Journal of Machine Learning Research 21 (2020) 1-68 Submitted 4/19; Revised 12/19; Published 1/20

Optimal Bipartite Network Clustering

Zhixin Zhou ZHIXIN@UCLA.EDU
Department of Statistics
University of California
Los Angeles, CA 90095-1554, USA

Arash A. Amini AAAMINI@UCLA.EDU
Department of Statistics
University of California
Los Angeles, CA 90095-1554, USA

Editor: Matthias Hein

Abstract

We study bipartite community detection in networks, or more generally the network biclustering problem. We present a fast two-stage procedure based on spectral initialization followed by the application of a pseudo-likelihood classifier twice. Under mild regularity conditions, we establish the weak consistency of the procedure (i.e., the convergence of the misclassification rate to zero) under a general bipartite stochastic block model. We show that the procedure is optimal in the sense that it achieves the optimal convergence rate that is achievable by a biclustering oracle, adaptively over the whole class, up to constants. This is further formalized by deriving a minimax lower bound over a class of biclustering problems. The optimal rate we obtain sharpens some of the existing results and generalizes others to a wide regime of average degree growth, from sparse networks with average degrees growing arbitrarily slowly to fairly dense networks with average degrees of order $\sqrt{n}$. As a special case, we recover the known exact recovery threshold in the $\log n$ regime of sparsity. To obtain the consistency result, as part of the provable version of the algorithm, we introduce a sub-block partitioning scheme that is also computationally attractive, allowing for distributed implementation of the algorithm without sacrificing optimality. The provable algorithm is derived from a general class of pseudo-likelihood biclustering algorithms that employ simple EM type updates. We show the effectiveness of this general class by numerical simulations.

Keywords: Bipartite networks; stochastic block model; community detection; biclustering; network analysis; pseudo-likelihood; spectral clustering.

1. Introduction

Network analysis has become an active area of research over the past few years, with applications and contributions from many disciplines including statistics, computer science, physics, biology and social sciences. A fundamental problem in network analysis is detecting and identifying communities, also known as clusters, to help better understand the underlying structure of the network. The problem has seen rapid advances in recent years with numerous breakthroughs in modeling, theoretical understanding, and practical applications (Fortunato and Hric, 2016). In particular, there has been much excitement and progress in understanding and analyzing the stochastic block model (SBM) and its variants. We refer to Abbe (2017) for a recent survey of the field. Much of this effort, especially on the theoretical side has been focused on the unipartite (or symmetric) case, while the bipartite counterpart, despite numerous practical applications, has received comparatively less attention. Of course, there has been lots of activity in terms of modeling and algorithm development for bipartite clustering both in the context of networks (Zhou et al., 2007; Larremore et al., 2014; Wyse et al., 2017; Rohe et al., 2012; Razaee et al., 2019) as well as other domains such as topic modeling and text mining (Dhillon, 2001, 2003) and biological applications (Cheng and Church, 2000; Madeira et al., 2010). But much of this work either lacks theoretical investigations or has not considered the issue of statistical optimality.

In this paper, we consider the community detection, or clustering, in the bipartite setting with a focus on deriving fundamental theoretical limits of the problem. The main goal is to propose computationally feasible algorithms for bipartite network clustering that exhibit provable statistical optimality. We will focus on the bipartite version of the SBM which is a natural model for bipartite networks with clusters. SBM is a stochastic network model where the probability of edge formation depends on the latent (unobserved) community assignment of the nodes, often referred to as node labels. The goal of the community detection problem is to recover these labels given an instance of the network. This is a non-trivial task since, for example, maximum likelihood estimation involves a search over exponentially many labels.

Community detection in bipartite SBM is closely related to the biclustering problem, for which many algorithms have been developed over the years (Hartigan, 1972; Cheng and Church, 2000; Tanay et al., 2002; Gao et al., 2016). On the other hand, in recent years, various algorithms have been proposed for clustering in unipartite SBMs, including global approaches such as spectral clustering (Rohe et al., 2011; Krzakala et al., 2013; Lei et al., 2015; Fishkind et al., 2013; Vu, 2018; Massoulié, 2014; Yun and Proutiere, 2014; Chin et al., 2015; Bordenave et al., 2015; Gulikers et al., 2017; Pensky et al., 2019; Zhou and Amini, 2019) and convex relaxations via semidefinite programs (SDPs) (Amini et al., 2018; Hajek et al., 2016a; Bandeira, 2015; Guédon and Vershynin, 2016; Montanari and Sen, 2016; Ricci-Tersenghi et al., 2016; Agarwal et al., 2017; Perry and Wein, 2017), as well as local methods such as belief propagation (Decelle et al., 2011), Bayesian MCMC (Nowicki and Snijders, 2001) and variational Bayes (Celisse et al., 2012; Bickel et al., 2013), greedy profile likelihood (Bickel and Chen, 2009; Zhao et al., 2012) and pseudo-likelihood maximization (Amini et al., 2013), among others. A limitation of spectral clustering approaches is that they are often not optimal on their own, and the SDPs have the drawback of not being able to fit the full generality of SBMs.

Various algorithms can further improve the clustering accuracy, and adapt to the generality of SBM. Profile likelihood maximization was proposed and analyzed in Bickel and Chen (2009), but the underlying optimization problem is computationally infeasible and the approach only applicable to networks of limited size. Pseudo-likelihood ideas were used in Amini et al. (2013) to derive EM type updates to maximize a surrogate to the likelihood of the SBM. We extend the ideas of Amini et al. (2013) to the bipartite setting and greatly improve their analysis by showing that these pseudo-likelihood approaches can achieve minimax optimal rates in a wide variety of settings.

In the unipartite setting, there has been interesting recent advancements in understanding optimal recovery in the semi-sparse regime where the (expected) average network degree, d_{av}, is allowed to grow slowly to infinity as the number of nodes, n, grows. In a series of papers (Mossel et al., 2015; Abbe et al., 2016; Hajek et al., 2016a,b), the optimal threshold for exact recovery, also known as strong consistency, was established for simple planted partition models. In Abbe and Sandon (2015), the problem of strong consistency was considered for a general SBM and the optimal threshold for strong consistency was established. In subsequent work (Zhang et al., 2016; Gao et al., 2017, 2018), the results were extended to include weak consistency, i.e., requiring the fraction of misclassified nodes to go to zero, rather than drop to exactly zero (as in strong consistency), and rates of optimal convergence were established, up to a slack in the exponent. To achieve the more relaxed consistency results, Gao et al. (2017) limited the model to what we refer to as strongly assortative SBM; see Amini et al. (2018) for a definition.

Our work is inspired by the insightful analysis of Abbe and Sandon (2015) and Gao et al. (2017). We extend these ideas by presenting results that are strictly sharper and more general that what has been obtained so far. In short, we only assume that the clusters are distinguishable (in the sense of Chernoff divergence) and the network is not very dense, i.e. $d_{\text{av}} = O(\sqrt{n})$ where d_{av} denotes the expected average degree, and n is the number of nodes. Our results establish minimax optimal rates below this $\sqrt{n}$ regime and above the sparse regime $d_{\text{av}} = O(1)$. In particular, we obtain precise rates of weak consistency when d_{av} grows arbitrarily slowly. We require $d_{\text{av}} = O(\sqrt{n})$ to allow Poisson approximations on the degrees of nodes restricted to large subsets. In the regimes denser than $\sqrt{n}$, existing work easily establishes exact recovery under suitable distinguishability conditions. We make more detailed comparisons with existing work in Section 3.

Contributions. Establishing these results require a fair amount of technical and algorithmic novelty. Here, we highlight some of the features of our approach:

1. Existing minimax rates of convergence for the misclassification error are known for what we refer to as the *nearly assortative model* where the probability of connection is $\geq a/n$ within clusters and $\leq b/n$ outside clusters. The existing results establish an error rate that belongs to an interval:

$$\text{Error} \in \left[e^{-(1+o(1))I}, e^{-(1-o(1))I}\right], \quad \text{as } I \to \infty,$$

for some $o(1)$ terms that are positive. Here, I is related to the Bhattacharyya distance (also known as the Hellinger affinity) of Bernoulli variables with probabilities a/n and b/n. This type of result originally appeared in Zhang et al. (2016) and propagated to many subsequent works (Gao et al., 2018, 2017; Zhang and Zhou, 2017; Xu et al., 2017; Chien et al., 2018). This rate, however, is not sharp since the slack term $e^{o(1)I}$ could be unbounded (because $o(1) > 0$ and $I \to \infty$). Another shortcoming of these results are their limitations to the simple nearly assortative setting. Our result sharpens and generalizes this known minimax rate to

$$\text{Error} = e^{-I-R}, \quad \text{for some } 0 < R \asymp \log I$$

for the general class of all SBMs under a mild distinguishably assumption on the rows and columns of the edge probability matrix. Furthermore, the I in our result takes the form of a Chernoff exponent among Poisson vectors, which is the form necessary for the general SBM.

2. In order to achieve these sharp rates, we introduce an efficient sub-block (or sub-graph) partitioning scheme that generalizes the partitioning idea of Chin et al. (2015). Our scheme allows one to break down the costly spectral initialization, by applying it to smaller subblocks, without losing optimality. If done in parallel, spectral clustering on subblocks is computationally cheaper than performing a spectral decomposition of the entire matrix. The resulting algorithm is naturally parallelizable, hence can be deployed in a *distributed fashion* and scaled to very large networks.

3. Our algorithms are modifications of a natural EM algorithm on mixtures of Poisson vectors, hence very familiar from a statistical perspective. Although other (optimal) algorithms in the literature are preforming more or less similar operations, the link to EM algorithms and mixture modeling is quite clear in our work. In Section 4.2, we provide the general blueprint of the algorithms based on the pseudo-likelihood and block compression ideas (Algorithm 1). We then construct a provable version by combining with the sub-block partitioning idea (Section 5).

4. To get the sharper rate, analyzing a single step of an EM algorithm is not enough; we thus analyze the second step too. We will show that the first step gets us from a good (but crude) initial rate γ_1 to the fast rate $\approx \exp(-I/Q)$ where Q is the number of subblocks. This rate is in the vicinity of the optimal rate and repeating the iteration once more, with the more accurate labels gets us to the minimax error rate $\exp(-I-R)$.

5. Among the technical contributions are the uniform consistency of the likelihood ratio classifier (LRC) over a subset of the parameters close to the truth (Lemma 2), sharp approximations for the Poisson-binomial distributions (Appendix F.4), and the extension (and elucidation) of a novel technique of Abbe and Sandon (2015) for deriving error exponents (Appendix F.3). The uniform consistency result for LRCs allows us to tolerate some dependence among the estimates from iteration to iteration (Sections 5 and Appendix B.1).

6. In contrast to the unipartite case, the bipartite setting allows us to introduce an oracle version of the problem that reveals the nature of the optimal rates in community detection and how they relate to classical hypothesis testing and mixture modeling. In particular, we answer the curious question of why the Chernoff exponent of a (binary) hypothesis testing problem controls the misclassification rate in community detection and network clustering. The oracle also provides an upper bound on the performance (i.e., a lower bound on the misclassification rate) of any algorithm. See Section 1 and Proposition 1 for details.

The rest of the paper is organized as follows. We introduce the model and the biclustering oracle in Section 2. We present our main results in Section 3, including an upper bound on the error rate of the algorithm and a matching minimax lower bound. The general pseudo-likelihood algorithms are presented in Section 4 and a provable version in Section 5. In Section 6, we demonstrate the numerical performance of the methods. The proofs appear in Appendices B through F. Extra comments on the results and proof techniques can be found in Appendix A.

2. Network biclustering

We start by introducing the network biclustering problem based on stochastic block modeling, and set up some notation. We then discuss how a biclustering oracle with side information can optimally recover the labels. These ideas will be the basis for our algorithms.

2.1. Bipartite block model

We work with a bipartite network that can be represented by a biadjacency matrix $A \in \{0,1\}^{n\times m}$, where for simplicity, the nodes on the two sides are indexed by the sets $[n]$ and $[m]$. We assume that there are K and L communities for the two sides respectively, and the membership of the nodes to these communities are given by two vectors $y = (y_i) \in [K]^n$ and $z = (z_j) \in [L]^m$. Thus, $y_i = k$ if node i on side 1 belongs to community $k \in [K]$. We call y_i and z_j the labels of nodes i and j, respectively. We often treat these labels as binary vectors as well, using the identification $[K] \simeq \{0,1\}^K$ via the one-hot encoding, that is $y_i = k \iff y_{ik} = 1,\ y_{ik'} = 0,\ k' \neq k$.

Given the labels y and z, and a *connectivity* matrix $P \in [0,1]^{K \times L}$ (also known as the edge probability matrix), the general bipartite stochastic block model (biSBM) assumes that: A_{ij} are Bernoulli variables, independent over $(i,j) \in [n] \times [m]$ with mean parameters,

$$\mathbb{E}[A_{ij}] = y_i^T P z_j = P_{k\ell}, \quad \text{if } y_i = k,\ z_j = \ell. \tag{1}$$

We denote this model compactly as $A \sim \mathrm{SBM}(y,z,P)$. It is helpful to consider the Poisson version of the model too, which is denoted as $A \sim \mathrm{pSBM}(y,z,P)$. This is the same model as the Bernoulli SBM, with the exception that each entry A_{ij} is drawn (independently) from a Poisson variate with the mean given in (1). These two models behave very closely when the entries of P are small enough. Throughout, we treat y, z and P as unknown deterministic parameters. The goal of network biclustering is to recover these three parameters given an instance of A.

In fact, as we will see, the parameters P themselves are not that important. What matters is the set of (Poisson) *mean parameters* which are derived from P and the sizes of the communities. In order to define these parameters, let $n(z) = (n_1(z), \dots, n_L(z)) \in \mathbb{N}^L$, be the number of nodes in each of the communities of side 2. That is, $n_\ell(z) = \sum_{j=1}^m 1\{z_j = \ell\} = \sum_{j=1}^m z_{j\ell}$. We also let $\pi_\ell(z) = n_\ell(z)/m$ be the proportion of nodes in the ℓth community of side 2. Similar notations, namely $n(y) \in \mathbb{N}^K$ and $\pi(y) \in [0,1]^K$, denote the community sizes and proportions of side 1. The *row mean parameters* are defined as

$$\Lambda = (\lambda_{k\ell}) = (P_{k\ell}\, n_\ell(z)) = P\, \mathrm{diag}(n(z)) \in \mathbb{R}^{K \times L} \tag{2}$$

where $\mathrm{diag}(v)$ for a vector $v = (v_k)$ is a diagonal matrix with diagonal entries v_k. The column mean parameters can be defined similarly,

$$\Gamma^T = \big(n_k(y) P_{k\ell}\big) = \mathrm{diag}(n(y)) P \in \mathbb{R}^{K \times L}. \tag{3}$$

Note the transpose in the above definition, i.e., $\Gamma \in \mathbb{R}^{L \times K}$. This convention allows us to define information measures based on rows of matrices Λ and Γ in a similar fashion, as will be discussed in Section 3. Although the rates we derive are controlled by the Poison parameters defined above, we always assume that the true distribution is the Bernoulli SBM and any Poisson approximation will be carefully derived.

2.2. Biclustering oracle with side information

The key idea behind the algorithms, as well as our consistency arguments is the following observation: Assume that we have prior knowledge of P and the column labels z, but not the row labels y. For each row, we can sum the columns of A according to their column memberships, i.e., we can perform the (ideal) *block compression* $b_{i\ell}^* := \sum_j A_{ij} z_{j\ell}$. The vector $b_{i*}^* = (b_{i1}^*, \dots, b_{iL}^*)$ contains the same information for recovering the community of i, as the original matrix A—i.e., it is a sufficient statistic. Assume that we are under the pSBM model (i.e., the Poisson SBM). Then, b_{i*}^* has the distribution of a vector of independent Poisson variables. More precisely,

$$b_{i*}^* \sim \mathbb{Q}_k := \prod_{\ell=1}^{L} \mathrm{Poi}(\lambda_{k\ell}), \quad \text{if}, \quad y_i = k, \tag{4}$$

where $\lambda_{k\ell}$ are the row mean parameters defined in (2). Note that the distributions $\mathbb{Q}_k$, $k = 1, \dots, K$ are known under our simplifying assumptions. The problem of determining the row labels, thus, reduces to deciding which of these K known distributions generated b^*_{i*}. Whether node i belongs to a particular community k can be decided optimally by performing a likelihood ratio (LR) test of $\mathbb{Q}_k$ against each of $\mathbb{Q}_r$, $r \neq k$.

The above LR test is at the heart of the algorithms discussed in Sections 4 and 5. The difficulty of the biclustering problem (relative to a simple mixture modeling) is that in practice, we do not know in advance either y or Λ—hence neither the exact test statistics (b^*_{i*}) nor the distributions $\{\mathbb{Q}_k\}$ are known. We thus proceed by a natural iterative procedure: Based on the initial estimates of y and z, we obtain estimates of (b^*_{i*}) and $\{\mathbb{Q}_k\}$, perform the approximate LR test to obtain better estimates of z, and then repeat the procedure over the columns to obtain better estimates of y. These new label estimates lead to better estimates of (b^*_{i*}) and $\{\mathbb{Q}_k\}$, and we can repeat the process.

We refer to the algorithm that has access to the true column labels z and parameters Λ, and performs the optimal LR tests, as the *oracle classifier*. The misclassification rate of this oracle gives a lower bound on the misclassification rate of any biclustering algorithm. The performance of the oracle, in turn, is controlled by the error exponent of the simple hypothesis testing problems $\mathbb{Q}_k$ versus $\mathbb{Q}_r, r \neq k$, as detailed in Proposition 1. This line of reasoning reveals the origin of the information quantities I_{kr} and $I^{\text{col}}_{\ell r}$—defined in (8) and (9)—that control the optimal rate of the biclustering problem. Note that the bipartite setup has the advantage of disentangling the row and column labels, so that a non-trivial oracle exists. It does not make much sense to assume known column labels in the unipartite SBM, since by symmetry we then know the row labels too, hence, nothing more is left to estimate. On the other hand, due to the close relation between the bipartite and unipartite problems, the above argument also sheds light on why the error exponent of a hypothesis test is the key factor controlling optimal misclassification rates of community detection in unipartite SBMs.

