tilde{q}$. Since ψ is a stationary point, the above update gives $\psi = g(\xi)$.

We consider the following cases.

Case 1: $\zeta_2^2 = \Omega(1)$. Since $\zeta_1(1-\zeta_1) \geq \zeta_2^2$, it is easy to see that (55) implies that $\tilde{p} > \frac{p_0+q_0}{2} > \tilde{q}$, thus $\tilde{p} - \tilde{q} = \Omega(\rho_n)$, $\tilde{t} = \Omega(1)$, $\tilde{p} < \tilde{\lambda} < \tilde{q}$. It follows then $\tilde{b}_i = O(\rho_n)$, and $|\tilde{a}_i| = \Omega(\rho_n)$ for $i \in \mathcal{C}_1$ or $i \in \mathcal{C}_2$ (or both). In any of these cases, $\|w\| = O(\rho_n\sqrt{n}) = o(\sqrt{n})$.

Case 2: $\zeta_2 = o(1)$. Note that $\psi^T(\mathbf{1}-\psi) \geq 0$ implies that $\zeta_1(1-\zeta_1) - \frac{\|w\|^2}{n} \geq \zeta_2^2$. If $\|w\|^2 = o(n)$, we are done. If $\|w\|^2 = \Omega(n)$, $\zeta_1(1-\zeta_1) = \Omega(1)$. In this case, $\tilde{p} = \frac{p_0+q_0}{2} + O(\rho_n\zeta_2^2)$, and similarly for $\tilde{q}$. It follows then that $\tilde{t} = O(\zeta_2^2) = o(1)$, $\tilde{\lambda} = \frac{p_0+q_0}{2} + o(\rho_n)$ (we defer the details to (59)- (63)). Also note that $\tilde{b}_i = O(\rho_n\zeta_2^2)$. When $n|\tilde{a}_i| \gg \tilde{b}_i$, $g(\xi_i) = g(n\tilde{a}_i) + o(1)$. Since $g(n\tilde{a}) \in \text{span}\{u_1, u_2\}$, this implies that $\|w\| = o(\sqrt{n})$. When $n|\tilde{a}_i| \asymp \tilde{b}_i$, $\xi_i = o(1)$, and so we have $\|w\| = o(\sqrt{n})$ again. ■

Proof [Proof of Theorem 16] Let $a = (p_0+q_0)/2$. By (10), define $\kappa_1 := 4t^{(1)}\left(\zeta_1 - \frac{1}{2}\right)(a - \lambda^{(1)})$ and $\kappa_2 = 4t^{(1)}\zeta_2\frac{p_0-q_0}{2}$. Consider the initial distribution $\psi_i^{(0)} \overset{iid}{\sim} f_\mu$, where f is a distribution supported on $(0,1)$ with mean μ. Note that we have the following:

$$\begin{aligned}\zeta_1 &= \frac{(\psi^{(0)})^T\mathbf{1}}{n} = \mu + O_P(1/\sqrt{n}), \\ \zeta_2 &= \frac{(\psi^{(0)})^T u_2}{n} = O_P(1/\sqrt{n}).\end{aligned} \tag{57}$$

Now using (12), recall that

$$\begin{aligned}p^{(1)} &= \frac{p_0+q_0}{2} + \underbrace{\underbrace{\frac{(p_0-q_0)(\zeta_2^2 - x/2n^2)}{\zeta_1^2 + (1-\zeta_1)^2 - x/n^2}}_{\epsilon_1'} + O_P(\sqrt{\rho_n}/n)}_{\epsilon_1}, \\ q^{(1)} &= \frac{p_0+q_0}{2} - \underbrace{\underbrace{\frac{(p_0-q_0)(\zeta_2^2 + y/2n^2)}{2\zeta_1(1-\zeta_1) - y/n^2}}_{\epsilon_2'} - O_P(\sqrt{\rho_n}/n)}_{\epsilon_2}.\end{aligned} \tag{58}$$

This gives

$$\epsilon_1 = \epsilon_1' + O_P\left(\frac{\sqrt{\rho_n}}{n}\right) = O_P\left(\frac{\rho_n}{n}\right) + O_P\left(\frac{\sqrt{\rho_n}}{n}\right) = O_P\left(\frac{\sqrt{\rho_n}}{n}\right),$$

$$\epsilon_2 = \epsilon_2' + O_P\left(\frac{\sqrt{\rho_n}}{n}\right) = O_P\left(\frac{\sqrt{\rho_n}}{n}\right).$$

We will use the following logarithmic inequalities for $a > \epsilon > 0$:

$$\frac{2\epsilon}{a+\epsilon} \le \log\frac{a+\epsilon}{a-\epsilon} \le \frac{2\epsilon}{a-\epsilon}. \tag{59}$$

Now we have

$$\begin{aligned} t^{(1)} &= \frac{1}{2}\left(\log\left(\frac{a+\epsilon_1}{a-\epsilon_2}\right) + \log\left(\frac{1-a+\epsilon_2}{1-a-\epsilon_1}\right)\right), \\ 2t^{(1)} &\ge \frac{\epsilon_1+\epsilon_2}{a+\epsilon_1} + \frac{\epsilon_1+\epsilon_2}{1-a+\epsilon_2} \ge \frac{(\epsilon_1+\epsilon_2)}{(a+\epsilon_1)(1-a+\epsilon_2)}, \\ 2t^{(1)} &\le \frac{(\epsilon_1+\epsilon_2)}{(a-\epsilon_2)(1-a-\epsilon_1)}. \end{aligned} \tag{60}$$

For $\lambda^{(1)}$, if $\epsilon_1 + \epsilon_2 \ge 0$, we have

$$\lambda^{(1)} = \frac{\log\frac{1-q^{(1)}}{1-p^{(1)}}}{\log\frac{p^{(1)}}{q^{(1)}} + \log\frac{1-q^{(1)}}{1-p^{(1)}}} \le \frac{\epsilon_1+\epsilon_2}{1-a-\epsilon_1}\Big/\left(\frac{\epsilon_1+\epsilon_2}{a+\epsilon_1} + \frac{\epsilon_1+\epsilon_2}{1-a-\epsilon_1}\right) = a+\epsilon_1. \tag{61}$$

$$\lambda^{(1)} \ge \frac{\epsilon_1+\epsilon_2}{1-a+\epsilon_2}\Big/\left(\frac{\epsilon_1+\epsilon_2}{a-\epsilon_2} + \frac{\epsilon_1+\epsilon_2}{1-a+\epsilon_2}\right) = a-\epsilon_2. \tag{62}$$

If $\epsilon_1 + \epsilon_2 \le 0$,

$$\begin{aligned} \lambda^{(1)} &= \frac{\log\frac{1-q^{(1)}}{1-p^{(1)}}}{\log\frac{p^{(1)}}{q^{(1)}} + \log\frac{1-q^{(1)}}{1-p^{(1)}}} \ge \frac{\epsilon_1+\epsilon_2}{1-a-\epsilon_1}\Big/\left(\frac{\epsilon_1+\epsilon_2}{a+\epsilon_1} + \frac{\epsilon_1+\epsilon_2}{1-a-\epsilon_1}\right) = a+\epsilon_1, \\ \lambda^{(1)} &\le \frac{\epsilon_1+\epsilon_2}{1-a+\epsilon_2}\Big/\left(\frac{\epsilon_1+\epsilon_2}{a-\epsilon_2} + \frac{\epsilon_1+\epsilon_2}{1-a+\epsilon_2}\right) = a-\epsilon_2. \end{aligned} \tag{63}$$

The above analysis shows $t^{(1)} = O_P(\frac{1}{n\sqrt{\rho_n}})$, $|a - \lambda^{(1)}| = O_P(\frac{\sqrt{\rho_n}}{n})$.

We next try to generalize the above calculations for any iteration s. For convenience we assume A has self loops, which makes no difference to the asymptotics. Note that, for some $|\xi'| < \xi$, since $g''(0) = 0$,

$$\psi = g(\xi) = \frac{1}{2} + \frac{1}{4}\xi + g'''(\xi')\frac{\xi^3}{3!} = \frac{1}{2} + \frac{1}{4}\xi + O(\xi^3) \tag{64}$$

using the fact that $g'''(\xi) = O(1)\ \forall \xi$. Substituting, we have:

$$\begin{aligned} \zeta_1^{(s)} &= \frac{1}{n}\left\langle \psi^{(s)}, \mathbf{1}\right\rangle = \frac{1}{2} + \frac{1}{4n}\left\langle \xi^{(s)}, \mathbf{1}\right\rangle + O\left(\frac{\|(\xi^{(s)})^3\|_2}{\sqrt{n}}\right) \\ &= \frac{1}{2} + \frac{t^{(s)}}{n}\left\langle (A - \lambda^{(s)}J)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}), \mathbf{1}\right\rangle + O\left(\frac{\|(\xi^{(s)})^3\|_2}{\sqrt{n}}\right), \end{aligned} \tag{65}$$

using the update equation for $\xi^{(s)}$ in (11) and assuming A has self loops for convenience. Here using the decomposition $A = P + (A - P)$,

$$\begin{aligned}\left\langle (P - \lambda^{(s)} J)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}), \mathbf{1} \right\rangle &= n^2 \left(\frac{p_0 + q_0}{2} - \lambda^{(s)} \right) (\zeta_1^{(s-1)} - 1/2) \\ \left\langle (A - P)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}), \mathbf{1} \right\rangle &\leq \sqrt{n} \|A - P\|_{op} \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 \\ &= O_P(\sqrt{n^2 \rho_n}) \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2,\end{aligned}$$