2.3. Notation on misclassification rates

Let Π_n be the set of permutations on $[n]$. The (average) misclassification rate between two sets of (column) labels $\widehat{y}$ and y is given by

$$\operatorname{Mis}(\widehat{y}, y) := \min_{\sigma \in \Pi_n} \frac{1}{n} \sum_{i=1}^{n} 1\{\sigma(\widehat{y}_i) \neq y_i\}. \tag{5}$$

Letting σ^* be a minimizer in (5), the misclassification rate over cluster k is

$$\operatorname{Mis}_k(\widehat{y}, y) := \frac{1}{n_k(y)} \sum_{i: y_i = k} 1\{\sigma^*(\widehat{y}_i) \neq y_i\} = \frac{|i : \sigma^*(\widehat{y}_i) \neq k,\, y_i = k|}{n_k(y)}, \tag{6}$$

using the cardinality notation to be discussed shortly. Note that (6) is not symmetric in its arguments. We will also use the notation $\sigma^*(\widehat{y} \to y)$ to denote an optimal permutation in (5). When $\operatorname{Mis}(\widehat{y}, y)$ is sufficiently small, this optimal permutation will be unique. It is also useful to define the *direct misclassification rate* between $\widehat{y}$ and y, denoted as $\operatorname{dMis}(\widehat{y}, y)$, which is obtained by setting the permutation in (5) to the identity. In other words, $\operatorname{dMis}(\widehat{y}, y)$ is the normalized Hamming distance between $\widehat{y}$ and y. With $\sigma^* = \sigma^*(\widehat{y} \to y)$, we have $\operatorname{Mis}(\widehat{y}, y) = \operatorname{dMis}(\sigma^*(\widehat{y}), y)$.

We note that

$$\mathrm{Mis}(\widehat{y},y) \;=\; \sum_{k\in[K]} \pi_k(y)\;\mathrm{Mis}_k(\widehat{y},y) \;\le\; \max_{k\in[K]} \mathrm{Mis}_k(\widehat{y},y), \tag{7}$$

as well as $\max_{k\in K} \mathrm{Mis}_k(\widehat{y},y) \le \mathrm{Mis}(\widehat{y},y)/\min_{k'} \pi_{k'}(y)$. We can similarly define the misclassification rate of an estimate $\widehat{z}$ relative to z. Our goal is to derive efficient algorithms for obtaining $\widehat{y}$ and $\widehat{z}$ that have minimal misclassification rates asymptotically, as the number of nodes grow.

Other notation. We write w.h.p. as an abbreviation for "with high probability", meaning that the event holds with probability $1-o(1)$. To avoid ambiguity, we assume that all parameters, including m, are functions of n. All limits and little o notations are under $n\to\infty$. For example, $f(n)=o(g(n))$ denotes $\lim_{n\to\infty} f(n)/g(n)=0$. We write $\mathbb{Z}_Q=\mathbb{Z}/Q\mathbb{Z}$ to denote a cyclic group of order Q. Our convention regarding solutions of optimization problems, whenever more than one exists, is to choose one uniformly at random. We use the shorthand notation $|i: y_i=k| := |\{i: y_i=k\}|$ for cardinality of sets, where $i\in[n]$ is implicit, assuming that y is a vector of length n. For example, if $\widehat{y},y\in[K]^n$, we have the identity $|i:\;\widehat{y}_i\neq y_i| = \sum_{k\in[K]} |i:\; y_i=k,\,\widehat{y}_i\neq k|$. It is worth noting that we use *community* and *cluster* interchangeably in this paper, although some authors prefer to use community for the assortative clusters, and use "cluster" to refer to any general group of nodes. We will not follow this convention and no assortativity will be implicitly assumed.

3. Main results

Let us start with some assumptions on the mean parameters. Recall the row and column mean parameter matrices Λ and Γ defined in (2) and (3). Let $\Lambda_{\min}$ and $\|\Lambda\|_\infty$ be the minimum and maximum value of the entries of Λ, respectively, and similarly for Γ. We assume

$$\frac{\|\Lambda\|_\infty}{\Lambda_{\min}} \vee \frac{\|\Gamma\|_\infty}{\Gamma_{\min}} \;\le\; \omega, \tag{A1}$$

for some $\omega>0$. That is, ω measures the deviation of the entries of the mean matrices from uniform. We assume that the sizes of the clusters are bounded as

$$\frac{1}{\beta K} \le \pi_k(y) \le \frac{\beta}{K} \quad\text{and}\quad \frac{1}{\beta L} \le \pi_\ell(z) \le \frac{\beta}{L} \tag{A2}$$

for all $k\in[K]$ and $\ell\in[L]$. We will assume (A1) and (A2) throughout the paper. The following key quantity controls the misclassification rate:

$$I_{kr} := I_{kr}(\Lambda) := \sup_{s\,\in\,(0,1)} \sum_{\ell=1}^{L} (1-s)\lambda_{k\ell} + s\lambda_{r\ell} - \lambda_{k\ell}^{1-s}\lambda_{r\ell}^{s}, \tag{8}$$

for $k,r\in[K]$. We can think of $I(\Lambda) := (I_{kr}(\Lambda)) \in \mathbb{R}_+^{K\times K}$, as an operator acting on pairs of rows of a matrix $\Lambda\in\mathbb{R}_+^{K\times L}$, say λ_{k*} and λ_{r*}, producing a $K\times K$ pairwise information matrix. We often refer to the function of s being maximized in (8) as $s\mapsto I_s$, with some abuse of notation, dropping the dependence on k and r and assuming that they are fixed. This function is strictly concave over $\mathbb{R}$ whenever $\lambda_{k*}\neq\lambda_{r*}$, and we have $I_0=I_1=0$.

Recalling the product Poisson distributions $\{\mathbb{Q}_k\}$, I_{kr} given in (8) is the Chernoff exponent in testing the two hypotheses $\mathbb{Q}_k$ and $\mathbb{Q}_r$ (Chernoff, 1952). The difference with the classical setting in which the Chernoff exponent appears is the regime we work in, where we are effectively testing based on a sample of size of 1 and instead, let $I_{kr} \to \infty$. The column information matrix is defined similarly

$$I_{\ell\ell'}^{\text{col}} := I_{\ell\ell'}(\Gamma) = \sup_{s\in(0,1)} \sum_{k=1}^{K} (1-s)\Gamma_{\ell k} + s\Gamma_{\ell' k} - \Gamma_{\ell k}^{1-s}\Gamma_{\ell' k}^{s}, \tag{9}$$

for all $\ell, \ell' \in [L]$. We let $I_{\min} := \min_{k\neq r} I_{kr}$ and $I_{\min}^{\text{col}} := \min_{\ell\neq\ell'} I_{\ell\ell'}^{\text{col}}$. Another set of key quantities in our analysis are:

$$\varepsilon_{kr} := \max_{\ell\in[L]} \Big(\frac{\lambda_{k\ell}}{\lambda_{r\ell}} \vee \frac{\lambda_{r\ell}}{\lambda_{k\ell}}\Big) - 1, \quad \varepsilon_k := \min_{r\in[K]} \varepsilon_{kr}, \quad \text{and} \quad \varepsilon := \min_{k\in[K]} \varepsilon_k. \tag{10}$$

The relation with hypothesis testing is formalized in the following proposition:

Proposition 1. *Consider the likelihood ratio (LR) testing of the null hypothesis $\mathbb{Q}_k$ against $\mathbb{Q}_r$, based on a sample of size 1. Let $\Lambda = [\lambda_{k*}; \lambda_{r*}] \in \mathbb{R}_+^{2\times L}$. Assume that as $\Lambda_{\min} \to \infty$, (a) $\liminf \varepsilon_{kr} > 0$, and (b) $\omega = O(1)$. Then, there exist constants C and C' such that*

$$\mathbb{P}(\textit{Type I error}) + \mathbb{P}(\textit{Type II error}) \begin{cases} \leq & C\exp\big(-I_{kr} - \frac{1}{2}\log\Lambda_{\min}\big), \\ \geq & \exp\big(-I_{kr} - \frac{L}{2}(\log\Lambda_{\min} + C')\big). \end{cases} \tag{11}$$

See Corollary 6 and Appendix G.6 for the proof. Any hypothesis testing procedure can be turned into a classifier, and a bound on the error of the hypothesis test (for a sample of size 1) translates into a bound on the misclassification rate for the associated classifier. This might not be immediately obvious, and we provide a formal statement in Lemma 3. Proposition 1 thus provides a precise bound on the misclassification rate of the *LR classifier* for deciding between $\mathbb{Q}_k$ and $\mathbb{Q}_r$. The dependence on L in the lower bound is later removed in Zhou and Li (2018), subsequent to our current work.

The significance of the Chernoff exponent of the hypothesis test in controlling the rates is thus natural, given the full information about $\{\mathbb{Q}_k\}$ and the test statistics. What is somewhat surprising is that almost the same bound holds when no such information is available a priori. Our main result below is a formalization of this claim. In our assumptions, we include a parameter $Q \in \mathbb{N}$ that controls the number of sub-blocks when partitioning, the details of which are discussed in Section 5. Under the following two assumptions:

$$(Q^2\log Q)\beta^2\omega^3 KL(K\vee L)\log(K\vee L)(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty)^2 = O(n\wedge m), \text{ and} \tag{A3}$$

$$(Q\log Q)^2\beta^3\omega^2(K\vee L)^3(\alpha\vee\alpha^{-1})(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty) = o\big((I_{\min}\wedge I_{\min}^{\text{col}})^2\big), \tag{A4}$$

where $\alpha := m/n$, there is an algorithm that achieves almost the same rate as the oracle:

Theorem 1 (Main result)**.** *Consider a bipartite SBM (Section 2.1) satisfying* (A1)–(A4). *Then, as $I_{min} \wedge I_{min}^{col} \to \infty$ and $\Lambda_{\min} \to \infty$, the row labels $\widehat{y}$ outputted by Algorithm 3 in Section 5 satisfy for some $\zeta = o(1)$,*

$$\text{Mis}_k(\widehat{y}, y) = O\Big(\omega\sum_{r\neq k}\Big(1 + \frac{1}{\varepsilon_{kr}}\Big)\exp\Big(-I_{kr} - \Big(\frac{1}{2} - \zeta\Big)\log\Lambda_{\min}\Big)\Big) \tag{12}$$

for every $k \in [K]$, with high probability. Similar bounds holds for $\widehat{z}$ w.r.t. z.

By modifying the sequence $\zeta = o(1)$, one can replace the big O with the small o in (12) to obtain an equivalent statement (e.g., factor out a term $e^{-\sqrt{\log \Lambda_{\min}}} = o(1)$ and change ζ to $\zeta' := \zeta + (\log \Lambda_{\min})^{-1/2} = o(1)$). Let us discuss the assumptions of Theorem 1. The only real assumptions are (A3) and (A4). The other two, namely (A1) and (A2), are more or less the definitions of ω and β. The main constraints on ω and β are encoded in (A3) and (A4) in tandem with the other parameters.

Remark 1. In the first reading, one can take $\beta, \omega, Q = O(1)$, $n \asymp m$ and $\|\Lambda\|_\infty \asymp \|\Gamma\|_\infty$. In this setting, (A3) is a very mild sparsity condition, implying that the degrees should not grow faster than $\sqrt{n}$. Condition (A4) guarantees that the information quantities grow fast enough so that the clusters are distinguishable. We only need (A4) for Algorithm 3 which uses a spectral initialization. In Section 5.2.1, we present Theorem 3 for the likelihood-updating portion of the algorithm, assuming that a good initialization is provided irrespective of the algorithm used. Theorem 3 only requires a weakened version of assumption (A4); see (A4′) in Section 5.2.1.

Depending on the behavior of ε_{kr}, the rate obtained in Theorem 1 can exhibit different regimes which are summarized in Corollary 1 below. Consider the additional assumption:

$$\max_{k,r\,\in\,[K]} \omega\Big(1 + \frac{1}{\varepsilon_{kr}}\Big) = O(1). \tag{A5}$$

Corollary 1. *Under the same assumptions as Theorem 1, w.h.p., for all $k \in [K]$,*

$$\mathrm{Mis}_k\,(\widehat{y},\, y) = o\Big(\sum_{r\neq k} \exp(-I_{kr})\Big). \tag{13}$$

If in addition we assume (A5)*, then for some $\zeta = o(1)$, w.h.p., for all $k \in [K]$,*

$$\mathrm{Mis}_k\,(\widehat{y},\, y) = O\Big(\sum_{r\neq k} \exp\Big(-I_{kr} - \Big(\frac{1}{2} - \zeta\Big)\log \Lambda_{\min}\Big)\Big). \tag{14}$$

Remark 2. Consider the oracle version of the biclustering problem where the connectivity matrix P and the true column labels z are given. Then, the optimal row clustering reduces to the likelihood ratio tests in Proposition 1. That is, given the row sums within blocks as sufficient statistics, we compare the likelihoods at two different mean parameters. By Proposition 1, the optimal misclassification rate for the oracle problem is

$$\mathbb{E}\big[\,\mathrm{Mis}_k\,(\widehat{y},\, y)\big] = O\Big(\sum_{r\neq k} \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min}\Big)\Big), \tag{15}$$

where the sum over r is due to the need to compare against all other clusters. The gap between $1/2$ and $1/2 - \zeta$ is not avoidable when stating high probability results, due to the Markov inequality; see Lemma 3 for the details. This error rate coincides with (14), which merely loses a constant due to the unknown mean parameters and column labels. The rate is sharp up to a factor of $\exp(-\frac{1}{2}(L-1)\log \Lambda_{\min})$ according to the lower bound in Proposition 1.

The argument in Remark 2 can be formalized as the following minimax lower bound:

Theorem 2 (Minimax lower bound)**.** *Consider the parameter space*

$$\begin{aligned}\mathcal{S} := \mathcal{S}(n,m,K,L,I^*) := \big\{(y,z,P) \mid\ & y \in \{0,1\}^{n\times K}, z \in \{0,1\}^{m\times L},\ P \in [0,1]^{K\times L},\\ & \Lambda \textit{ is defined based on } z \textit{ and } P \textit{ according to } (2),\\ & I_{min}(\Lambda) \ge I^*,\ \varepsilon \ge \varepsilon^* \textit{ where } \varepsilon \textit{ is defined in } (10),\\ & y, z \textit{ and } \Lambda \textit{ satisfy } (\mathrm{A1}) \textit{ to } (\mathrm{A4}).\big\},\end{aligned}$$

and assume that there exists $(y,z,P) \in \mathcal{S}$ such that $I_{min}(\Lambda) = I^*$. Further assume that $\beta, \omega > 1$, $\varepsilon^* > 0$ are constants and $K \le \exp(c\,L)$ for some constant $c > 0$. Then, for large n, the minimax risk over $\mathcal{S}$ satisfies

$$\inf_{\hat{y}} \sup_{(y,z,P)\in\mathcal{S}} \mathbb{E}[\,\mathrm{Mis}(\hat{y},y)\,] \;\ge\; \exp\big(-I^* - L(\log I^* + C)\big). \tag{16}$$

Theorem 2 is a non-asymptotic result, i.e., we fix n (and hence m). In this case, assumption (A4) in the definition of the parameter space $\mathcal{S}$ should be interpreted by fixing a vanishing sequence in advance. Note that in defining $\mathcal{S}$, only Λ and not Γ, is required to satisfy (A3)–(A4).

In order to better understand the rates in Corollary 1, let us consider some examples that clarify our results relative to the previous literature.

Example 1. Consider a simple planted partition model where

$$n = m, \quad K = L, \quad P_{kk} = \frac{a}{n}, \quad P_{k\ell} = \frac{b}{n},\ \forall k \ne \ell.$$

Then, $\lambda_{kk} \in [\frac{a}{\beta K}, \frac{\beta a}{K}]$ and $\lambda_{k\ell} \in [\frac{b}{\beta K}, \frac{\beta b}{K}]$ when $k \ne \ell$. Applying (8) with $s = 1/2$,

$$I_{kr} \ge \frac{1}{2}\sum_{\ell}(\sqrt{\lambda_{k\ell}} - \sqrt{\lambda_{r\ell}})^2 \ge \frac{(\sqrt{a}-\sqrt{b})^2}{\beta K}.$$

Assume that (A3) and (A4) hold, that is (using $\|\Lambda\|_\infty \le \beta a/K$),

$$\beta^4\omega^3(K\log K)a^2 = O(n) \quad \text{and} \quad \beta^6\omega^2K^4a = o\Big((\sqrt{a}-\sqrt{b})^4\Big).$$

Further assume that $\beta\omega^2K^3 = o(a \wedge b)$. Then w.h.p., we have

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = o\Big(\exp\Big(-\frac{(\sqrt{a}-\sqrt{b})^2}{\beta K}\Big)\Big). \tag{17}$$

For the details of (17), see Appendix D.3. In particular, if

$$\liminf_{n\to\infty} \frac{(\sqrt{a}-\sqrt{b})^2}{\beta K \log n} \;\ge\; 1, \tag{18}$$

we obtain $\mathrm{Mis}_k\left(\widehat{y},\, y\right) = o(1/n)$ w.h.p., that is, we have exact recovery of the labels by Algorithm 3. (Whenever misclassification rate drops below $1/n$, it should be exactly zero.) This result holds without any assumption of assortativity, i.e., it holds whether $a > b$ or $b > a$.

Example 2. Suppose that $P := \widetilde{P}(\log n)/n$ where $\widetilde{P}$ is a symmetric constant matrix, $n = m$, $K = L$, and $y = z$. Then, K, ω and ε_{kr} are constants, and

$$\lambda_{k\ell} = \widetilde{\lambda}_{k\ell} \log n, \quad \text{where } \widetilde{\lambda}_{k\ell} := \widetilde{P}_{k\ell}\pi_k(y), \quad \text{and } I_{kr} = \tilde{I}_{kr} \log n$$

where $\tilde{I}_{kr}$ is defined based on $\widetilde{\lambda}_{k\ell}$ and $\widetilde{\lambda}_{r\ell}$ as in (8). Assuming in addition that $\pi(y)$ is constant, both $\widetilde{\lambda}_{kr}$ and $\tilde{I}_{kr}$ are constants. In this regime, our key assumptions (A3) and (A4) are satisfied. By Corollary 1, w.h.p., we have

$$\text{Mis}_k\left(\widehat{y},\, y\right) = o\Big(\exp\Big(-\min_{r\neq k} \tilde{I}_{kr} \log n\Big)\Big) = o\Big(n^{-\min_{r\neq k} \tilde{I}_{kr}}\Big). \tag{19}$$

As a consequence if $\min_{k\neq r} \tilde{I}_{kr} \geq 1$, then $\text{Mis}_k\left(\widehat{y},\, y\right) = o(1/n)$ w.h.p., that is we have exact recovery by Algorithm 3.