where the first line follows from $P - \lambda^{(s)} J = \left(\frac{p_0 + q_0}{2} - \lambda^{(s)} \right) \mathbf{1}\mathbf{1}^T + \frac{p_0 - q_0}{2} u_2 u_2^T$. It follows then

$$\begin{aligned}&|\zeta_1^{(s)} - \frac{1}{2}| \\ \leq &\frac{4|t^{(s)}|}{n} \left(n^2 \left(\frac{p_0 + q_0}{2} - \lambda^{(s)} \right) |\zeta_1^{(s-1)} - 1/2| + O_P(\sqrt{n^2 \rho_n}) \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 \right) + O\left(\frac{\|(\xi^{(s)})^3\|_2}{\sqrt{n}} \right) \\ = &4|t^{(s)}| (n \left(\frac{p_0 + q_0}{2} - \lambda^{(s)} \right) |\zeta_1^{(s-1)} - 1/2| + O_P(\sqrt{\rho_n}) \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2) + O\left(\frac{\|\xi^{(s)}\|_2^3}{\sqrt{n}} \right)\end{aligned} \tag{66}$$

since $\|v^3\|_2 = \sqrt{\sum_i v_i^6} \leq \|v\|_2 \|v\|_\infty^2 \leq \|v\|_2^3$ for any v. Similarly, we have:

$$\begin{aligned}\zeta_2^{(s)} &= \frac{1}{n} \left\langle \psi^{(s)}, u_2 \right\rangle = \frac{1}{4n} \left\langle \xi^{(s)}, u_2 \right\rangle + O\left(\frac{\|(\xi^{(s)})^3\|_2}{\sqrt{n}} \right), \\ &= \frac{t^{(s)}}{n} \left\langle (A - \lambda^{(s)} J)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}), u_2 \right\rangle + O\left(\frac{\|(\xi^{(s)})^3\|_2}{\sqrt{n}} \right),\end{aligned} \tag{67}$$

$$\begin{aligned}|\zeta_2^{(s)}| &\leq \frac{|t^{(s)}|}{n} \left(\frac{n^2 (p_0 - q_0)}{2} |\zeta_2^{(s-1)}| + O_P(\sqrt{n^2 \rho_n}) \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 \right) + O\left(\frac{\|(\xi^{(s)})^3\|_2}{\sqrt{n}} \right) \\ &= |t^{(s)}| (O(n\rho_n) |\zeta_2^{(s-1)}| + O_P(\sqrt{\rho_n}) \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2) + O\left(\frac{\|\xi^{(s)}\|_2^3}{\sqrt{n}} \right)\end{aligned} \tag{68}$$

For the norm of $\xi^{(s)}$,

$$\begin{aligned}\|\xi^{(s)}\|_2 \leq 4|t^{(s)}| \Bigg(&n^{3/2} \left(\left| \left(\frac{p_0 + q_0}{2} - \lambda^{(s)} \right) (\zeta_1^{(s-1)} - 1/2) \right| + O(\rho_n) |\zeta_2^{(s-1)}| \right) \\ &+ O_P(\sqrt{n\rho_n}) \|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 \Bigg)\end{aligned} \tag{69}$$

using the same eigen-decomposition on P.

To bound $t^{(s)}$, we can first define $\epsilon_1^{(s)}$ and $\epsilon_2^{(s)}$ in the same way as (58), where the order terms come from the second part of (54) (and an analogous equation for $q^{(1)}$, with

general $\psi^{(s-1)}$ replacing ψ). Then provided $\zeta_1^{(s-1)}$ is bounded away from 0 and 1, and $\epsilon_1^{(s)}, \epsilon_2^{(s)} = o_P(\rho_n)$, by (60),

$$|t^{(s)}| = O_P(\zeta_2^{(s-1)})^2 + O(\frac{1}{n^2\rho_n})\left((\psi^{(s-1)})^T(A-P)\psi^{(s-1)} + (\mathbf{1}-\psi^{(s-1)})^T(A-P)(\mathbf{1}-\psi^{(s-1)})\right)$$
$$+ O_P(\frac{1}{n^2\rho_n})(\psi^{(s-1)})^T(A-P)(\mathbf{1}-\psi^{(s-1)}), \tag{70}$$

where for any ψ,

$$\begin{aligned}\psi^T(A-P)\psi &= \frac{1}{4}\mathbf{1}^T(A-P)\mathbf{1} + \mathbf{1}^T(A-P)(\psi - \frac{1}{2}\mathbf{1}) + (\psi-\frac{1}{2}\mathbf{1})^T(A-P)(\psi-\frac{1}{2}\mathbf{1})\\ &= O_P(\sqrt{n^2\rho_n})(1+\|\psi-\frac{1}{2}\mathbf{1}\|_2) + O_P(\sqrt{n\rho_n})\|\psi-\frac{1}{2}\mathbf{1}\|_2^2\\ &= O_P(\sqrt{n^2\rho_n})(1+\|\psi-\frac{1}{2}\mathbf{1}\|_2)\end{aligned}$$

since $\|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\| \le \sqrt{n}$. Similarly

$$\begin{aligned}\psi^T(A-P)\mathbf{1} &= (\psi-\frac{1}{2}\mathbf{1})^T(A-P)\mathbf{1} + \frac{1}{2}\mathbf{1}^T(A-P)\mathbf{1}\\ &= O_P(\sqrt{n^2\rho_n})(1+\|\psi-\frac{1}{2}\mathbf{1}\|_2).\end{aligned}$$

The upper bound on $t^{(s)}$ becomes:

$$|t^{(s)}| = O_P\left(|\zeta_2^{(s-1)}|^2\right) + O_P\left(\frac{1}{\sqrt{n^2\rho_n}}\right)(1+\|\psi^{(s-1)}-\frac{1}{2}\mathbf{1}\|_2)$$

In a similar way to bound $\lambda^{(s)}$, note that defining general $\epsilon_1^{(s)}, \epsilon_2^{(s)}$ in (60)-(63), as long as $\zeta_1^{(s-1)}$ is bounded away from 0 and 1, and $\epsilon_1^{(s)}, \epsilon_2^{(s)} = o_P(\rho_n)$, we have:

$$\left|\frac{p_0+q_0}{2} - \lambda^{(s)}\right| = O_P\left(\rho_n t^{(s)}\right)$$

Finally,

$$\begin{aligned}\|\psi^{(s)} - \frac{1}{2}\mathbf{1}\|_2 &= \frac{1}{4}\|\xi^{(s)}\|_2 + O\left(\|(\xi^{(s)})^3\|_2\right)\\ &= \frac{1}{4}\|\xi^{(s)}\|_2 + O\left(\|(\xi^{(s)})\|_2^3\right)\end{aligned} \tag{71}$$

For $s=1$, we have the following:

$$\begin{aligned}&t^{(1)} = O_P(\frac{1}{n\sqrt{\rho_n}}), \frac{p_0+q_0}{2} - \lambda^{(1)} = O_P(\frac{\sqrt{\rho_n}}{n}),\\ &\|\xi^{(1)}\|_2, \|\psi^{(1)} - \frac{1}{2}\mathbf{1}\|_2 = O_P(1),\\ &|\zeta_1^{(1)} - 1/2|, |\zeta_2^{(1)}| = O_P\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

where the second line follows from (69), (71), noting $\zeta_1^{(0)} = O_P(1)$, $\zeta_2^{(0)} = O_P(n^{-1/2})$. The last line follows from (66) and (68).

For $s = 2$, note that the above bounds imply $\zeta_1^{(1)}$ is bounded away from 0 and 1, and $\epsilon_1^{(s)}, \epsilon_2^{(s)} = o_P(\rho_n)$. Using the same set of equations again, we have:

$$\begin{aligned} & t^{(2)} = O_P(\frac{1}{n\sqrt{\rho_n}}), \frac{p_0+q_0}{2} - \lambda^{(2)} = O_P(\frac{\sqrt{\rho_n}}{n}) \\ & \|\xi^{(2)}\|_2, \|\psi^{(2)} - \frac{1}{2}\mathbf{1}\|_2 = O_P(\sqrt{\rho_n}) \\ & |\zeta_1^{(2)} - 1/2|, |\zeta_2^{(2)}| = O_P\left(\sqrt{\frac{\rho_n}{n}}\right) \end{aligned} \tag{72}$$

In general, once $\|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 = O_P(1)$, $|\zeta_1^{(s-1)} - 1/2|$ and $|\zeta_2^{(s-1)}| = O_P(1/\sqrt{n})$ we have $t^{(s)} = O_P(1/n\sqrt{\rho_n})$, $(p_0+q_0)/2 - \lambda^{(s)} = O_P(\sqrt{\rho_n}/n)$, $\|\xi^{(s)}\|_2 = O_P(\sqrt{\rho_n})$, $|\zeta_1^{(s)} - 1/2|, |\zeta_2^{(s)}|$ are both $o_P(1/\sqrt{n})$ and $\|\psi^{(s)} - \frac{1}{2}\mathbf{1}\|_2 = O_P(\sqrt{\rho_n})$.