In addition to the above more or less familiar setups (cf. Section 3.1), our results determine the optimal rate for a much wider range of parameter settings. As an example, consider the following setting of very slow degree growth:

Example 3. Consider the setup of Example 2 but with $\log n$ replaced with $\log\log n$ in the definition of P. In this case, the expected average-degree grows very slowly as $\log\log n$, and it is known that exact recovery is not possible in this regime. However, our results establish the following minimax optimal rate:

$$\text{Mis}(\widehat{y}, y) \asymp \frac{1}{(\log n)^{\alpha}} \frac{1}{\log\log n}$$

where $\alpha = \min_{k\neq r} \tilde{I}_{kr}$ and this rate is achieved by Algorithm 3.

3.1. Comparison with existing results

Let us now compare with the work of Gao et al. (2017) and Abbe and Sandon (2015) whose results are closest to ours. Both papers consider the symmetric (unipartite) SBM, but the results can be argued to hold in the bipartite setting as well. The setup of Example 1 is more or less what is considered in Gao et al. (2017). They have shown that, when $a > b$, there is an algorithm with misclassification error bounded by

$$\exp\Big(-\frac{(1-o(1))(\sqrt{a}-\sqrt{b})^2}{\beta K}\Big). \tag{20}$$

We have sharpened this rate to (17) under assumption (A3) (i.e., assuming the average degree grows slower than $O(\sqrt{n})$). Bound (20) implies that when

$$\liminf_{n\to\infty} \frac{(\sqrt{a}-\sqrt{b})^2}{\beta K \log n} > 1,$$

one has exact recovery. Our bound on the other hand, imposes the relaxed condition (18).

We note that the results in Gao et al. (2017) are derived for a more general class of (assortative) models than that of Example 1, namely, the class with connectivity matrix satisfying $P_{kk} \geq a/n$ and $P_{k\ell} \leq b/n$ for $k \neq \ell$. The rate obtained in Gao et al. (2017) uniformly over this class is dominated by that of the hardest within this class which is the model of Example 1. For other members of this class, neither their rate (20), nor the one we gave in (17) is optimal. The optimal rate in those cases is given by the general form of Theorem 1 and is controlled by the general form of I_{kr} in (8). In other words, Algorithm 3 that we present is *rate adaptive* over the class considered in Gao et al. (2017), achieving the optimal rate simultaneously for each member of the class.

A key in our approach is to apply the likelihood-type algorithm twice, in contrast to the single application in Gao et al. (2017). After the second stage we obtain much better estimates of the labels and parameters relative to the initial values, allowing us to establish the sharper form of the bound. Another key difference is the result in Lemma 2(b) which provides a better error rate than the classical Chernoff bound, using a very innovative technique introduced in Abbe and Sandon (2015). Moreover, we keep track of the balance parameter β in (A2), allowing it to go to infinity slowly. Last but not least, assortativity is a key assumption in Gao et al. (2017), while our result does not rely on it. Besides consistency, our provable algorithm is computationally more efficient. To obtain initial labels, we apply the spectral clustering on very few subgraphs (8 to be exact). However, the provable version of the algorithm in Gao et al. (2017) applies the spectral clustering for each single node on the rest of the graph excluding that node. If the cost of running the spectral clustering on a network of n nodes is C_n, then our approach costs $\approx 8C_{n/8}$ while that of Gao et al. (2017) costs roughly nC_{n-1}. Our algorithm thus has a significant advantage in computational complexity when $n \to \infty$. To be fair, the algorithm in Gao et al. (2017) was for the symmetric SBM, which has the extra complication of dependency in A due to symmetry. Our comparison here is mostly with Corollary 3.1 in Gao et al. (2017). In addition, Gao et al. (2017) have a result (their Theorem 5) for when ω grows arbitrarily fast which is not covered by our result. See Appendix A.2 for comments on the symmetric case and dependence on ω.

The problem of exact recovery for a general SBM has been considered in Abbe and Sandon (2015), again for the case of a symmetric SBM, though the results are applicable to the bipartite setting (with $y = z$). The model and scaling considered in Abbe and Sandon (2015) is the same as that of Example 2, and they show that exact recovery of all labels is possible if (and only if) $\min_{k,r:k\neq r} \widetilde{I}_{kr} \geq 1$ which is the same result we obtain in Example 2 for Algorithm 3. Thus, our result contains that of Abbe and Sandon (2015) as a special case, namely in the $\log n$ regime of degree growth, with other parameters (such as K and the normalized connectivity matrix) kept constant. The results and algorithms of Abbe and Sandon (2015) do not apply to the general model in our paper. First, they only consider the regime $P \sim \log n/n$, i.e., the degree grows as fast as $\log n$, while we allow the degree to grow in the range from "arbitrarily slowly" up to "as fast as $O(\sqrt{n})$". Second, as discussed in Appendix A.1, their edge splitting idea cannot be used to derive the results in this paper, and we introduce the block partitioning to supply the independent copies necessary for the analysis.

Finally, we note that Example 3 with a general nonassortative matrix $\widetilde{P}$ has no counterpart in the literature. Existing results are not capable of providing any guarantees for such setups.

4. Pseudo-likelihood approach

In this section, after introducing the local and global mean parameters which will be used throughout the paper, we present our general pseudo-likelihood approach to biclustering.

4.1. Local and global mean parameters

Let us define the following operator that takes an adjacency matrix A, and row and column labels $\tilde{y}$ and $\tilde{z}$, and outputs the corresponding (unbiased) estimate of its mean parameters:

$$[\mathscr{L}(A, \tilde{y}, \tilde{z})]_{k\ell} = \frac{1}{n_k(\tilde{y})} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} 1\{\tilde{y}_i = k, \tilde{z}_j = \ell\}, \quad k \in [K],\ \ell \in [L]. \tag{21}$$

Note that $\mathscr{L}(A, \tilde{y}, \tilde{z})$ is a $K \times L$ matrix with nonnegative entries. In general, we let

$$\hat{\Lambda} = (\hat{\lambda}_{k\ell}) := \mathscr{L}(A, \tilde{y}, \tilde{z}), \tag{22}$$

$$\Lambda(\tilde{y}, \tilde{z}) = (\lambda_{k\ell}(\tilde{y}, \tilde{z})) := \mathscr{L}(\mathbb{E}[A], \tilde{y}, \tilde{z}), \tag{23}$$

for any row and column labels $\tilde{y}$ and $\tilde{z}$. Here, $\hat{\Lambda}$ is the estimate of the true row mean matrix. Its expectation is $\mathbb{E}[\hat{\Lambda}] = \Lambda(\tilde{y}, \tilde{z})$ due to the linearity of $\mathscr{L}$. We call $\Lambda(\tilde{y}, \tilde{z})$, the *(global) row mean parameters* associated with labels $\tilde{y}$ and $\tilde{z}$. (We do not explicitly show the dependence of $\hat{\Lambda}$ on the labels, in contrast to the mean parameters.) We have the following key identity

$$\Lambda(\tilde{y}, \tilde{z}) \mid_{\tilde{y}=y,\ \tilde{z}=z} = \ \Lambda \tag{24}$$

where Λ is the *true* (global) row mean parameter matrix defined in (2). In words, (24) states that the global row mean parameters associated with the true labels y and z, are the true such parameters. We will also use parameters such as $\Lambda(y, \tilde{z})$ which are obtained based on the true row labels y and generic column labels $\tilde{z}$.

We also need local versions of all these definitions which are obtained based on submatrices of A. More precisely, let $A^{(q',q)}$ be a submatrix of A, and let $y^{(q')}$ and $z^{(q)}$ be the corresponding subvectors of z and y (i.e., corresponding to the same row and column index sets used to extract the submatrix). Here, q and q' range over $[Q] = \{1, \dots, Q\}$, creating a partition of A into Q^2 subblocks. We call

$$\Lambda^{(q',q)}(\tilde{y}, \tilde{z}) := (\lambda_{k\ell}^{(q',q)}(\tilde{y}, \tilde{z})) := \mathscr{L}(\mathbb{E}[A^{(q',q)}],\, \tilde{y}^{(q')},\, \tilde{z}^{(q)}), \tag{25}$$

the *local row mean parameters* associated with submatrix $A^{(q',q)}$ and sublabels $y^{(q')}$ and $z^{(q)}$. The corresponding estimates are defined similarly, by replacing $\mathbb{E}[A^{(q',q)}]$ with $A^{(q',q)}$. We will mostly work with submatrices obtained from a partition $A^{(q',q)}$, $q', q \in [Q]$ of A into (nearly) equal-sized blocks—the details of which are described in Section 5. In such cases,

$$\Lambda^{(q',q)}(\tilde{y}, \tilde{z}) \approx \frac{1}{Q} \Lambda(\tilde{y}, \tilde{z}), \quad \forall q', q \in [Q]$$

assuming that each subblock in the partition has nearly similar cluster proportions: $n(z^{(q)}) \approx n(z)$. This is the case, for example, for a random partition as we show in Appendix B.2.

Of special interest is when we replace both $\tilde{y}$ and $\tilde{z}$ with true labels y and z. In such cases, $\Lambda^{(q',q)}$ does not depend on q'. More precisely, we have for any $q \in [Q]$,

$$\lambda_{k\ell}^{(q',q)}(y, z) = P_{k\ell}\, n_\ell(z^{(q)}), \quad \forall q' \in [Q], \tag{26}$$

where $n_\ell(z^{(q)})$ is the number of labels in class ℓ in $z^{(q)}$, consistent with our notation for the full label vectors. We often write $\Lambda^{(q)} = (\lambda_{k\ell}^{(q)})$ as a shorthand for $\Lambda^{(q',q)}(y, z)$ which is justified by the above discussion. These will be called the *true* local row mean parameters (associated with column q subblock in the partition).

4.2. General pseudo-likelihood algorithm

Let us now describe our main algorithm based on the pseudo-likelihood (PL) idea, which is a generalization of the approach in Amini et al. (2013) to the bipartite setup. The pseudo-likelihood algorithm (PLA) is effectively an EM algorithm applied to the approximate mixture of Poissons obtained from the block compression of the adjacency matrix. It relies on some initial estimates of the row and column labels to perform the first block compressions (for both rows and columns). The initialization is often done by spectral clustering and will be discussed in Section 5.2.2. Subsequent block compressions are performed based on the label updates at previous steps of PLA.

Let us assume that we have obtained labels $\tilde{y}$ and $\tilde{z}$ as estimates of the true labels y and z. We focus on the steps of PLA for recovering the row labels. Let us define an operator $\mathcal{B}(A;\tilde{z})$ that takes approximate columns labels and produces the corresponding *column compression* of A:

$$\mathcal{B}(A;\tilde{z}) := \boldsymbol{b}(\tilde{z}) := \big(b_{i\ell}(\tilde{z})\big) \in \mathbb{Z}_+^{n\times L}, \quad b_{i\ell}(\tilde{z}) := \sum_{j=1}^m A_{ij} 1\{\tilde{z}_j = \ell\}. \tag{27}$$

The distribution of $b_{i\ell}(\tilde{z})$ is determined by the row class of i. It is not hard to see that

$$\mathbb{E}[b_{i\ell}(\tilde{z})] = \lambda_{k\ell}(y,\tilde{z}) = \lambda_{k\ell}(\tilde{y},\tilde{z})|_{\tilde{y}=y}, \quad \text{if } y_i = k, \tag{28}$$

where $\lambda_{k\ell}(\tilde{y},\tilde{z})$ are the (global) row mean parameters defined in (23).

Now consider an operator $\mathcal{L}(\boldsymbol{b};\tilde{y})$ that, given the column compression $\boldsymbol{b}$ and the initial estimate of the row labels $\tilde{y}$, produces estimates of the *(row) mean parameters* $\lambda_{k\ell}(y,\tilde{z})$:

$$\mathcal{L}(\boldsymbol{b};\tilde{y}) := \hat{\Lambda} := [\hat{\lambda}_{k\ell}] \in \mathbb{R}_+^{K\times L}, \quad \hat{\lambda}_{k\ell} := \frac{1}{n_k(\tilde{y})} \sum_{i=1}^n b_{i\ell} 1\{\tilde{y}_i = k\}. \tag{29}$$

Note that if $\tilde{y} = y$, we have $\mathbb{E}[\hat{\lambda}_{k\ell}] = \lambda_{k\ell}(y,\tilde{z})$. The definition of the estimates in (29) are consistent with those of (22) due to the following identity:

$$\mathcal{L}(\mathcal{B}(A;\tilde{z});\tilde{y}) = \mathscr{L}(A,\tilde{y},\tilde{z})$$

which holds for any row labels $\tilde{y}$ and column labels $\tilde{z}$. Let us write

$$\pi(\tilde{y}) := (\pi_k(\tilde{y})), \quad \pi_k(\tilde{y}) := \frac{1}{n}\sum_{i=1}^n 1\{\tilde{y}_i = k\} \tag{30}$$

for the estimate of (row) class priors based on $\tilde{y}$. We note that operations $\mathcal{B}$ and $\mathcal{L}$ remain valid even if $\tilde{y}$ and $\tilde{z}$ are *soft labels* with a minor modification. By a soft row label $\tilde{z}_j \in [0,1]^L$ we mean a probability vector on $[L]$: $\tilde{z}_{j\ell} \geq 0$ and $\sum_{\ell=1}^L \tilde{z}_{j\ell} = 1$, which denotes a soft assignment to each row cluster. To extend (27) to soft row labels, it is enough to replace $1\{\tilde{z}_j = \ell\}$ with $\tilde{z}_{j\ell}$. Extending (29) to soft column labels $\tilde{y}$ is done similarly.

Given any block compression $\boldsymbol{b} = (b_{i\ell})$, any estimate $\hat{\Lambda}$ of the (row) mean parameters and any estimate $\widetilde{\pi} \in [0,1]^K$ of the (row) class prior, consider the operator that outputs the *(row) class posterior* assuming that the rows of $\boldsymbol{b}_i$ approximately follow $\sum_k \widetilde{\pi}_k \prod_\ell \mathrm{Poi}(\hat{\lambda}_{k\ell})$:

$$\mathcal{F}(\boldsymbol{b},\hat{\Lambda},\widetilde{\pi}) := (\widehat{\pi}_{ik}) \in [0,1]^{n\times K}, \quad \widehat{\pi}_{ik} := \frac{\widetilde{\pi}_k \prod_{\ell=1}^L \varphi(b_{i\ell},\hat{\lambda}_{k\ell})}{\sum_{k'=1}^K \widetilde{\pi}_{k'} \prod_{\ell=1}^L \varphi(b_{i\ell},\hat{\lambda}_{k'\ell})}, \tag{31}$$

Algorithm 1 Pseudo-likelihood biclustering, meta algorithm

```
1:  Initialize row and column labels $\tilde{y}$ and $\tilde{z}$.
2:  while $\tilde{y}$ and $\tilde{z}$ have not converged do
3:      $\boldsymbol{b} \leftarrow \mathcal{B}(A; \tilde{z})$
4:      while $\hat{\Lambda}$ and $\tilde{y}$ not converged (optional) do
5:          $\hat{\Lambda} \leftarrow \mathcal{L}(\boldsymbol{b}; \tilde{y})$
6:          Option 1: $\widetilde{\pi} \leftarrow \mathbf{1}$, or option 2: $\widetilde{\pi} \leftarrow \pi(\tilde{y})$
7:          $\tilde{y} \leftarrow \mathcal{F}(\boldsymbol{b}, \hat{\Lambda}, \widetilde{\pi})$
8:          (Optional) Convert $\tilde{y}$ to hard labels.
9:      end while
10:     Repeat lines 3–9 with appropriate modifications to update $\tilde{z}$ and columns parameters
    (by changing $A$ to $A^T$ and swapping $\tilde{z}$ and $\tilde{y}$.)
11:     (Optional) Convert $\tilde{y}$ and $\tilde{z}$ to hard labels if they are not.
12: end while
```

Algorithm 2 Simplified pseudo-likelihood clustering

```
Input: Initial column labels $\tilde{z}$, and $\tilde{\Lambda}$ that estimates $\Lambda$.
Output: Estimate of row labels $\widehat{y}$.
$\boldsymbol{b} \leftarrow \mathcal{B}(A; \tilde{z})$
$\hat{\Lambda} \leftarrow \mathcal{L}(\boldsymbol{b}; \tilde{y})$
$\widehat{y} \leftarrow \mathcal{F}(\boldsymbol{b}, \tilde{\Lambda}, \mathbf{1})$
Convert $\widehat{y}$ to hard labels, by computing MAP estimates.
```

where $\varphi(x, \lambda) = \exp(x \log \lambda - \lambda)$ is the Poisson likelihood (up to constants). In practice, we only use $\pi(\tilde{y})$ or a flat prior $\mathbf{1}$ as the estimated prior $\widetilde{\pi}$ in this step; similarly, we only use a block compression that is based on the estimates of row labels, i.e., $b_{i\ell} = b_{i\ell}(\tilde{z})$ for some $\tilde{z} \in [n]^L$. Note that $\mathcal{F}$ outputs soft-labels as new estimates of y. We can convert $(\widehat{\pi}_{ik})$ to hard labels if needed.

Algorithm 1 summarizes the general blueprint of PLA which proceeds by iterating the three operators (27), (29) and (31). An optional conversion from soft to hard labels is performed by MAP assignment per row. With option 2 in step 6, the inner loop on lines 4–8 is the EM algorithm for a mixture of Poisson vectors. We can also remove the inner loop and perform iterations 5–8 only once. In total, Algorithm 1 has (at least) 6 possible versions, depending on whether we include each of the steps 8 or 11 (for the soft to hard label conversion) and whether to implement the inner loop till convergence or only for one step. We provide empirical results for two of these versions in Section 6. In practice, we recommend keeping the labels soft throughout, and only run the inner loop for a few iterations; maybe even once, if the computational cost is of significance.

4.3. Likelihood ratio classifier

A basic simplified building block of the PLA is given in Algorithm 2. This operation—which will play a key role in the development of the provable version of the algorithm in Section 5—can be equivalently described as a *likelihood ratio classifier* (LRC). Let us write the joint Poisson likelihood (up to a constant) as:

$$\Phi(x,\lambda) = \prod_{\ell=1}^{L} \varphi(x_\ell, \lambda_\ell) = \prod_{\ell=1}^{L} \exp(x_\ell \log \lambda_\ell - \lambda_\ell), \quad x \in \mathbb{R}^L,\ \lambda \in \mathbb{R}_+^L, \tag{32}$$

and the corresponding likelihood ratio as:

$$\Psi(x;\lambda \mid \lambda') = \log \frac{\Phi(x,\lambda)}{\Phi(x,\lambda')} = \sum_{\ell=1}^{L} x_\ell \log \frac{\lambda_\ell}{\lambda'_\ell} + \lambda'_\ell - \lambda_\ell, \quad x \in \mathbb{R}^L,\ \lambda, \lambda' \in \mathbb{R}_+^L. \tag{33}$$

Recalling the column compression (27), the *likelihood ratio classifier*, based on initial row labels $\tilde{z}$ and an estimate $\tilde{\Lambda}$ of the row mean parameter matrix, is

$$[\mathrm{LR}(A, \tilde{\Lambda}, \tilde{z})]_i \ \in \ \underset{r \in [K]}{\mathrm{argmax}} \ \log \Phi(b_{i*}(\tilde{z}), \tilde{\lambda}_{r*}), \quad i \in [n]. \tag{34}$$

This operation gives us a refined estimate of the row labels (i.e., y). It is not hard to see that the output of Algorithm 2 is $\widehat{y} = \mathrm{LR}(A, \tilde{\Lambda}, \tilde{z})$.

5. Provable version

When analyzing Algorithm 2, we need the initial labels to be independent of the adjacency matrix. Hence, we cannot apply the initialization method (e.g., the spectral clustering) and the likelihood ratio classifier (Algorithm 2) on the same adjacency matrix. In this section, we introduce Algorithm 3 which partitions A into sub-blocks and operates iteratively on collections of these blocks to maintain the desired independence. Algorithm 3 is the version of the pseudo-likelihood algorithm for which our main result (Theorem 1) holds.

Let us assume that n and m are divisible by $2Q = 8$. This assumption is not necessary but helps simplify the notation. Let us write

$$\widehat{y} = \mathrm{rowSC}(A), \quad \widehat{z} = \mathrm{colSC}(A)$$

to denote labels obtained by applying the spectral clustering on rows and columns of the adjacency matrix A, respectively; see Section 5.2.2 for details. We have $\mathrm{colSC}(A) = \mathrm{rowSC}(A^T)$. We also recall the LR classifier defined in (34). For matrices (or vectors) A and B, we use $[A; B]$ to denote column concatenation and $[A\ B]$ to denote row concatenation.

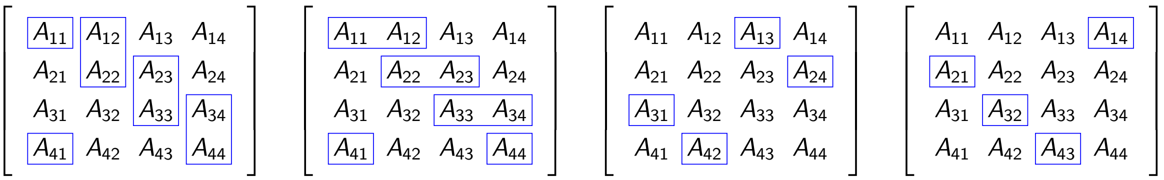
The general idea behind the partitioning scheme used in Algorithm 3, which is done by sequential sampling without replacement, is to ensure that in each step where the LR classifier is applied, the initial labels used are independent of the sub-block of the adjacency matrix under consideration. We do not require, however, that the initial labels be independent of the estimates of the mean parameters $\hat{\Lambda}$, since—as will be seen in Appendix B.1—we have uniform consistency of the LR classifier over all $\hat{\Lambda}$ close to the truth. For example, in step 7, that is, in the assignment $\tilde{y}^{(q)} \leftarrow \mathrm{LR}(A^{(q,q+2)}, \hat{\Lambda}^{(q+2)}, \tilde{z}^{(q+2)})$, the claim is that $\tilde{z}^{(q+2)}$—at that stage in the algorithm—is independent of $A^{(q,q+2)}$ but not necessarily of $\hat{\Lambda}^{(q+2)}$. This will become clear in the following discussion where we keep track of the dependence of various estimates through the algorithm. Note that in the description of Algorithm 3, we are using the computer coding convention for in-place assignments, e.g., $\tilde{z}^{(q)}$ gets updated in place and refers to different objects at different points in the algorithm.

Algorithm 3 Provable (parallelizable) version

```
1: Randomly partition the rows into 2 groups of equal size ($n/2$), so that

       $$A = [A_{\text{top}}\,;\,A_{\text{bottom}}]$$

2: Randomly partition the rows and columns of $A_{\text{bottom}}$ into 4 groups of equal size, so that
   we have 16 sub-adjacency matrix with dimension $(n/8) \times (m/4)$, i.e.

       $$A_{\text{bottom}} = \begin{bmatrix} A^{(1,1)} & A^{(1,2)} & A^{(1,3)} & A^{(1,4)} \\ A^{(2,1)} & A^{(2,2)} & A^{(2,3)} & A^{(2,4)} \\ A^{(3,1)} & A^{(3,2)} & A^{(3,3)} & A^{(3,4)} \\ A^{(4,1)} & A^{(4,2)} & A^{(4,3)} & A^{(4,4)} \end{bmatrix}.$$

   In each of the following steps, perform the stated operation for every $q \in \mathbb{Z}_4$:
3: Obtain initial row labels:              $[\tilde{y}^{(q-1)}\,;\,\tilde{y}'^{(q)}] \leftarrow \text{rowSC}\left([A^{(q-1,q)}\,;\,A^{(q,q)}]\right),\ \forall q.$
4: Obtain initial column labels:           $[\tilde{z}^{(q)}\ \tilde{z}'^{(q+1)}] \leftarrow \text{colSC}([A^{(q,q)}\ A^{(q,q+1)}]),\ \forall q.$
5: Get consistent (global) labels:         $\tilde{y} \leftarrow \textsc{Match}(\tilde{y}, \tilde{y}')$ and $\tilde{z} \leftarrow \textsc{Match}(\tilde{z}, \tilde{z}')$.
6: Update (local) row mean parameters:     $\hat{\Lambda}^{(q+2)} \leftarrow \mathscr{L}(A^{(q,q+2)}, \tilde{y}^{(q)}, \tilde{z}^{(q+2)}),\ \forall q.$
7: Update row labels:                      $\tilde{y}^{(q)} \leftarrow \text{LR}(A^{(q,q+2)}, \hat{\Lambda}^{(q+2)}, \tilde{z}^{(q+2)}),\ \forall q.$
8: Similarly update column labels $\tilde{z}$ as in steps 6 and 7.
9: Update (local) row mean parameters:     $\hat{\Lambda}^{(q+3)} \leftarrow \mathscr{L}(A^{(q,q+3)}, \tilde{y}^{(q)}, \tilde{z}^{(q+3)}),\ \forall q.$
10: Obtain (global) row mean parameters:   $\hat{\Lambda} \leftarrow \sum_q \hat{\Lambda}^{(q)}$.
11:                                        $\widehat{y}_{\text{top}} \leftarrow \text{LR}(A_{\text{top}}, \hat{\Lambda}, \tilde{z})$.
12: Swap $A_{\text{top}}$ and $A_{\text{bottom}}$, then repeat steps 2–11 to obtain $\widehat{y}_{\text{bottom}}$, except for step 5 where
    $\tilde{y}$ is matched to $\widehat{y}_{\text{top}}$ instead, i.e., $\tilde{y} \leftarrow \textsc{Match}(\tilde{y}, \widehat{y}_{\text{top}})$.
13:                                        $\widehat{y} \leftarrow [\widehat{y}_{\text{top}}\,;\,\widehat{y}_{\text{bottom}}]$.
14: Apply step 1 to 10 on $A^T$ to obtain $\widehat{z}$.
```

Figure 1 illustrates the partitions used in steps 2–9 of the algorithm. The collection of the submatrices in the partition is given a name in each case. For example, G_1^{col} consists of the four submatrices in Figure 1(a). Note that $\{G_1^{\text{col}}, G_2, G_3\}$ form a complete partition of the matrix A_{bottom} into disjoint blocks. Also, G_1^{col} and G_1^{row} involve the same elements of the matrix, i.e. they *cover* the same portion of A_{bottom}. Thus, $\{G_1^{\text{row}}, G_2, G_3\}$ is also a complete cover of A_{bottom} with disjoint blocks. Let us write G_1 for the common portion of A_{bottom} covered by G_1^{col} and G_1^{row}.

Steps 3 and 4 operate on blocks in G_1^{col} and G_1^{row} respectively, producing initial row and column labels. For example, in step 3, we apply row SC on each submatrix specified in Figure 1(a)



(a) G_1^{col} (Step 3) (b) G_1^{row} (Step 4) (c) G_2 (Steps 6, 7) (d) G_3 (Step 9)

Figure 1: The four stages of partitioning in Algorithm 3. In each case, the collection of submatrices in the partition is given a name which is used in the text. We have used to shorthand $A_{qq'} = A^{(q,q')}$ for simplicity. Block used in obtaining initial labels (a–b), in obtaining the first local parameter estimates (c), and in the first application of LR classifier (d).

and obtain the label vectors (from the leftmost submatrix to the rightmost one):

$$[\tilde{y}'^{(1)}\,;\,\tilde{y}^{(4)}],\ [\tilde{y}^{(1)}\,;\,\tilde{y}'^{(2)}],\ [\tilde{y}^{(2)}\,;\,\tilde{y}'^{(3)}], [\tilde{y}^{(3)}\,;\,\tilde{y}'^{(4)}]. \tag{35}$$

As a result of these steps, we obtain two sets of row labels $\tilde{y} = (\tilde{y}^{(q)} :\ q \in \mathbb{Z}_4)$ and $\tilde{y}' = (\tilde{y}'^{(q)} :\ q \in \mathbb{Z}_4)$, and similarly for the columns labels. Neither of $\tilde{y}$ or $\tilde{y}'$ is necessarily a consistent set of labels for the whole matrix, since the cluster labels for individual pieces $y^{(q)}$ and $\tilde{y}'^{(q)}$ need not match (e.g., cluster 1 in one piece could be labeled cluster 2 in another piece.). However, if the sub-block labels (35) are sufficiently close to the truth, we can use the overlap among them to find a global set of labels that are consistent with each block of $\tilde{y}$ and $\tilde{y}'$. This is what the Match operator in step 5 does, as will be detailed in Section 5.1. The resulting updated global row and column labels only depend on G_1 portion of A_{bottom}. Steps 6–13 go through the following phases:

First local parameter estimates (step 6): Having obtained good initial (global) row and column labels, in Step 6, we obtain estimates of the local mean parameters $\hat{\Lambda}^{(q+2)}$ for the submatrices in G_2 as in Figure 1(c). Note for example, that $\hat{\Lambda}^{(q+2)}$ computed in this step depends on blocks $A^{(q,q+2)}$ and on G_1 through $\tilde{z}^{(q+2)}$. Collectively, the estimates $\{\hat{\Lambda}^{(q+2)} :\ q \in \mathbb{Z}_4\}$ in Step 6 depend on $G_1 \cup G_2$ portion of A_{bottom}.

First LR classifier (steps 7–8): Using the estimates of the (local) row mean parameters, in Step 7, we apply the LR classifier, $\tilde{y}^{(q)} \leftarrow \text{LR}(A^{(q,q+2)}, \hat{\Lambda}^{(q+2)}, \tilde{z}^{(q+2)})$ to each of the submatrices in G_2 (in Figure 1(c)). Here, $\hat{\Lambda}^{(q+2)}$ depends on the same block $A^{(q,q+2)}$ on which we apply LR classifier, but the dependence is not problematic due to the uniform consistency of LR classifier in parameters (Lemma 2). However, we note that $\tilde{z}^{(q+2)}$ is a function of G_1 blocks of A_{bottom}, hence independent of $A^{(q,q+2)}$ which is key in our arguments. We will similarly apply the LR classifier on the columns of G_2, and obtain $\tilde{z}^{(q)}$. By the end of step 8, the updated labels $\tilde{y}$ and $\tilde{z}$ will depend on blocks in $G_1 \cup G_2$; these labels will be much more accurate (Mis $\approx \exp(-I/Q)$) than the initial labels obtained by spectral clustering.

Second parameter estimates (steps 9–10): Using the more accurate labels of step 8, we obtain the local mean parameters $\hat{\Lambda}^{(q+3)}$ in step 9 for the submatrices in G_3 (Figure 1(d)). This step is similar to step 6, but because of the much more accurate labels, the parameter estimates are much more accurate too. Since the global mean parameter is the sum of local mean parameters, i.e. $\Lambda = \sum_{q \in [Q]} \Lambda^{(q)}$, we use $\hat{\Lambda} := \sum_q \hat{\Lambda}^{(q)}$ to estimate Λ in step 10. We recall that the true local mean parameters do not depend on the block row index; see (26).

Second LR classifier (step 11): Using the more accurate estimates of (global) row mean parameters $\hat{\Lambda}$ from step 10 and the more accurate labels $\tilde{z}$ in step 8, in step 11 we apply the LR classifier $\widehat{y}_{\text{top}} \leftarrow \text{LR}(A_{\text{top}}, \hat{\Lambda}, \tilde{z})$ on A_{top}. We note that A_{top} in this step is independent of $\tilde{z}$ (as well as $\hat{\Lambda}$). This second LRC application is what brings us from very accurate labels (Mis $\approx \exp(-I/Q)$) to almost optimal (Mis $\approx \exp(-I)$); see Appendix C.

Bottom half (steps 12–13): The same process is repeated in step 12, after swapping the top and bottom halves of A, to get the bottom portion of the row labels. Matching the labels $\tilde{y}$ from the spectral clustering in step 5 with $\widehat{y}_{\text{top}}$ guarantees that $\widehat{y}_{\text{top}}$ and $\widehat{y}_{\text{bottom}}$ have consistent labels. Thus, no extra matching is required when concatenating the two in step 13.

5.1. Matching step

Let us describe the details of the matching step in Algorithm 3. Although, the idea is intuitively clear, formally describing the procedure is fairly technical. In order to understand the idea, consider the two-block labels $\tilde{y}^{(q-1,q)} := [\tilde{y}^{(q-1)}\,;\, \tilde{y}'^{(q)}]$, for $q = 2, 3$, that is,

$$\tilde{y}^{(1,2)} := [\tilde{y}^{(1)}\,;\, \tilde{y}'^{(2)}], \quad \tilde{y}^{(2,3)} := [\tilde{y}^{(2)}\,;\, \tilde{y}'^{(3)}].$$

We will detail how these two sets of labels can be fused together to generate a set of consistent labels for the three-block true label vector $y^{(1,2,3)} := [y^{(1)}; y^{(2)}; y^{(3)}]$. The two (overlapping) two-blocks of the true label vector are denoted as

$$y^{(1,2)} := [y^{(1)}\,;\, y^{(2)}], \quad y^{(2,3)} := [y^{(2)}\,;\, y^{(3)}].$$

More generally, we let $y^{(q-1,q)} = [y^{(q-1)}\,;\, y^{(q)}]$, similar to the notation for estimated blocks.

Recall our notation $\sigma^*(\cdot \;\to\; \cdot)$ for (an) optimal permutation between two sets of labels (cf. Section 2.3). Finding σ^* is a linear assignment problem, with computational complexity $O(K \vee L)^3$); see Burkard and Cela (1999). Let us define

$$\sigma_{q-1,q} := \sigma^*\big(\tilde{y}^{(q-1,q)} \to y^{(q-1,q)}\big), \quad \sigma_q := \sigma^*(\tilde{y}^{(q)} \to y^{(q)}), \quad \sigma'_q := \sigma^*(\tilde{y}'^{(q)} \to y^{(q)}). \tag{36}$$

Thus, for example we have

$$\sigma_{1,2} = \sigma^*(\tilde{y}^{(1,2)} \;\to\; y^{(1,2)}), \quad \sigma_2 = \sigma^*(\tilde{y}^{(2)} \to y^{(2)}), \quad \sigma'_3 = \sigma^*(\tilde{y}'^{(3)} \to y^{(3)}),$$

and so on, as depicted in Figure 2(a). In other words, each of these permutations is the optimal permutation from the corresponding block of the underlying estimated label to that of the truth. Let us write $\tilde{y}^{(1,2)} \approx y^{(1,2)}$ to mean that the two sets of labels are sufficiently close (to be made precise later).