We can further derive a contraction result from $s = 2$ onward. Since the rates in (72) hold for $s \geq 2$, and applying (64) to $\psi^{(s)}$,

$$\begin{aligned} \|\psi^{(s)} - \frac{1}{2}\mathbf{1}\|_2 &\leq \frac{1}{4}\|\xi^{(s)}\|_2 + O(\|(\xi^{(s)})^3\|_2) \\ &\leq \frac{1}{4}\|\xi^{(s)}\|_2 + O_P(\rho_n^{3/2}). \end{aligned} \tag{73}$$

For $\xi^{(s)}$,

$$\xi^{(s)} = 4t^{(s)}\left((P - \lambda^{(s)}J)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1}) + (A - P)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1})\right),$$

where $\|(P - \lambda^{(s)}J)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1})\|_2 = O_P(\rho_n)\|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2$, and $\|(A-P)(\psi^{(s-1)} - \frac{1}{2}\mathbf{1})\|_2 = O_P(\sqrt{n\rho_n})\|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2$. It follows from (73) and the rate of $t^{(s)}$,

$$\|\psi^{(s)} - \frac{1}{2}\mathbf{1}\|_2 \leq O_P(1/\sqrt{n})\|\psi^{(s-1)} - \frac{1}{2}\mathbf{1}\|_2 + O_P(\rho_n^{3/2}).$$

■

Proof [Proof of Theorem 17] Under the current initialization,

$$\begin{aligned} \zeta_1 &= \frac{1}{2} + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^{(0)}, \\ \zeta_2 &= \Delta\mu + \frac{1}{n}\sum_{i\in\mathbf{1}_{\mathcal{C}_1}}\epsilon_i^{(0)} - \frac{1}{n}\sum_{i\in\mathbf{1}_{\mathcal{C}_2}}\epsilon_i^{(0)}. \end{aligned} \tag{74}$$

Define the event $\mathcal{A}_1 = \{\frac{1}{n}\sum_{i\in\mathbf{1}_{\mathcal{C}_1}}\epsilon_i^{(0)}, \frac{1}{n}\sum_{i\in\mathbf{1}_{\mathcal{C}_2}}\epsilon_i^{(0)} = O(\sqrt{\log n/n})\}$, then by Bernstein's inequality, this event happens with probability at least $1 - \exp(-\Theta(\log n))$.

Recall ϵ_1 and ϵ_2 from Eq (58). Define $\mathcal{A}_2 = \{(\psi^{(0)})^T(A-\tilde{P})\psi^{(0)}, (1-\psi^{(0)})^T(A-\tilde{P})(1-\psi^{(0)}) = O(n\sqrt{\rho_n \log n})\}$. By a similar Bernstein's inequality, $\mathcal{A}_2$ has probability at least $1-\exp(-\Theta(\log n))$. Then under $\mathcal{A}_1 \cap \mathcal{A}_2$,

$$\epsilon_1 = (p_0-q_0)\frac{\zeta_2^2 - \frac{x}{2n^2}}{\frac{1}{2}+O(\sqrt{\log n/n}) - \frac{x}{n^2}} + O\left(\frac{\sqrt{\rho_n \log n}}{n}\right)$$
$$\epsilon_2 = (p_0-q_0)\frac{\zeta_2^2 + \frac{y}{2n^2}}{\frac{1}{2}+O(\sqrt{\log n/n}) - \frac{y}{n^2}} + O\left(\frac{\sqrt{\rho_n \log n}}{n}\right),$$

using (74) and the same decomposition as in (54). The lower bound (15) on $\Delta\mu$ implies $\zeta_2^2 \gg \sqrt{\log n/n}$, it follows $0 < \epsilon_1, \epsilon_2 < a$ since $|\zeta_2| \le \frac{1}{2}$, and $\epsilon_1, \epsilon_2 = \Theta((p_0-q_0)\zeta_2^2)$. Then by (60)-(63),

$$t^{(1)} = \Theta(\frac{\epsilon_1+\epsilon_2}{\rho_n}) = \Theta\left((p_0-q_0)\zeta_2^2/\rho_n\right)$$
$$|a-\lambda^{(1)}| \le \max\{\epsilon_1, \epsilon_2\}. \tag{75}$$

Next define

$$\kappa_1 = 4t^{(1)}(\zeta_1 - \frac{1}{2})(a-\lambda^{(1)})$$
$$\kappa_2 = 4t^{(1)}\zeta_2\frac{(p_0-q_0)}{2}. \tag{76}$$

Using (74) - (76), under $\mathcal{A}_1 \cap \mathcal{A}_2$,

$$\kappa_1+\kappa_2 = 4t^{(1)}\left(\Delta\mu \cdot \frac{p_0-q_0}{2} + O(\rho_n\sqrt{\log n/n})\right),$$
$$\kappa_1-\kappa_2 = 4t^{(1)}\left(-\Delta\mu \cdot \frac{p_0-q_0}{2} + O(\rho_n\sqrt{\log n/n})\right).$$

From (10) and adding the noise term from the sample version of the update,

$$\xi_i^{(1)} = n(\kappa_1+\sigma_i\kappa_2) + b_i^{(0)} + nr_i^{(0)}, \tag{77}$$