The first claim is that $\tilde{y}^{(1,2)} \approx y^{(1,2)}$ implies that the underlying sub-blocks have the same optimal permutation to the truth as the original two-block label, i.e.,

$$\tilde{y}^{(1,2)} \approx y^{(1,2)} \implies \sigma_1 = \sigma'_2 = \sigma_{1,2}$$

and similarly $\tilde{y}^{(2,3)} \approx y^{(2,3)} \implies \sigma_2 = \sigma'_3 = \sigma_{2,3}$. The second claim is that each sub-block has "almost" the same misclassification error as the bigger two-block. To see this, recall the *direct misclassification rate* introduced in Section 2.3, i.e., misclassification rate without applying any permutation (or equivalently with the identity permutation). We have

$$\text{dMis}\left(\sigma_{2,3}(\tilde{y}^{(2,3)}), y^{(2,3)}\right) = \text{Mis}\left(\tilde{y}^{(2,3)}, y^{(2,3)}\right) \leq \varepsilon. \tag{37}$$

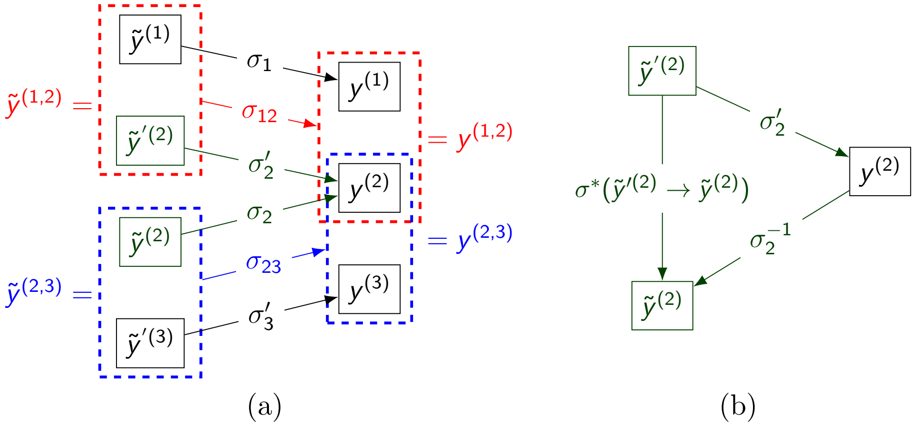


Figure 2: Pictorial depiction of the matching step. (a) Two-block and sub-block optimal permutations to the truth. When $\tilde{y}^{(1,2)} \approx y^{(1,2)}$, we have $\sigma_1 = \sigma_2' = \sigma_{1,2}$ and similarly $\tilde{y}^{(2,3)} \approx y^{(2,3)}$ implies $\sigma_2 = \sigma_3' = \sigma_{2,3}$. (b) Commutative diagram depicting how the missing permutation $\sigma_2^{-1} \circ \sigma_2'$ can be obtained by matching observed labels $\tilde{y}'^{(2)}$ and $\tilde{y}^{(2)}$. See Section 5.1 for details.

where the inequality is by assumption (ε being the rate achieved by the spectral clustering algorithm). A similar expression holds with $(2,3)$ replaced with $(1,2)$. Now (37) implies

$$\mathrm{dMis}\left(\sigma_2(\tilde{y}^{(2)}), y^{(2)}\right) = \mathrm{dMis}\left(\sigma_{2,3}(\tilde{y}^{(2)}), y^{(2)}\right) \le 2\varepsilon = \varepsilon' \tag{38}$$

where the equality uses $\sigma_2 = \sigma_{2,3}$. To see the inequality, let $n_{2,3}$, n_2 and n_3 be the lengths of $y^{(2,3)}$, $y^{(2)}$ and $y^{(3)}$. Then,

$$\mathrm{dMis}\left(\sigma_{2,3}(\tilde{y}^{(2,3)}), y^{(2,3)}\right) = \frac{n_2}{n_{2,3}} \mathrm{dMis}\left(\sigma_{2,3}(\tilde{y}^{(2)}), y^{(2)}\right) + \frac{n_3}{n_{2,3}} \mathrm{dMis}\left(\sigma_{2,3}(\tilde{y}'^{(3)}), y^{(3)}\right)$$

and the result follows since we have $n_2 = n_3 = n_{2,3}/2$ by construction. Note that dMis has the property of being easily distributed over sub-blocks as opposed to Mis. Similarly, we obtain $\mathrm{dMis}\left(\sigma_3'(\tilde{y}^{(3)}), y^{(3)}\right) \le \varepsilon'$ considering the second components of $\tilde{y}^{(2,3)}$ and $y^{(2,3)}$. Applying the same argument to indices $(1,2)$, we conclude similarly that $\mathrm{dMis}\left(\sigma_1(\tilde{y}^{(1)}), y^{(1)}\right) \le \varepsilon'$ and $\mathrm{dMis}\left(\sigma_2'(\tilde{y}'^{(2)}), y^{(2)}\right) \le \varepsilon'$.

Now consider the following three-block vector, undergoing the transformation

$$\begin{bmatrix} \sigma_1(\tilde{y}^{(1)}) \\ \sigma_2(\tilde{y}^{(2)}) \\ \sigma_3'(\tilde{y}^{(3)}) \end{bmatrix} \to \begin{bmatrix} \sigma_2^{-1} \circ \sigma_1(\tilde{y}^{(1)}) \\ \sigma_2^{-1} \circ \sigma_2(\tilde{y}^{(2)}) \\ \sigma_2^{-1} \circ \sigma_3'(\tilde{y}^{(3)}) \end{bmatrix} \overset{=}{\to} \begin{bmatrix} \sigma_2^{-1} \circ \sigma_1(\tilde{y}^{(1)}) \\ \tilde{y}^{(2)} \\ \tilde{y}^{(3)} \end{bmatrix} \overset{=}{\to} \begin{bmatrix} \sigma_2^{-1} \circ \sigma_2'(\tilde{y}^{(1)}) \\ \tilde{y}^{(2)} \\ \tilde{y}^{(3)} \end{bmatrix}.$$

The leftmost vector has dMis of at most ε' relative to $y^{(1,2,3)}$ by the previous arguments, and since Mis $\le$ dMis, we have the same bound on Mis rate for the leftmost vector. The first transformation keeps the same Mis rate since we are applying a single permutation σ_2^{-1} to all elements. The second transformation is in fact an equality, using $\sigma_3' = \sigma_2$ established earlier. The third transformation/equality follows similarly by $\sigma_1 = \sigma_2'$. Thus, if we can recover $\sigma_2^{-1} \circ \sigma_2'$ from the data, we can construct a consistent three-block label having Mis $\le \varepsilon'$.

The third and final claim is that this is possible, and in fact we have

$$\sigma_2^{-1} \circ \sigma_2' = \sigma^*(\tilde{y}'^{(2)} \to \tilde{y}^{(2)}) \tag{39}$$

Algorithm 4 SC-RRE

```
Apply degree-reduction regularization $A$ to obtain $A_{\mathrm{re}}$, as discussed in Zhou and Amini (2019).
Obtain $A_{\mathrm{re}}^{(K\wedge L)} = \widehat{Z}_1\hat{\Sigma}\widehat{Z}_2^T$, the $(K\wedge L)$-truncated SVD of $A_{\mathrm{re}}$.
Output $\mathscr{K}(\widehat{Z}_1\hat{\Sigma})$ where $\mathscr{K}$ is an isometry-invariant $\kappa$-approximate $k$means algorithm.
```

that is, $\sigma_2^{-1}\circ\sigma_2'$ can be obtained (assuming ε' is sufficiently small) by optimally matching $\tilde{y}'^{(2)}$ to $\tilde{y}^{(2)}$, both of which we observe in practice. See the commutative diagram in Figure 2(b). In order to make the above argument precise, we need to justify the first and third claims. We will discuss the details in Appendix B.4. The above matching process can be repeated over all the two-blocks $\tilde{y}^{(q-1,q)}$ to get a consistent set of global labels whose overall misclassification rate is no more than twice that of the original two-blocks (cf. ε' versus ε).

5.2. Results for Algorithm 3

5.2.1. General initialization

Before studying the spectral initialization, let us give a general bound on the misclassification rate of Algorithm 3, assuming sufficiently good initial labels. Assume that the initial labels obtained in steps 3 and 4 of the algorithm are γ_1-good in the sense

$$\mathrm{Mis}(\tilde{y}, y) \le \frac{\gamma_1}{\beta K}, \quad \mathrm{Mis}(\tilde{z}, z) \le \frac{\gamma_1}{\beta L}, \tag{B3}$$

with γ_1 satisfying

$$\gamma_1 \;\le\; \frac{1}{384\beta^2\omega}\Big(\frac{I_{\min}}{8L\|\Lambda\|_\infty} \wedge \frac{I^{\mathrm{col}}_{\min}}{16K\|\Gamma\|_\infty}\Big). \tag{40}$$

Any other initialization algorithm besides spectral clustering can be used, as long as the above guarantee on its output holds. We also need the following weaker version of (A4):

$$\beta\omega(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty) = o\Big(\Big[\frac{I_{\min}\wedge I^{\mathrm{col}}_{\min}}{Q\log Q(K\vee L)}\Big]^a\Big), \text{ for some } a>0. \tag{A4$'$}$$

Theorem 3. *Assume that the model parameters satisfy $I_{min}\wedge I^{col}_{min}\to\infty$, $\Lambda_{\min}\to\infty$, (A1), (A2), (A3) and (A4′), and the initial labels satisfy (40). Then, for some $\zeta = o(1)$, $\widehat{y}$ outputted by Algorithm 3 satisfies*

$$\mathrm{Mis}_k\,(\widehat{y},\, y) = O\Big(\omega\sum_{r\neq k}\Big(1+\frac{1}{\varepsilon_{kr}}\Big)\exp\big(-I_{kr} - \Big(\frac{1}{2}-\zeta\Big)\log\Lambda_{\min}\big)\Big) \tag{41}$$

for every $k\in[K]$ with probability $1-o(1)$.

We refer to Section 3 for the definition of the parameters involved in the rate given in (41).

5.2.2. Spectral initialization

Theorem 3 requires initial labels that satisfy (40). A spectral algorithm, namely SC-RRE given in Algorithm 4 can provide such initialization. (The acronym stands for reduced-rank efficient spectral clustering.) The algorithm is presented and analyzed in Zhou and Amini (2019). One performs a truncated SVD of rank $r := K \wedge L$ on a regularized version of the adjacency matrix, A_{re}, to obtain $\widehat{Z}_1 \hat{\Sigma} \widehat{Z}_2^T$, where $\hat{\Sigma}$ is the diagonal matrix retaining the top r largest singular values of A_{re}, and $\widehat{Z}_1 \in \mathbb{R}^{n \times r}$ and $\widehat{Z}_2 \in \mathbb{R}^{n \times r}$ are the matrices of the corresponding singular vectors. One then runs a k-means type algorithm on $\widehat{Z}_1 \hat{\Sigma}$, rather than $\widehat{Z}_1$ which is the more common approach in spectral clustering. This allows one to state consistency results without a reference to the gap in the spectrum of A_{re}, while still retaining the attractive feature of the latter approach, namely, the computational efficiency of running k-means on a matrix of reduced dimension. The "isometry-invariant" qualification used in Algorithm 4 means that the k-means algorithm should only be sensitive to the pairwise distances of the data points. We refer to Zhou and Amini (2019) for a detailed discussion. In particular, one has the following bound on the misclassification rate of SC-RRE:

Theorem 4 (Zhou and Amini (2019))**.** *Let $\alpha = m/n$ and $\Lambda_\wedge^2 := \min_{t \neq s} \|\Lambda_{s*} - \Lambda_{t*}\|^2$. Consider the spectral algorithm given in Algorithm 4, assume* (A2) *and that for a sufficiently small $C_1 > 0$,*

$$\beta^2 KL(K \wedge L)\, \alpha \frac{\|\Lambda\|_\infty}{\Lambda_\wedge^2} \leq C_1 (1+\kappa)^{-2}. \tag{42}$$

Then the algorithm outputs estimated row labels $\tilde{y}$ satisfying

$$\text{Mis}(\tilde{y}, y) \;\leq\; C_1^{-1}\, (1+\kappa)^2 \beta L(K \wedge L)\, \alpha \Big(\frac{\|\Lambda\|_\infty}{\Lambda_\wedge^2} \Big).$$

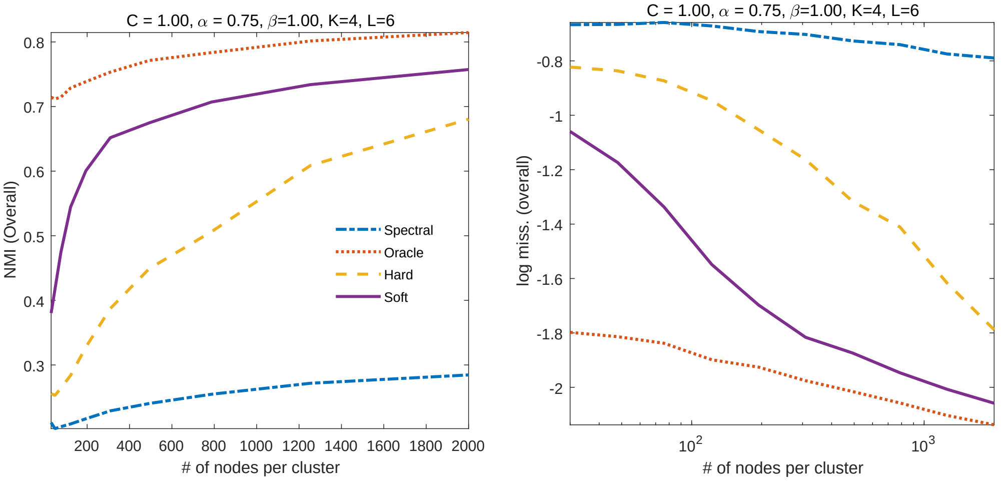
Here, Λ_{s*} refers to the sth row of the mean parameter matrix Λ (cf. Section 2.1). Combining Theorems 3 and 4, one obtains Theorem 1. Some work is required to translate the bound of the Theorem 4 to be applicable to sub-blocks. See Section D.1 for details.

6. Simulations

We provide some simulation results to corroborate the theory. We generate from the SBM of Section 2.1 with the following connectivity matrix

$$P = C \frac{\big[\log(mn)\big]^\alpha}{\sqrt{mn}} B, \quad B = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 6 & 1 \\ 3 & 4 & 5 & 6 & 1 & 2 \\ 4 & 5 & 6 & 1 & 2 & 3 \end{bmatrix}. \tag{43}$$

Note that B does not have any clear assortative or dissortative structure. We let $n = Kn_0$ and $m = Ln_0$, and we vary n_0. All clusters (both row and column) will have the same number of nodes n_0. By changing α, we can study different regimes of sparsity. In particular, when $\alpha \in (0,1)$, we are in the regime where weak recovery is possible but not exact (or strong) recovery. We consider both the misclassification rate, and the normalized mutual information (NMI) as measures of performance. NMI is a measure of accuracy which is between 0 and



(a) (b)

Figure 3: Plots of (a) the (overall) NMI and (b) the corresponding log. misclassification rate, for the SBM model with connectivity matrix (43). The four algorithms considered are the `Spectral` clustering of Algorithm 4, `Soft` and `Hard` versions of Algorithm 1 and the `Oracle` algorithm of Section 2.2.

1 (=perfect match). The NMI is quite sensitive to mismatch and tends to reveal discrepancies between methods more clearly. Figure 3(a) shows the overall NMI versus n_0. Figure 3(b) illustrates the corresponding log. misclassification rates.

We have considered four algorithms: (1) `Spectral`: the spectral clustering of Algorithm 4. (2) `Soft`: Algorithm 1 with flat prior, no inner loop and no conversion to hard labels. (3) `Hard`: Algorithm 1 with flat prior, no inner loop and conversion to hard labels after each label computation. (4) `Oracle`: The oracle classifier discussed in Section 2.2 and Remark 2: Assuming the knowledge of z and Λ, we obtain $\widehat{y}$ by the likelihood ratio classifier, and similarly obtain $\widehat{z}$, assuming the knowledge of y and Γ.

Figure 3 shows the results for $\alpha = .75$ (a regime where no exact recovery is possible) and $C = 1$. Both the soft and hard versions of Algorithm 1 are initialized with the spectral clustering and both significantly improve over it. The soft version of Algorithm 1 also outperforms the hard version as one would expect: soft labels carry more information between iterations. It is interesting to note that, as predicted by the theory, the slope for the log-misclassification rate of Algorithm 1 approaches that of the oracle, which is especially clear for the soft version in Figure 3(b). Simulation results for various other settings can be found in Appendix H, showing qualitatively similar behavior.

Acknowledgement

We thank Yunfeng Zhang for helpful discussions.

Appendix A. Additional comments

A.1. Comments on edge splitting

One needs independent versions of the adjacency matrix in different stages of the algorithm. To achieve this goal, *edge splitting* was introduced in Abbe and Sandon (2015). The idea is that one can regard the two (or more) graphs obtained from edge splitting to be nearly independent. To be specific, let $\mathbb{P}_1$ be the joint probability measure corresponding to a pair of graphs G_1 and G_2 generated independently with connectivity matrices qP and $(1-q)P$. Let $\mathbb{P}_2$ be the joint probability measure on G_1 and G_2 obtained by edge splitting from a single SBM with connectivity matrix P, assigning every edge independently to either G_1 or G_2 with probabilities q and $1-q$. Then, $\mathbb{P}_1$ and $\mathbb{P}_2$ have the same marginal distributions. Having a vanishing total variation between $\mathbb{P}_1$ and $\mathbb{P}_2$ is necessary for further analysis which, as was pointed out by (Abbe and Sandon, 2015, pp. 46-47), is equivalent to showing that under $\mathbb{P}_1$, G_1 and G_2 do no share any edge, with high probability. Letting $\widetilde{P}_{\min} = \min_{k\ell} \widetilde{P}_{k\ell}$,

$$\mathbb{P}_1(G_1 \text{ and } G_2 \text{ do not share edges}) \le \Big(1 - \frac{(1-q)q\widetilde{P}^2_{\min}(\log n)^2}{n^2}\Big)^{n^2}$$

which is strictly bounded away from 1 unless $(1-q)q\widetilde{P}^2_{\min}(\log n)^2 = o(1)$, that is, the connectivity matrix of either G_1 or G_2 should vanish faster than $1/n$. Our consistency result will not hold in this regime. Thus, edge splitting cannot be used to derive the results in this paper, and we introduce the block partitioning idea to supply us with the independent copies necessary for analysis. Another technical issue about edge splitting is discussed in Remark 4.

A.2. Discussion

Our results do not directly apply to the symmetric case, due to the dependence between the upper and lower triangular parts of the adjacency matrix A. However, a more sophisticated two-stage block partitioning scheme can be used to derive similar bounds under mild extra assumptions. One starts with an asymmetric partition into blocks of sizes $\{qn, (1-q)n\} \times \{qn, (1-q)n\}$, for $q = 1/Q \to 0$ very slowly. In the first stage, one applies a similar procedure as described in Algorithm 3 on the upper triangular portion of the large subblock $(1-q)n \times (1-q)n$, followed by the application of the LR classifier on the fat block $qn \times (1-q)n$ to obtain very accurate row labels of the small block $qn \times qn$.. One then repeats the process using the “leave-one-out” of Gao et al. (2017), but applied to small blocks $qn \times qn$ rather than individual nodes. We leave the details for a future work.