In (77), $b_i^{(0)}$ is of smaller order than the other terms and it suffices to consider $n(\kappa_1+\sigma_i\kappa_2+r_i^{(0)})$. By the argument in (27), $\mathcal{A}_3 = \left\{\max_i |r_i^{(0)}| = O\left(\sqrt{\frac{\rho_n}{n}\log n}\right)\right\}$ has probability at least $1-\exp(-\Theta(\log n))$. For any pair $i \in \mathcal{C}_1$ and $j \in \mathcal{C}_2$, under $\cap_{k=1}^3 \mathcal{A}_k$, we have

$$\begin{aligned}
&(\kappa_1+\kappa_2+r_i^{(0)})(\kappa_1-\kappa_2+r_j^{(0)}) \\
\le&(\kappa_1^2-\kappa_2^2) + O\left(\max(|r_i^{(0)}|, |r_j^{(0)}|)\max(|\kappa_1|, |\kappa_2|)\right) \\
\le&16(t^{(1)})^2\left(-(\Delta\mu)^2\frac{(p_0-q_0)^2}{4} + O(\rho_n^2\sqrt{\log n/n})\right) + 2t^{(1)}|\zeta_2|(p_0-q_0)O(\sqrt{\frac{\rho_n}{n}\log n}) \\
\le&Ct^{(1)}(p_0-q_0)|\Delta\mu|\left(-t^{(1)}|\Delta\mu|(p_0-q_0) + O(\sqrt{\frac{\rho_n}{n}\log n})\right)
\end{aligned}$$

$$\leq Ct^{(1)}(p_0-q_0)|\Delta\mu|\left(-|\Delta\mu|^3(p_0-q_0)^2/\rho_n+O(\sqrt{\frac{\rho_n}{n}\log n})\right)<0$$

for $|\Delta\mu| > C\left(\frac{\rho_n^{3/2}\sqrt{\log n}}{(p_0-q_0)^2\sqrt{n}}\right)^{1/3} = \Theta\left(\sqrt{\frac{\log n}{n\rho_n}}\right)^{1/3} \gg \left(\frac{\log n}{n}\right)^{1/4}$ for some general constant C large enough, independent of n and model parameters. Thus $n(\kappa_1+\kappa_2+r_i^{(0)})$ and $n(\kappa_1-\kappa_2+r_j^{(0)})$, for i,j in different blocks, have opposite signs.

We will now check if $n(\kappa_1+\sigma_i\kappa_2+r_i^{(0)})\to\infty$, and it suffices to lower bound $n(|\kappa_2|-|\kappa_1|-\max_i|r_i^{(0)}|)$. By (75), (76),

$$\begin{aligned} n(|\kappa_2|-|\kappa_1|-\max_i|r_i^{(0)}|) &\geq n\left(C|\Delta\mu|^3(p_0-q_0)-O(\sqrt{\frac{\rho_n}{n}\log n})\right)\\ &\geq \Omega(\sqrt{n\rho_n\log n})\to\infty \end{aligned}$$

Thus $n(\kappa_1+\sigma_i\kappa_2+r_i^{(0)})$ is growing to infinity with an order bounded below by $\Omega(\sqrt{n\rho_n\log n})$, with probability at least $1-\exp(-\Theta(\log n))$.

If $n(\kappa_1+\kappa_2+r_i^{(0)})>0$, since $\psi_i^{(1)}=g(n(\kappa_1+\sigma_i\kappa_2)+b_i^{(0)}+nr_i^{(0)})$, we have $\psi^{(1)}=\mathbf{1}_{\mathcal{C}_1}+O(\exp(-\Omega(\sqrt{n\rho_n\log n}))$ with probability at least $1-\exp(-\Theta(\log n))$. The case $\kappa_1+\kappa_2+r_i^{(0)}<0$ is similar.

For later iterations, WLOG assume $z_0=\mathbf{1}_{\mathcal{C}_1}$, and A has self-loops for convenience. First noting that the error term in $\psi^{(1)}$ is uniform, we can write $\|\psi^{(1)}-\mathbf{1}_{\mathcal{C}_1}\|_1=n\eta_n$, where $\eta_n=O(\exp(-\Omega(\sqrt{n\rho_n\log n})))$ with high probability. We have $\zeta_1^{(1)}=\frac{1}{2}+O(\eta_n)$, $\zeta_2^{(1)}=\frac{1}{2}-O(\eta_n)$. To obtain $p^{(2)},q^{(2)}$, we observe in Eq (58), $x=n-\Theta(n\eta_n)$, $y=\Theta(n\eta_n)$. Using the decomposition in (54), define $\mathcal{A}_4=\{\|A-P\|_{op}=O(\sqrt{n\rho_n})\}$ which has probability at least $1-n^{-r}$, $r>0$, then under $\mathcal{A}_4$,

$$|\psi^T(A-P)\psi|\leq\|A-P\|_{op}\|\psi\|_2^2=O_P(n^{3/2}\rho_n^{1/2})$$

for general ψ. It follows then

$$\begin{aligned} p^{(2)} &= \frac{p_0+q_0}{2}+\frac{p_0-q_0}{2}(1+O(\eta_n))+O(\sqrt{\rho_n/n})\\ &= p_0-O(\rho_n\eta_n)+O(\sqrt{\rho_n/n}),\\ q^{(2)} &= \frac{p_0+q_0}{2}-\frac{p_0-q_0}{2}(1+O(\eta_n))+O(\sqrt{\rho_n/n})\\ &= q_0+O(\rho_n\eta_n)+O(\sqrt{\rho_n/n}). \end{aligned} \tag{78}$$

To bound $t^{(2)}$ and $\lambda^{(2)}$, note that

$$\begin{aligned} \log\frac{p^{(2)}}{q^{(2)}} &= \log\frac{p^{(2)}-p_0+p_0}{q^{(2)}-q_0+q_0}=\log\frac{p_0}{q_0}+O(\eta_n)+O(1/\sqrt{n\rho_n}),\\ \log\frac{1-q^{(2)}}{1-p^{(2)}} &= \log\frac{1-q_0-q^{(2)}+q_0}{1-p_0-p^{(2)}+p_0}=\log\frac{1-q_0}{1-p_0}+O(\rho_n\eta_n)+O(\sqrt{\rho_n/n}), \end{aligned} \tag{79}$$

then

$$t^{(2)} = t_0 + O(\eta_n) + O(1/\sqrt{n\rho_n}),\ \lambda^{(2)} = \lambda_0 + O(\rho_n\eta_n) + O(\sqrt{\rho_n/n}). \tag{80}$$

To show that contraction starts from this iteration, we can use arguments similar to the last part of the proof of Theorem 9. Writing

$$\xi^{(2)} = 4t^{(2)}(P - \lambda^{(2)}J)(\psi^{(1)} - \frac{1}{2}\mathbf{1}) + 4t^{(2)}(A - P)(\psi^{(1)} - \frac{1}{2}\mathbf{1}), \tag{81}$$

the signal part is constant for $i \in \mathbf{1}_{\mathcal{C}_1}$ and $i \in \mathbf{1}_{\mathcal{C}_2}$. Denote

$$\begin{aligned} s_1^{(1)} &= (p_0 - \lambda^{(2)}) \sum_{i\in\mathcal{C}_1}(\psi_i^{(1)} - \frac{1}{2}) + (q_0 - \lambda^{(2)}) \sum_{i\in\mathcal{C}_2}(\psi_i^{(1)} - \frac{1}{2}) \\ s_2^{(1)} &= (q_0 - \lambda^{(2)}) \sum_{i\in\mathcal{C}_1}(\psi_i^{(1)} - \frac{1}{2}) + (p_0 - \lambda^{(2)}) \sum_{i\in\mathcal{C}_2}(\psi_i^{(1)} - \frac{1}{2}). \end{aligned}$$

It is easy to see that when n is large,

$$\begin{aligned} s_1^{(1)} &\geq (p_0 - q_0)\left(\frac{n}{4} - \frac{\eta_n n}{2}\right) + (\lambda_0 - \lambda^{(2)})\eta_n n \\ &\geq C_0(p_0 - q_0)n, \end{aligned} \tag{82}$$

for some general constant $C_0 < 1/4$, independent of n and model parameters. and similarly $s_2^{(1)} \leq -C_0(p_0 - q_0)n$. Next in Eq (40), taking $x_0 = C_0 t_0 n(p_0 - q_0)$, using (80), (82),

$$\begin{aligned} \mathbb{1}\{\xi_i^{(2)} \leq x_0\} &\leq \mathbb{1}\left\{4t^{(2)}s_1^{(1)} + 4t^{(2)}(A - P)_{i,\cdot}(z_0 - \frac{1}{2}\mathbf{1}) \leq 2x_0\right\} \\ &\quad + \mathbb{1}\left\{4t^{(2)}(A - P)_{i,\cdot}(\psi^{(1)} - z_0) \leq -x_0\right\} \\ &\leq \exp\left\{-2C_0 t_0(p_0 - q_0)n + O(\sqrt{n\rho_n \log n}) + O(n\rho_n\eta_n)\right\} \\ &\quad + \mathbb{1}\left\{4t^{(2)}(A - P)_{i,\cdot}(\psi^{(1)} - z_0) \leq -x_0\right\}. \end{aligned} \tag{83}$$

Then using the same argument as in Eq (43), for large enough n, under $\mathcal{A}_4$,

$$\|\psi^{(2)} - z_0\|_1 \leq n\exp(-c_1 t_0(p_0 - q_0)n) + \frac{c_2\rho_n}{(p_0 - q_0)^2 n}\|\psi^{(1)} - z_0\|_1.$$

for some constants c_1, c_2 independent of n and model parameters. The same argument works for later iterations under $\cap_{k=1}^4 \mathcal{A}_k$, thus the high probability statement holds uniformly for all iterations with probability at least $1 - n^{-r}$, $0 < r$.