It was also shown by (Gao et al., 2017, Theorem 5) that their equivalent of condition (A1) can be removed by modifying the algorithm. In their setting, without assuming $a \asymp b$, a misclassification rate of $\exp(-(1-\varepsilon)I)$ is achievable, where $\varepsilon \in (0,1)$ is a variable in the new version of their algorithm. If those arguments can be extended to the general block model, it will be possible to relax the requirements on ω in (A3) and (A4). When $K, L = O(1)$, one can completely remove sparsity condition (A3) using a much sharper Poisson-binomial approximation than what we have used in this paper. Finally, we suspect that our result could be generalized beyond SBMs to biclustering arrays where the row and column sums over clusters follow Poissonian central limit theorems. We will explore these ideas in the future.

A.3. PL naming

We have borrowed the name pseudo-likelihood (PL) from Amini et al. (2013) based on which the algorithms in this paper are derived. In Amini et al. (2013), the setup is that of the symmetric SBM, and in order to treat the full likelihood as the product of independent (over nodes $i = 1, \dots, n$) of the mixture of Poisson vectors, one has to ignore the dependence among the upper and lower triangular parts of the adjacency matrix, making the PL naming more inline with the traditional use of the term. In our bipartite setup, there is no such dependence to ignore, but we have kept the name PL for consistency with Amini et al. (2013) and ease of use. We interpret the "pseudo" nature of the likelihood as the approximation used in the block compression stage (with imperfect labels) and in replacing Poisson-binomial distribution with the Poisson.

Appendix B. Preliminary analysis

We start by analyzing the properties of the operators introduced in Sections 4.1 and 4.2, for some fixed (deterministic) initial labels $\tilde{y}$ and $\tilde{z}$. We assume that these labels satisfy:

$$\operatorname{Mis}(\tilde{y}, y) \le \frac{\gamma}{\beta K}, \quad \operatorname{Mis}(\tilde{z}, z) \le \frac{\gamma}{\beta L}.$$

We call such labels γ-*good* as in Section 5.2.1. Throughout, $\tilde{\Lambda}$ will be used to denote a generic deterministic approximation of the true row mean parameter Λ. The *relative* ℓ_∞ *ball* of radius δ centered at Λ, that is,

$$\mathscr{B}_\Lambda(\delta) := \{\tilde{\Lambda} : \|\tilde{\Lambda} - \Lambda\|_\infty \le \delta \|\Lambda\|_\infty\}, \tag{44}$$

will play a key role in our arguments. For sufficiently small δ and true Λ, $\mathscr{B}_\Lambda(\delta)$ will be the set of δ-*good* row mean parameters.

B.1. Fixed label analysis

We first present the analysis assuming that all the operations are performed on the entire adjacency matrix A. In Appendix B.2, these results are extended to be applicable to sub-blocks of A. Recall the definitions of the mean parameters an their estimates from Section 2.1. In particular, we recall that $\lambda_{k*}(y, \tilde{z})$ is the mean of $b_{i*}(\tilde{z})$ for any node i with $y_i = k$. These mean parameters form the kth row of $\Lambda(y, \tilde{z})$. Our first main lemma illustrate that whenever the initial labels $\tilde{z}$ and $\tilde{y}$ are γ-good, then the parameters $\Lambda(y, \tilde{z})$ as well as the corresponding estimates $\hat{\Lambda}$ defined in (22) are close to the truth, that is Λ.

Lemma 1 (Parameter consistency)**.** *Let $C_\gamma = C_{\gamma,\beta} = \beta^2\gamma/(1-\gamma)$, assume that $6C_\gamma\omega \le 1$, and let $h_c(\tau) := \frac{3}{4c}\tau \log\left(1 + \frac{2c}{3}\tau\right)$. Then under assumptions* (A1), (A2) *and* (B3)*, we have*

(a) $\|\Lambda(y, \tilde{z}) - \Lambda\|_\infty \le C_\gamma \|\Lambda\|_\infty, \quad \|\Lambda(y, \tilde{z})\|_\infty \le 2\|\Lambda\|_\infty.$

(b) $\|\Lambda(\tilde{y}, \tilde{z}) - \Lambda(y, \tilde{z})\|_\infty \le 2\gamma \|\Lambda\|_\infty, \quad \|\Lambda(\tilde{y}, \tilde{z})\|_\infty \le 4\|\Lambda\|_\infty.$

(c) $\|\hat{\Lambda} - \Lambda(\tilde{y}, \tilde{z})\|_\infty \le 4\tau \|\Lambda\|_\infty$ *with probability at least* $1 - 2p_1$ *where*

$$p_1 = p_1(\tau;\, n, \Lambda_{\min}, \beta) := KL \exp\Big(-\frac{n\Lambda_{\min}\, h_1(\tau)}{4\beta K}\Big), \quad \forall \tau > 0, \tag{45}$$

(d) $\|\Lambda(y, \tilde{z}) - \Lambda\|_\infty \le C_\gamma \|\Lambda\|_\infty,$

and $\hat{\Lambda}$ *is as defined in* (22)*. In particular, all the estimates* $\Lambda(y, \tilde{z})$*,* $\Lambda(\tilde{y}, \tilde{z})$ *and* $\hat{\Lambda}$ *are within relative* ℓ_∞ *distance of at most* $4(C_\gamma + \tau)$ *from* Λ.

The lemma is proved in Appendix F.1. Note that the lemma implies that $\hat{\Lambda} \in \mathscr{B}_\Lambda(4(C_\gamma+\tau))$ with the stated probability.

Our second key lemma shows that the LR classifiers in (34) are uniformly dominated, over $\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta)$, by a single (perturbed) classifier. To state this result, recall the block compression $\boldsymbol{b}(\tilde{z}) := \mathcal{B}(A;\tilde{z})$ given in (27), and define the following:

$$\begin{aligned}
Y_{ikr}(b_{i*},\tilde{\Lambda}) &:= \Psi(b_{i*};\tilde{\lambda}_{r*} \mid \tilde{\lambda}_{k*}) = \sum_{\ell=1}^{L} b_{i\ell} \log \frac{\tilde{\lambda}_{r\ell}}{\tilde{\lambda}_{k\ell}} + \tilde{\lambda}_{k\ell} - \tilde{\lambda}_{r\ell}, && (46)\\
Z_{ik}(b_{i*},\tilde{\Lambda}) &:= 1\{Y_{ikr}(\tilde{\Lambda}) \geq 0, \text{ for some } r \neq k\}. && (47)\\
S_k(\boldsymbol{b},\tilde{\Lambda}) &:= \frac{1}{n_k(y)} \sum_{i:y_i=k} Z_{ik}(b_{i*},\tilde{\Lambda}), && (48)
\end{aligned}$$

where Ψ is the Poisson log-likelihood ratio defined in (33). Thus, Y_{ikr} is the (pseudo) log-likelihood ratio, for $k,r \in [K]$, measuring the relative likelihood of row i having label k. We note that $Y_{ikr}(\tilde{\Lambda}) < 0, \forall r \neq k$ implies $\widehat{y}_i := (\mathrm{LR}(A,\tilde{\Lambda},\tilde{z}))_i = k$. Thus, $S_k(b_{i*},\tilde{\Lambda})$ is the misclassification rate for the LR classifier over the kth row-class, i.e., $\mathrm{Mis}_k(\widehat{y},y)$. Let

$$J_{kr} = L\|\Lambda\|_\infty / I_{kr}. \tag{49}$$

Recalling definitions of ε_{kr}, ω and β from Section 3, set

$$\begin{aligned}
\eta' &:= \eta'(\delta;\Lambda) = 8\omega\delta L\|\Lambda\|_\infty = 8\omega\delta J_{kr} I_{kr}, && (50)\\
\eta_{kr} &:= \eta_{kr}(\delta;\omega,\beta,m,\Lambda)\\
&= 21\delta\omega L\|\Lambda\|_\infty + \frac{5\beta L^2\|\Lambda\|_\infty^2}{m} + \log\Big[11\omega\Big(\frac{1}{\varepsilon_{kr} - 2\omega(1+\varepsilon_{kr})\delta} + 1\Big)\Big] - \frac{1}{2}\log\Lambda_{\min}. && (51)
\end{aligned}$$

We have the following key lemma:

Lemma 2 (Uniformity of LRC in mean parameters)**.** *Fix any row label $\tilde{z}$ and let $\boldsymbol{b} = \boldsymbol{b}(\tilde{z})$ be the corresponding column compression. Let $\Lambda' = \Lambda(y,\tilde{z})$ be the row mean parameter associated with $\boldsymbol{b}$. Assume (A1), (A2), and $\Lambda' \in \mathscr{B}_\Lambda(\delta)$ with $3\,\omega\delta < 1$. Then, for all $k,r \in [K]$, $k \neq r$, and all $i : y_i = k$, we have the following bounds:*

(a) With η' defined as in (50)*,*

$$\mathbb{P}\big(\,\exists\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta),\ Y_{ikr}(b_{i*},\tilde{\Lambda}) \geq 0\,\big) \ \leq\ \exp(-I_{kr} + \eta'). \tag{52}$$

(b) If in addition $\varepsilon_{kr} - 2\omega\delta > 0$, then with η_{kr} defined as in (51)*,*

$$\mathbb{P}\big(\,\exists\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta),\ Y_{ikr}(b_{i*},\tilde{\Lambda}) \geq 0\,\big) \ \leq\ \exp\big(-I_{kr} + \eta_{kr}\big). \tag{53}$$

The proof of Lemma 2(b) appears in Appendix F.5, and that of part (a) in Appendix G.5.

Remark 3 (Typical setting)**.** In the error exponent in Lemma 2(b), i.e. $-I_{kr} + \eta_{kr}$, the first three terms in (51) are positive and constitute the undesirable part of the bound. Our goal is to keep these terms dominated at the final stage of the algorithm, i.e., make them $o(\log\Lambda_{\min})$, by making δ sufficiently small. For now, let us introduce a simple *typical setting* to give some idea of the order of η_{kr}. In the first reading, one can consider the case where $\beta, \omega = O(1)$, $I_{kr} \asymp I \to \infty$ for all k, r and some I, and assume that $L\|\Lambda\|_\infty / I = O(1)$ and (A5) holds. In this setting, $J_{kr} = O(1)$ and we have $\eta_{kr} = C(\delta + m^{-1}I)I - \frac{1}{2}\log \Lambda_{\min}$ for some constant C. Keeping these typical orders in mind will be helpful in understanding the statements of the subsequent results.

It is also worth noting that we always have $J_{kr} \geq \frac{1}{2}$. which follows from the general bound $I_{kr} \leq 2L\|\Lambda\|_\infty$. Another important quantity is C_γ in Lemma 1, which in the typical setting behaves as $C_\gamma \asymp \gamma$ when $\gamma \to 0$.

Combining Lemma 2 with the Markov inequality, we can get uniform control on the misclassification rate of the LR classifier in its parameter argument (i.e., $\hat{\Lambda}$):

Lemma 3. *Fix $k \in [K]$ and $\tilde{z} \in [L]^m$. Let $\hat{\Lambda} \in \mathbb{R}_+^{K \times L}$ be any random matrix and set $\widehat{y}(\tilde{z}) := \mathrm{LR}(A, \hat{\Lambda}, \tilde{z})$. Assume that* (52) *holds. Then, for any $u \in \mathbb{R}$, we have*

$$\mathrm{Mis}_k\big(\widehat{y}(\tilde{z}), y\big) \leq \sum_{r \neq k} \exp\big(-I_{kr} + \eta' + u\big),$$

with probability at least $1 - e^{-u} - \mathbb{P}\big(\hat{\Lambda} \notin \mathscr{B}_\Lambda(\delta)\big)$. The result is also true if we replace η' by η_{rk} when (53) *holds.*

Remark 4. Edge splitting (ES) was proposed in Abbe and Sandon (2015) to generate nearly independent copies from a single network. One might ask whether combining the edge splitting idea with Lemma 3 is enough to give us a result similar to Theorem 1. In ES, edges are randomly assigned to two graphs G_1 and G_2, with probabilities q and $1-q$. The new graphs G_1 and G_2 will follow a SBM with a reduced connectivity matrix (by a factor of q and $1-q$ respectively). Hence, the corresponding parameters Λ and I are reduced by the same factor; for example I will be scaled to qI for G_1. Let us consider the typical setting where $\beta, K, L, \omega, \varepsilon_{kr} = O(1)$ and $I_{kr} \asymp I$ for all k, r and some I; assume the connectivity matrix is symmetric, i.e., $\Lambda = \Gamma$ and $I = I^{\mathrm{col}}$. Let $\tilde{z}$ and $\tilde{y}$ be the labels obtained by performing biclustering on G_1. Lemma 3 in the best case scenario, with the most favorable version of η_{kr}—i.e., ignoring the first three positive terms in (51)—gives a misclassification rate

$$\max\{\mathrm{Mis}\big(\tilde{y}, y\big), \mathrm{Mis}\big(\tilde{z}, z\big)\} \;\leq\; \gamma_2 := \sum_{r \neq k} \exp\Big(-qI_{kr} - \frac{1}{2}\log(q\Lambda_{\min}) + v\Big)$$

for some $v \to \infty$, w.h.p.. In the second stage, given the labels $\tilde{z}$ and $\tilde{y}$, we obtain an estimate of the (row) mean parameters based on G_2, using the natural estimator $\hat{\Lambda}_2 = \mathscr{L}(G_2, \tilde{y}, \tilde{z})$. We then obtain the second stage labels $y(\tilde{z}) := \mathrm{LR}(G_2, \hat{\Lambda}, \tilde{z})$. Let $\Lambda_2 = (1-q)\Lambda$ be the row mean parameter of G_2. By Lemma 1, $\hat{\Lambda}_2 \in \mathscr{B}_{\Lambda_2}(\delta)$ w.h.p for some $\delta \geq \gamma_2$. By Lemma 3, and the perturbation of information (Lemma 7) we have

$$\mathrm{Mis}\big(\widehat{y}(\tilde{z}), y\big) \leq \gamma_3 := \sum_{r \neq k} \exp\Big(-(1-q)I_{kr} + C(1-q)\delta\|\Lambda\|_\infty - \frac{1}{2}\log \Lambda_{\min} + u\Big)$$

for some $u \to \infty$ w.h.p.. To obtain result (14) in Corollary 1, we at least hope to have

$$qI_{kr} + C(1-q)\gamma_2\|\Lambda\|_\infty = o(\log \Lambda_{\min}).$$

So we need $qI_{kr} = o(\log \Lambda_{\min})$ and $(1-q)\gamma_2\|\Lambda\|_\infty = o(\log \Lambda_{\min})$. Assume that we have $qI_{kr} = o(\log \Lambda_{\min})$. Then,

$$\gamma_2 = \sum_{r \neq k} \exp\Big(-qI_{kr} - \frac{1}{2}\log(q\Lambda_{\min}) + v\Big) = O(\Lambda_{\min}^{-1/2-o(1)}/\sqrt{q}).$$

However, this is not sufficient to show $(1-q)\gamma_2\|\Lambda\|_\infty = o(\log \Lambda_{\min})$. Therefore, applying edge splitting and Lemma 3 does not lead to the main result of this paper.

B.2. Analysis on subblocks

We now extend the analysis of Appendix B.1 to be applicable to the sub-blocks obtained by random partitioning. Some care needs to be taken since the true (row and column) mean parameters of the sub-blocks are changed by partitioning, due to the change in the distributions of the labels within each sub-block among the $K \times L$ classes. The deviations of the sub-block class proportions from the global version will be controlled by *a slack parameter* ξ which will be set at the final stage of the proof (see Appendix C.2). Throughout this section, assumptions (A1) and (A2) will be implicit in all the stated lemmas. We will also state the result for a general $2Q \times Q$ partitioning scheme, although $Q = 4$ is enough for the analysis of Algorithm 3.

Recall that the class priors $\pi_\ell(z)$ for the full labels are defined in (30). We will use the same notation for sublabels $z^{(q)}$, that is, $\pi_\ell(z^{(q)})$ is the proportion of labels in $z^{(q)}$ that lie in class ℓ. Note that we have

$$\pi_\ell(z) = \frac{n_\ell(z)}{m}, \quad \pi_\ell(z^{(q)}) = \frac{n_\ell(z^{(q)})}{m/Q}, \quad \text{hence,} \quad \frac{\pi_\ell(z^{(q)})}{\pi_\ell(z)} = Q\frac{n_\ell(z^{(q)})}{n_\ell(z)}, \tag{54}$$

since $z^{(q)}$ has length m/Q. We similarly we have $\pi_k(y^{(q)}) = n_k(y^{(q)})/(n/(2Q))$. We will work under the assumption that the partitioning scheme satisfies:

$$\max_{k,q} |\pi_k(y^{(q)}) - \pi_k(y)| \le \xi \quad \text{and} \quad \max_{\ell,q} |\pi_\ell(z^{(q)}) - \pi_\ell(z)| \le \xi, \tag{B4a}$$

$$\xi \le \min\Big(\frac{1}{2\beta K}, \frac{1}{2\beta L}\Big). \tag{B4b}$$

When these conditions hold, we call the scheme a *good partition*. We note that these conditions combined with (A2) give,

$$\Big|\frac{\pi_\ell(z^{(q)})}{\pi_\ell(z)} - 1\Big| \le \xi L\beta \le \frac{1}{2} \implies \frac{1}{2}\frac{1}{\beta L} \le \pi_\ell(z^{(q)}) \le \frac{3}{2}\frac{\beta}{L} \tag{55}$$

and similarly for $y^{(q)}$. It follows that both $z^{(q)}$ and $y^{(q)}$ satisfy (A1) with β replaced with 2β.

Each count $n_k(y^{(q)})$ follows a hypergeometric distribution with parameters $(n, n_k(y), n/(2Q))$, that is, the number of nodes labeled k, in a sample of size $n/(2Q)$, from a population of size n, with a total of $n_k(y)$ nodes labeled k. The concentration of the hypergoemtric distribution gives the following:

Lemma 4. (B4a) *holds for random partitioning, with probability at least* $1 - p_2$, *where*

$$p_2 = 2Q(K+L)\exp\big(-\min(n,m)\xi^2/Q\big). \tag{56}$$

The proof of this lemma and others in this section appear in Appendix G.1.