■

Appendix D. Additional simulation

We compare the effectiveness of the random initialization $\psi_i^{(0)} \overset{iid}{\sim}$ Bernoulli(1/2) with informative initializations obtained from spectral clustering and SDP. For SDP, we use the algorithm in Li et al. (2018) with the tuning parameter selected using the method in Cai et al. (2015). In Figure 6(a), we set the average degree of each graph to 20, $n = 400$, and vary the p_0/q_0 ratio; smaller ratios mean weaker signal. We run BCAVI with three types of initializations: $\psi_i^{(0)} \overset{iid}{\sim}$ Bernoulli(1/2) (blue), $\psi_i^{(0)}$ set to the result of running spectral clustering (black), and $\psi_i^{(0)}$ set to the result of running SDP (red). The plot shows the mean accuracy and standard deviation from 20 random graphs at each point. As expected, BCAVI initialized with spectral clustering and SDP have higher accuracy, although random initializations have quite reasonable performance in high signal regimes. Figure 6(b) is similar with average degree set to 30. In this denser case, the performance of random initializations improve, and become very close to the other two methods.

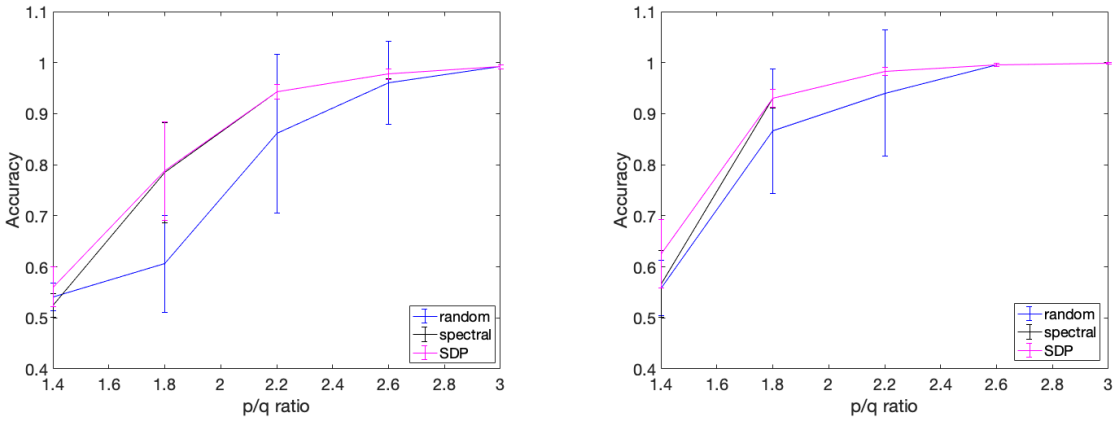


Figure 6: Average clustering accuracy of three types of initialization scheme for $n = 400$, average degree equals 20 (a) and 30 (b).

References

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.

Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, pages 1922–1943, 2013.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435. URL `http://dl.acm.org/`

citation.cfm?id=944919.944937.

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.

T Tony Cai, Xiaodong Li, et al. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, 2015.

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 Giordano, Tamara Broderick, and Michael I Jordan. Covariances, robustness and variational bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029, 2018.

Jake M. Hofman and Chris H. Wiggins. Bayesian approach to network modularity. *Physical review letters*, 100(25):258701, 2008.

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

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.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Advances in Neural Information Processing Systems*, pages 4116–4124, 2016.

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.

Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Xiaodong Li, Yudong Chen, and Jiaming Xu. Convex relaxation methods for community detection. *arXiv preprint arXiv:1810.00315*, 2018.

Song Mei, Yu Bai, Andrea Montanari, et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

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.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational Bayes. In *AISTATS*, 2018.

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

Flemming Topsøe. Some bounds for the logarithmic function. *Research report collection*, 7 (2), 2004. URL `http://vuir.vu.edu.au/17162/`.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL `http://dx.doi.org/10.1561/2200000001`.

Bo Wang and D. M. Titterington. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 577–584. AUAI Press, 2004.

Bo Wang and D. M. Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *AISTATS*, 2005.

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

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

T Westling and TH McCormick. Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics*, pages 1–12, 2019.

Burton Wu, Clare A McGrory, and Anthony N Pettitt. A new variational bayesian algorithm with application to human mobility pattern modeling. *Statistics and Computing*, 22(1): 185–203, 2012.

Ji Xu, Daniel J. Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2676–2684. Curran Associates, Inc., 2016.

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.