Lemma 5. *Under* (B4a) *and* (B4b), *the true local mean parameters* $\Lambda^{(q)} = (\lambda^{(q)}_{k\ell})$ *satisfy:*

$$\Big|\lambda^{(q)}_{k\ell} - \frac{\lambda_{k\ell}}{Q}\Big| \le (\xi L \beta)\frac{\lambda_{k\ell}}{Q} \le \frac{1}{2}\frac{\lambda_{k\ell}}{Q}, \quad \forall q, k, \ell. \tag{57}$$

In particular, $\Lambda^{(q)}_{\min} \ge \frac{1}{2Q}\Lambda_{\min}$, $\|\Lambda^{(q)}\|_\infty \le \frac{3}{2Q}\|\Lambda\|_\infty$ *and* $\Lambda^{(q)} \in \mathscr{B}_{\Lambda/Q}(\xi L\beta)$ *for all* $q \in [Q]$.

Our main lemma for the sub-blocks establishes the consistency of the local mean parameter estimates $\hat{\Lambda}^{(q',q)}$ for a *good* partitioning scheme. This lemma is an extension of Lemma 1. We recall the operator $\mathscr{L}$ from (21):

Lemma 6 (Local parameter consistency). *Let* $C_\gamma = \beta^2\gamma/(1-\gamma)$ *and* $h_c(\tau) := \frac{3}{4c}\tau\log\big(1+\frac{2c}{3}\tau\big)$ *as in Lemma 1 and assume that* $72\, C_\gamma\omega \le 1$. *Fix the underlying partition and fix* $q, q' \in [Q]$, *and labels* $\tilde{z}$ *and* $\tilde{y}$. *Let*

$$\hat{\Lambda}^{(q',q)} = \mathscr{L}(A^{(q',q)}, \tilde{y}^{(q')}, \tilde{z}^{(q)}).$$

Assume that the partition satisfies (B4a) *and* (B4b), *and the pairs* $(\tilde{z}^{(q)}, z^{(q)})$ *and* $(\tilde{y}^{(q)}, y^{(q)})$ *satisfy the misclassification rate in* (B3). *Then,*

$$\begin{aligned} \|\hat{\Lambda}^{(q',q)} - \Lambda^{(q)}\|_\infty &\le \big(24\, C_\gamma + 6\tau\big)\, \|\Lambda/Q\|_\infty, \quad \textit{and} \\ \|\hat{\Lambda}^{(q',q)} - \Lambda/Q\|_\infty &\le \big(24\, C_\gamma + 6\tau + \xi L\beta\big)\, \|\Lambda/Q\|_\infty \end{aligned}$$

with probability at least $1 - 2p_3$, *where*

$$p_3 = p_3(\tau;\, n, K, \Lambda_{\min}, Q) := KL\exp\Big(-\frac{n\Lambda_{\min}\, h_1(\tau)}{32Q^2\beta K}\Big). \tag{58}$$

We also have

(a) $\|\Lambda^{(q',q)}(y, \tilde{z}) - \Lambda^{(q)}\|_\infty \le 4C_\gamma\|\Lambda^{(q)}\|_\infty$.

(b) $\|\Lambda^{(q',q)}(\tilde{y}, \tilde{z}) - \Lambda^{(q',q)}(y, \tilde{z})\|_\infty \le 2\gamma\|\Lambda^{(q)}\|_\infty$.

(c) $\|\hat{\Lambda}^{(q',q)} - \Lambda^{(q',q)}(\tilde{y}, \tilde{z})\|_\infty \le 4\tau\|\Lambda^{(q)}\|_\infty$, *with probability at least* $1 - 2p_3$.

Remark 5. Similar results to those obtained above hold for the column parameters. Recall that the dual to the row mean parameters Λ are the column mean parameters Γ. The result of Lemma 5 can be translated to the column version by making the following substitutions $\Lambda \to \Gamma$, $Q \to 2Q$ and $L \leftrightarrow K$. For Lemma 6, in addition we need to make $n \to 4m$. (The reason for this is that in (117), in the proof, we need to replace $n/2Q$ with m/Q, and $\Lambda_{\min}/2Q$ with $\Gamma_{\min}/(4Q)$, and the combination of the aforementioned substitutions achieves this. We also note for future reference that the corresponding ω inflation by a factor of 3 remains true for column parameters.) After these substitutions, we obtain the same constant in (58), that is, p_3 has to be replaced with

$$p_3' := p_3(\tau; 4m, L, \Gamma_{\min}, 2Q) = p_3(\tau; m, L, \Gamma_{\min}, Q). \tag{59}$$

B.3. Perturbation of information

Recall the definition of Chernoff information from (8), and let us write $I_{kr} = I_{kr}(\Lambda)$ to explicitly show its dependence on the mean parameter matrix Λ. The following lemma, proved in Appendix G.1, bounds the perturbations of $I_{kr}(\Lambda)$ in Λ:

Lemma 7. *Under* (A1)*, for any* $\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta)$*, we have* $|I_{kr}(\tilde{\Lambda}) - I_{kr}(\Lambda)| \le 2\omega\delta L\|\Lambda\|_\infty$.

B.4. Analysis of the matching step

In this section, we fill in the details of the argument sketched in Section 5.1. Specifically, we need to give sufficient conditions so that the first and the third claims of Section 5.1 hold. We will use the following two lemmas. Recall the notation $\sigma^*(\tilde{y} \to y)$ introduced in Section 2.3 to denote the optimal permutation from the set of labels $\tilde{y}$ to another set y.

Lemma 8. *Let* $\tilde{y}, y \in [K]^n$*, and assume that* $\mathrm{dMis}(\tilde{y}, y) < \frac{1}{2}\min_k \pi_k(y)$*. Then,*

(a) $\sigma^*(\tilde{y} \to y) = \mathrm{id}$*, the identity permutation, and this optimal permutation is unique, and*

(b) $\pi_k(\tilde{y}) > \frac{1}{2}\pi_k(y)$ *for all* k*.*

Note that Lemma 8 implies that if $\mathrm{dMis}(\sigma(\tilde{y}), y) < \frac{1}{2}\min_k \pi_k(y)$ for some permutation σ, then $\sigma^*(\tilde{y} \to y) = \sigma$.

Lemma 9. *Consider three sets of labels* $y, \tilde{y}, \tilde{y}' \in [K]^n$*, and assume that*

$$\max\{\, \mathrm{Mis}(\tilde{y}, y),\, \mathrm{Mis}(\tilde{y}', y) \,\} < \frac{1}{4}\min_k \pi_k(\tilde{y}).$$

Let $\sigma = \sigma^*(\tilde{y} \to y)$ *and* $\sigma' = \sigma^*(\tilde{y}' \to y)$*. Then,* $\sigma^{-1} \circ \sigma' = \sigma^*(\tilde{y}' \to \tilde{y})$*.*

The first claim of Section 5.1 follows from Lemma 8, under the further assumption:

$$\mathrm{Mis}(\tilde{y}^{(q-1,q)}, y^{(q-1,q)}) < \frac{1}{32\beta K}, \quad q \in [Q]. \tag{60}$$

Using the permutation notations (36) of Section 5.1, we have:

Corollary 2. *Under assumptions* (A2), (B4a)*,* (B4b) *and* (60)*,* $\sigma_{q-1,q} = \sigma_{q-1}$ *for all* $q \in [Q]$*.*

The third and final claim of Section 5.1 follows from Lemmas 8 and 9, by applying them to the sub-block labels $y^{(2)}, \tilde{y}^{(2)}, \tilde{y}'^{(2)}$:

Corollary 3. *Under assumptions* (A2), (B4a)*,* (B4b) *and* (60)*,* $\sigma_q^{-1} \circ \sigma'_q = \sigma^*(\tilde{y}'^{(q)} \to \tilde{y}^{(q)})$ *for all* $q \in [Q]$*.*

The proofs of the results of this section are deferred to Appendix G.1.

Appendix C. Proof of Theorem 3

We start with the high-level analysis of Algorithm 3 in Appendix C.1. This analysis is parametrized by many parameters such as ξ , τ_1, τ_1^{col}, τ_2, etc. This allows us to give the high-level idea of the mechanics of the proof without making the arguments obscured by the expressions ultimately chosen for these parameters. In Appendix C.2, we make specific choices about these parameters and finish the proof of Theorem 3.

C.1. Parametrized analysis of Algorithm 3

We now have all the pieces for analyzing Algorithm 3. Let $\tilde{y}_{\text{step}\,5}$ and $\tilde{z}_{\text{step}\,5}$ be the labels from step 5 of of Algorithm 3. As before, in all the lemmas stated, (A1) and (A2) will be implicitly assumed. Consider the following event:

$$\mathfrak{A}_\gamma := \Big\{ \tilde{y}^{(q)}_{\text{step}\,5} \text{ and } \tilde{z}^{(q)}_{\text{step}\,5} \text{ satisfy (B3) with parameter } \gamma, \text{ for all } q \in [Q] \Big\}.$$

We implicitly assume that clusters in $\tilde{z}_{\text{step}\,5}$ and $\tilde{y}_{\text{step}\,5}$ are relabeled according to optimal permutation relative to the truth. In other words, $\tilde{z}_{\text{step}\,5}$ and $\tilde{y}_{\text{step}\,5}$ in the above event are not the raw output of the algorithm, but the relabeled versions (which we do not have access to in practice, but are well-defined and can be used in the proof.) When γ is sufficiently small, this implies that community k in $\tilde{z}_{\text{step}\,5}$ is the same as community k in $\tilde{z}$, for all $k \in [Q]$.

Let Π be the random partition used in Algorithm 3, and let $\mathfrak{P}$ be the event that Π satisfies condition (B4a). By Lemma 4, we have $\mathbb{P}(\mathfrak{P}) \ge 1 - p_2$ where p_2 is given in (56). For the most part, we will work on events of the form $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$. Let us also establish some terminology. By the probability "on an event $\mathfrak{P}$", we mean the probability under the restricted measure $\mathbb{P}_{\mathfrak{P}} := \mathbb{P}(\cdot \cap \mathfrak{P})$. For example, if $\mathfrak{D} = \{\text{property X holds}\}$, we will say that "property X fails" on $\mathfrak{P}$ with probability at most q if $\mathbb{P}(\mathfrak{D}^c \cap \mathfrak{P}) \le q$. In this case, if $\mathfrak{P}$ holds with high probability, say $\ge 1 - p_2$, and q is small, then $\mathfrak{D}$ holds with high probability as well: $\mathbb{P}(\mathfrak{D}) \ge 1 - q - p_2$.

Let $\hat{\Lambda}^{(q)}_{\text{step}\,6} = \mathscr{L}(A^{(q-2,q)}, \tilde{y}^{(q-2)}, \tilde{z}^{(q)})$, $q \in \mathbb{Z}/Q\mathbb{Z}$, be the first local parameter estimates obtained in step 6 of Algorithm 3 (it is easier to work with the shifted index), and let

$$\delta_1 := 24\, C_{\gamma_1} + 6\tau_1 + \xi L \beta. \tag{61}$$

A better name for δ_1, and τ_1 would be δ_1^{row}, and similarly τ_1^{row} contrasting with δ_1^{col} and τ_1^{col} defined later in (64). However, for simplicity, we drop the "row" qualifier here. Recall that ξ is a parameter controlling the tail probability related to the random partition, while τ_1 will be controlling the tail probability $p_3(\tau_1)$ related to the local parameter estimates in Lemma 6. These parameters will be optimized at the end of the argument (see Appendix C.2).

Lemma 10 (First local parameters). *Assume* (B4b) *and* $72\, C_{\gamma_1}\omega \le 1$*, and let* δ_1 *be as defined in* (61)*. Then, on event* $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$*,*

$$\hat{\Lambda}^{(q)}_{\textit{step 6}} \in \mathscr{B}_{\Lambda^{(q)}}(\delta_1), \quad \forall q \in \mathbb{Z}_Q,$$

fails with probability at most $2Q\, p_3$*, where* $p_3 = p_3(\tau_1)$ *as given in* (58)*.*

Proof. Conditioning on blocks G_1 (cf. Section 5) of the (bottom) adjacency matrix A_{bottom}—denoted as $A^{(G_1)}_{\text{bottom}}$—the distribution of blocks $A^{(q-2,q)}, q \in \mathbb{Z}_Q$ used in defining $\hat{\Lambda}^{(q)}_{\text{step}\,6}$ is not changed. Under this conditioning, both initial labels $\tilde{y}_{\text{step}\,5}$ and $\tilde{z}_{\text{step}\,5}$ are deterministic, hence the results of Appendix B.2 apply. We will apply Lemma 6 to $\hat{\Lambda}^{(q)}_{\text{step}\,6}$. Let us verify the conditions of the lemma. On $\mathfrak{A}_{\gamma_1}$, for all $q \in [Q]$, the sublabel pairs $(\tilde{z}^{(q)}_{\text{step}\,5}, z^{(q)})$ and $(\tilde{y}^{(q)}_{\text{step}\,5}, y^{(q)})$ satisfy (B3). On $\mathfrak{P}$, condition (B4a) holds for the random partition and (B4b) holds by assumption. Recall that the random partition is independent of all else, hence conditioning on it does not change the distribution of blocks $A^{(q-2,q)}, q \in \mathbb{Z}_Q$ either. We may then apply

Lemma 6 to conclude that for every $q \in \mathbb{Z}_Q$, conditioned on the partition Π and $A^{(G_1)}_{\text{bottom}}$, the event $\{\hat{\Lambda}^{(q)}_{\text{step}\,6} \notin \mathscr{B}_{\Lambda^{(q)}}(\delta_1)\} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$ holds with probability $\le 2p_3$. Let us write

$$\mathfrak{D} = \big\{\hat{\Lambda}^{(q)}_{\text{step}\,6} \in \mathscr{B}_{\Lambda^{(q)}}(\delta_1),\ \forall q \in \mathbb{Z}_Q\big\}$$

which is the desired event in this lemma. Using the union bound, and removing the conditioning, we have $\mathbb{P}(\mathfrak{D}^c \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}) \le 2Qp_3$, unconditionally. The proof is complete. □

Next we consider the first LR classifier application. Let $\tilde{y}_{\text{step}\,7}$ be the row label estimates in step 7. That is, we have

$$\tilde{y}^{(q-2)}_{\text{step}\,7} = \mathrm{LR}\left(A^{(q-2,q)},\ \hat{\Lambda}^{(q)}_{\text{step}\,6},\ \tilde{z}^{(q)}_{\text{step}\,5}\right)$$

for which we have the following bound on misclassification rate:

Lemma 11 (First LR classifier)**.** *Under the assumptions of Lemma 10, further assume that* $9\omega\delta_1 < 1$. *Let* $\eta^{\mathit{step}\,7} := 2\eta'(\delta_1; \Lambda/Q)$ *where* $\eta'(\cdot)$ *is defined in* (51)*. Then, on event* $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$*,*

$$\mathrm{Mis}_k\left(\tilde{y}^{(q)}_{\mathit{step}\,7}, y^{(q)}\right) \ \le\ \sum_{r \ne k} \exp\Big(-\frac{I_{kr}}{Q} + \eta^{\mathit{step}\,7} + u\Big) =: \gamma^{\mathit{row}}_{2k}, \quad \forall q \in \mathbb{Z}_Q, \tag{62}$$

fails with probability at most $Q(e^{-u} + 2Q\,p_3)$ *where* $p_3 = p_3(\tau_1)$ *as given in* (58)*.*

Proof. Fix $q \in \mathbb{Z}_Q$ and consider $\tilde{y}^{(q-2)}$. As in the proof Lemma 10, we condition on blocks in G_1 so that $\tilde{z}^{(q)}_{\text{step}\,5}$ can be assumed deterministic. We will apply Lemma 2(a) to the subblock $A^{(q-2,q)}$. As discussed earlier, the corresponding ω is inflated to 3ω, hence we need $3(3\omega)\delta_1 < 1$ which we have assumed. We also note that $\Lambda^{(q-2,q)}(y,\tilde{z})$ and $\Lambda^{(q)}$ play the role of $\Lambda(y,\tilde{z})$ and Λ in Lemma 2(b), and we have the needed condition $\Lambda^{(q-2,q)}(y,\tilde{z}) \in \mathscr{B}_{\Lambda^{(q)}}(\delta_1)$ from Lemma 6. Let $b^{(q-2,q)}_{i*}$ be the row block compression of $A^{(q-2,q)}$ based on $\tilde{z}^{(q)}_{\text{step}\,5}$. Then, Lemma 2(a) gives

$$\mathbb{P}\Big(\Big\{\exists\,\tilde{\Lambda} \in \mathscr{B}_{\Lambda^{(q)}}(\delta_1),\ Y_{ikr}\big(b^{(q-2,q)}_{i*}, \tilde{\Lambda}\big) \ge 0\Big\} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\ \Big|\ A^{(G_1)}_{\text{bottom}}, \Pi\Big) \ \le\ \exp\big(-I^{(q)}_{kr} + \eta^{(q)}\big) \tag{63}$$

for all rows i (in row block $q-2$) with $y_i = k$. Here $\Lambda^{(q)}_{\min}$ is the minimum element of $\Lambda^{(q)}$, and

$$\begin{aligned} I^{(q)}_{kr} \ &:=\ I_{kr}(\Lambda^{(q)}) \quad &&\ge\ \frac{I_{kr}}{Q} - 2\omega\delta_1 L\|\Lambda/Q\|_\infty \\ \eta^{(q)} \ &:=\ \eta'(\delta_1;\ \Lambda^{(q)}) \ &&\le\ (1+\delta_1)\,\eta'(\delta_1;\ \Lambda/Q) \end{aligned}$$

where $\eta'(\delta_1; \Lambda^{(q)}) = 8\omega\delta_1 L\|\Lambda^{(q)}\|_\infty$ as defined in (51). The first inequality uses Lemma 7 and the second is obtained using the definition of $\eta'(\cdot)$ combined with $\Lambda^{(q)} \in \mathscr{B}_{\Lambda/Q}(\delta_1)$ (Lemma 5) which implies $\|\Lambda^{(q)}\|_\infty \le (1+\delta_1)\|\Lambda/Q\|_\infty$. By taking expectation in (63), the same bound holds unconditionally.

By Lemma 10, on event $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$, we have $\hat{\Lambda}^{(q)}_{\text{step}\,6} \notin \mathscr{B}_{\Lambda^{(q)}}(\delta_1)$ with probability at most $2Q\,p_3$. Then, applying Lemma 3, we conclude that

$$\mathrm{Mis}_k\left(\tilde{y}^{(q-2)}_{\text{step}\,7}, y^{(q-2)}\right) \ \le\ \sum_{r \ne k} \exp\big(-I^{(q)}_{kr} + \eta^{(q)} + u\big)$$

fails on $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$ with probability $\le e^{-u} + 2Q\,p_3$, for each $q \in [Q]$. Note that

$$-I_{kr}^{(q)} + \eta^{(q)} \;\le\; -\frac{I_{kr}}{Q} + (1 + \delta_1 + 9^{-1})\,\eta'(\delta_1;\; \Lambda/Q).$$

Since $9\omega\delta_1 < 1$ implies $\delta_1 < 9^{-1}$ (recall $\omega \ge 1$), we have $1 + \delta_1 + 9^{-1} < 2$. Combining with the previous bound and applying the union bound over q gives the result. □

Note that we have called the rate in (62) γ_2^{row} for the (column) misclassification rate based on the row information. This rate is faster than initial rate γ_1. Repeating the procedure in steps 6 and 7 for the column labels—as prescribed in step 8 in Algorithm 3–we obtain a similar rate for the misclassification rate of $\tilde{z}_{\text{step}\,8}^{(q)}$ relative to $z^{(q)}$ which we call γ_2^{col}. In deriving γ_2^{col}, we have to make the substitutions in Remark 5, and particular, $\Lambda \to \Gamma$ where Γ is the column mean parameters defined in Section 2.1. (A minor exception is when counting the number of blocks which will still be Q rather than $2Q$.) Recall the definition of the column information matrix $(I_{\ell r}^{\text{col}})$ from (9). Letting

$$\delta_1^{\text{col}} := 24\,C_{\gamma_1} + 6\tau_1^{\text{col}} + \xi K\beta, \tag{64}$$

we obtain the following counterpart of Lemma 11:

Corollary 4 (First LR classifier, column version)**.** *Under the assumptions of Lemma 10, further assume that $9\omega\delta_1^{col} < 1$. Let $\eta^{step\,8} := 2\eta'(\delta_1^{col}; \Gamma/(2Q))$ where $\eta'(\cdot)$ is defined in* (51)*. Then, on event $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$,*

$$\mathrm{Mis}_\ell\left(\tilde{z}_{step\,8}^{(q)}, z^{(q)}\right) \;\le\; \sum_{r \ne \ell} \exp\Big(-\frac{I_{\ell r}^{col}}{2Q} + \eta^{step\,8} + u^{col}\Big) =: \gamma_{2\ell}^{col}, \quad \forall q \in \mathbb{Z}_Q, \tag{65}$$

fails with probability at most $Q(e^{-u^{col}} + 2Q\,p_3')$, where $p_3' = p_3'(\tau_1^{col})$ as given in (59)*.*

Let $\gamma_2^{\text{col}} := \max_{k \in [K]} \gamma_{2k}^{\text{col}}$, $\gamma_2^{\text{row}} := \max_{\ell \in [L]} \gamma_{2\ell}^{\text{row}}$ and

$$\gamma_2 := \max\{\beta K \gamma_2^{\text{row}}, \beta L \gamma_2^{\text{col}}\}. \tag{66}$$

By (7), we have that (62) and (65) imply

$$\mathrm{Mis}\left(\tilde{y}_{\text{step}\,7}^{(q)}, y^{(q)}\right) \le \gamma_2^{\text{col}} \le \frac{\gamma_2}{\beta K}, \quad \mathrm{Mis}\left(\tilde{z}_{\text{step}\,8}^{(q)}, z^{(q)}\right) \le \gamma_2^{\text{row}} \le \frac{\gamma_2}{\beta L} \tag{67}$$

Thus, if we consider the following event:

$$\mathfrak{B}_\gamma := \Big\{\tilde{y}_{\text{step}\,7}^{(q)} \text{ and } \tilde{z}_{\text{step}\,8}^{(q)} \text{ satisfy (B3) with parameter } \gamma, \text{ for all } q \in [Q]\Big\},$$

after Step 8, we can work on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$ which holds with high probability: Combining Lemma 11 and Corollary 4, by union bound, $\mathbb{P}(\mathfrak{B}_{\gamma_2}^c \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}) \le Q(2e^{-u} + 2Q(p_3 + p_3'))$, hence

$$\begin{aligned} \mathbb{P}(\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}) \;&\ge\; \mathbb{P}\big(\mathfrak{B}_{\gamma_2} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\big) \\ &=\; \mathbb{P}\big(\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\big) - \mathbb{P}\big(\mathfrak{B}_{\gamma_2}^c \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\big) \\ &\ge\; 1 - \mathbb{P}(\mathfrak{A}_{\gamma_1}^c) - \mathbb{P}(\mathfrak{P}^c) - Q(2e^{-u} + Q(p_3 + p_3')). \end{aligned} \tag{68}$$

Let $\hat{\Lambda}_{\text{step}\,9}^{(q)} = \mathscr{L}(A^{(q-3,q)}, \tilde{y}^{(q-3)}, \tilde{z}^{(q)})$, $q \in \mathbb{Z}_Q$, be the second local parameter estimates obtained in step 9 of Algorithm 3. Let

$$\delta_2 := 24C_{\gamma_2} + 6\tau_2. \tag{69}$$

Lemma 12 (Second local parameters)**.** *Assume* (B4b) *and* $72\,C_{\gamma_2}\omega \le 1$*, and let* δ_2 *be as defined in* (69)*. Then, on event* $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$*,*

$$\hat{\Lambda}^{(q)}_{step\ 9} \;\in\; \mathscr{B}_{\Lambda^{(q)}}(\delta_2), \quad \forall q \in \mathbb{Z}_Q$$

fails with probability at most $2Q\,p_3$*, where* p_3 *is given in* (58)*.*

Proof. Conditioning on blocks $G_1 \cup G_2$ (cf. Section 5) of the adjacency matrix A, the distribution of blocks $A^{(q-3,q)}$ used in defining $\hat{\Lambda}^{(q)}_{\text{step}\,9}$ is not changed. Under this conditioning, both initial labels $\tilde{y}_{\text{step}\,7}$ and $\tilde{z}_{\text{step}\,8}$ are deterministic, hence the results of Appendix B.2 apply. On $\mathfrak{B}_{\gamma_2}$, for all $q \in \mathbb{Z}_Q$, the sublabel pairs $(\tilde{z}^{(q)}_{\text{step}\,8}, z^{(q)})$ and $(\tilde{y}^{(q)}_{\text{step}\,7}, y^{(q)})$ satisfy (B3). The rest of the proof follows that of Lemma 10. □

The key is that δ_2 is much smaller than δ_1, due to $\gamma_2 \ll \gamma_1$ (typically), i.e., the second parameter estimates are much more accurate. Let $\hat{\Lambda}_{\text{step}\,10} = \sum_q \hat{\Lambda}^{(q)}_{\text{step}\,9}$ be the estimate of the global mean parameters obtained in step 10 of Algorithm 3. According to Lemma 12, on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$,

$$\|\hat{\Lambda}^{(q)}_{\text{step}\,9} - \Lambda/Q\|_\infty \le \delta_2 \|\Lambda/Q\|_\infty, \; \forall q \quad \text{hence,} \quad \|\hat{\Lambda}_{\text{step}\,10} - \Lambda\|_\infty \le \delta_2 \|\Lambda\|_\infty \tag{70}$$

fails with probability $\le 2Q\,p_3$, where we have used triangle inequality. That is, on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$, we have $\hat{\Lambda}_{\text{step}\,10} \in \mathscr{B}_\Lambda(\delta_2)$ with high probability.

Remark 6. Note that we could have used $Q\hat{\Lambda}^{(q)}_{\text{step}\,9}$ (for any $q \in \mathbb{Z}_Q$) as our estimate $\hat{\Lambda}_{\text{step}\,10}$, leading to the same bound as in (70). The results would be the same, though in practice, we expect the version given in the Algorithm 3 to perform better. We also note that on $\mathfrak{B}_{\gamma_2}$, the sublabels $(\tilde{z}^{(q)}_{\text{step}\,8},\; q \in \mathbb{Z}_Q)$ automatically define a consistent global label vector $\tilde{z}_{\text{step}\,8}$, and similarly for row labels $\tilde{y}_{\text{step}\,7}$.

Lemma 13 (Second LR classifier)**.** *Under the assumptions of Lemma 12, further assume that* δ_2 *defined in* (69) *satisfies* $3\omega\delta_2 < 1$ *and* $6C_{\gamma_2}\omega \le 1$*. Let* $\eta^{step\ 11}_{kr} := \eta_{kr}\big(\delta_2;\ \omega, \beta, m, \Lambda\big)$*. Then, on event* $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$*,*

$$\text{Mis}_k\left(\widehat{y}_{top},\, y_{top}\right) \;\le\; \sum_{r \neq k} \exp\Big(-I_{kr} + \eta^{step\ 11}_{rk} + v\Big) =: \gamma_3, \tag{71}$$

fails with probability at most $e^{-v} + 2Q\,p_3$ *where* $p_3 = p_3(\tau_2)$ *as given in* (58)*. The same result holds for* $\eta^{step\ 11}_{kr} = \eta'(\delta_2; \Lambda)$*.*

Proof. As in the proof Lemma 12, we condition on blocks in $G_1 \cup G_2$ so that $\tilde{z}_{\text{step}\,8}$ can be assumed deterministic. We will apply Lemma 2(b) to A_{top}. Let Top $\subset [n]$ denote the row indices of A_{top}. Since all the columns are present in A_{top}, we can directly apply Lemma 2(b) (in contrast to the argument in Lemma 11), that is, the relevant row mean parameters are $\Lambda(y, \tilde{z})$ and Λ—the same as those for the whole matrix A. The needed condition $\Lambda(y, \tilde{z}) \in \mathscr{B}_\Lambda(\delta_2)$ is supplied by Lemma 1. Let $b^{\text{step}\,11}_{i*} = b_{i*}(\tilde{z}_{\text{step}\,8})$ be the block compression in step 11 of the algorithm. Then, Lemma 2(b) gives (after conditioning on $A^{(G_1 \cup G_2)}_{\text{bottom}}$ and then removing the conditioning as in (63))

$$\mathbb{P}\Big(\Big\{\exists\, \tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta_1),\; Y_{ikr}\big(b^{\text{step}\,11}_{i*}, \tilde{\Lambda}\big) \ge 0\Big\} \cap \mathfrak{B}_{\gamma_2}\Big) \;\le\; \exp\Big(-I_{kr} + \eta^{\text{step}\,11}_{kr}\Big). \tag{72}$$

for any $i \in \text{Top}$ with $y_i = k$. By (70), on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$, we have $\hat{\Lambda}_{\text{step}\,10} \notin \mathscr{B}_\Lambda(\delta_2)$ with probability at most $2Q\,p_3$. Then, applying Lemma 3, we conclude (71) as desired. The last statement of the theorem follows if we apply Lemma 2(a) in place of Lemma 2(b) throughout. □

The same exact bound holds for $\widehat{y}_{\text{bottom}}$ in step 12, with the same probability. Hence, by union bound, the same bound on misclassification rate holds for the final row labels $\widehat{y}$ in step 13, with probability inflated by a factor of 2; that is, $\text{Mis}_k(\widehat{y}, y) \le \gamma_3$ fails on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$, with probability at most $2(e^{-v} + 2Q\,p_3)$.

To summarize, under the conditions of the lemmas, we have

$$\begin{aligned} \mathbb{P}\big(\,\text{Mis}_k(\widehat{y}, y) > \gamma_3\big) &\le \mathbb{P}\big(\{\,\text{Mis}_k(\widehat{y}, y) > \gamma_3\} \cap \mathfrak{B}_{\gamma_2} \cap \mathfrak{P}\big) + \mathbb{P}\big((\mathfrak{B}_{\gamma_2} \cap \mathfrak{P})^c\big) \\ &\le 2\big(e^{-v} + 2Q\,p_3(\tau_2)\big) + \mathbb{P}\big((\mathfrak{B}_{\gamma_2} \cap \mathfrak{P})^c\big) \\ &\le 2\big(e^{-v} + 2Q\,p_3(\tau_2)\big) + \mathbb{P}(\mathfrak{A}^c_{\gamma_1}) + \mathbb{P}(\mathfrak{P}^c) + Q\big(2e^{-u\wedge u^{\text{col}}} + Q(p_3(\tau_1) + p_3'(\tau_1^{\text{col}})\big) \end{aligned} \tag{73}$$

where γ_3 is the rate given in (71) and the second inequality uses (68).

C.2. Choosing the parameters

It remains to choose the parameters, τ_1, τ_2, ξ, etc. to simultaneously achieve the desired rate for γ_3 and ensure that the probability in (73) is $o(1)$.

Proof of Theorem 3. **First row LR classifier.** Let us write $\tau_1^{\text{row}} = \tau_1$ for clarity. Under our assumptions, we will have $\gamma_2 \le \gamma_1 \le 1/2$ so that $C_{\gamma_i} \le 2\beta^2\gamma_i$ for $i = 1, 2$, recalling the definition of $C_\gamma = \beta^2\gamma/(1-\gamma)$. In Lemma 11, we defined (recall (61))

$$\eta^{\text{step}\,7} = 8\delta_1 \omega L \|\Lambda/Q\|_\infty \;\le\; \big(384\beta^2\gamma_1 + 48\tau_1^{\text{row}} + 8\beta L\xi\big)\,\omega L\|\Lambda/Q\|_\infty. \tag{74}$$

By (40), $384\beta^2\gamma_1\omega L\|\Lambda/Q\|_\infty \le I_{\min}/(8Q)$. Take

$$\tau_1^{\text{row}} = \frac{I_{\min}}{384\omega L\|\Lambda\|_\infty}, \quad \xi = \frac{I_{\min} \wedge I_{\min}^{\text{col}}}{64\beta\omega(K \vee L)^2(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty)}, \quad u = \frac{I_{\min}}{8Q}, \tag{75}$$

where u is the parameter in (62). Then from (74) we have

$$\eta^{\text{step}\,7} \le \frac{I_{\min}}{8Q} + \frac{I_{\min}}{8Q} + \frac{I_{\min}}{8Q} = \frac{3I_{\min}}{8Q}, \quad \eta^{\text{step}\,7} + u \le \frac{I_{\min}}{2Q}.$$

Hence Lemma 11 implies that on event $\mathfrak{P}$,

$$\text{Mis}_k\left(\tilde{y}^{(q)}_{\text{step}\,7}, y^{(q)}\right) \;\le\; \gamma_{2k}^{\text{row}} := \sum_{r \ne k} \exp\Big(-\frac{I_{kr}}{Q} + \frac{I_{\min}}{2Q}\Big) \le K\exp\Big(-\frac{I_{\min}}{2Q}\Big), \quad \forall q \in \mathbb{Z}_Q \tag{76}$$

fails with probability at most $Q(e^{-u} + 2Q\,p_3(\tau_1^{\text{row}}))$. By (A4$'$), $Q\log Q = o(I_{\min})$, hence $Qe^{-u} = o(1)$. By (A3),

$$\begin{aligned} Q^2 p_3(\tau_1^{\text{row}}) &= Q^2 KL\exp\Big(-\frac{n\Lambda_{\min}\,h_1(\tau_1^{\text{row}})}{32Q^2\beta K}\Big) \\ &\le Q^2 KL\exp\Big(-\frac{nI_{\min}^2}{256(384^2)Q^2\beta KL^2\omega^3\|\Lambda\|_\infty}\Big) = o(1) \end{aligned} \tag{77}$$

where we have used the definition (58) of p_3, $h_1(\tau) \ge \tau^2/8$ for $\tau \le 1$ and $\|\Lambda\|_\infty/\Lambda_{\min} \le \omega$. Moreover, (A4′) implies $(I_{\min} \wedge I_{\min}^{\text{col}})/(K \vee L) \to \infty$, hence eventually $(I_{\min} \wedge I_{\min}^{\text{col}})/(K \vee L) \ge 1$ which gives

$$\begin{aligned}\mathbb{P}(\mathfrak{P}^c) = p_2(\xi) &= 2Q(K+L)\exp\big(-(n\wedge m)\xi^2/Q\big)\\ &\le 2Q(K+L)\exp\Big(-\frac{n\wedge m}{64^2 Q\beta^2\omega^2(K\vee L)^2(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty)^2}\Big) = o(1)\end{aligned} \tag{78}$$

where the last implication follows from (A3).

First column LR classifier. We can apply a similar argument to $\tilde{z}_{\text{step}\,8}$. Let

$$\tau_1^{\text{col}} = \frac{I_{\min}^{\text{col}}}{768\omega K\|\Gamma\|_\infty}, \quad u^{\text{col}} = \frac{I_{\min}^{\text{col}}}{16Q},$$

with ξ defined as in (75). By (40), and a similar argument, we obtain $\eta^{\text{step}\,8} + u^{\text{col}} \le I_{\min}^{\text{col}}/(4Q)$. By Corollary 4, on event $\mathfrak{P}$,

$$\text{Mis}_\ell\big(\tilde{z}_{\text{step}\,8}^{(q)}, z^{(q)}\big) \;\le\; \gamma_{2\ell}^{\text{col}} \;\le\; \sum_{r\neq\ell}\exp\Big(-\frac{I_{\ell r}^{\text{col}}}{2Q} + \frac{I_{\min}^{\text{col}}}{4Q}\Big) \le L\exp\Big(-\frac{I_{\min}^{\text{col}}}{4Q}\Big), \quad \forall q \in \mathbb{Z}_Q, \tag{79}$$

fails with probability at most $Q(e^{-u^{\text{col}}} + 2Q\,p_3')$, where $Q^2p_3' = Q^2p_3'(\tau_1^{\text{col}}) = o(1)$ by (A3) and $Qe^{-u^{\text{col}}} = o(1)$ by (A4′), similar to how we argued for the row labels.

Second row LR classifier. Recalling γ_2 from (66) and combining with (76) and (79),

$$\gamma_2 \;\le\; \max\Big(\beta K^2\exp\Big(-\frac{I_{\min}}{2Q}\Big), \beta L^2\exp\Big(-\frac{I_{\min}^{\text{col}}}{4Q}\Big)\Big) = o\Big(\frac{\beta(K\vee L)^2}{(I_{\min}\wedge I_{\min}^{\text{col}})^b}\Big) \tag{80}$$

for any $b > 0$, as $I_{\min} \wedge I_{\min}^{\text{col}} \to \infty$. By Lemma 13 and (51),

$$\begin{aligned}\eta_{kr}^{\text{step}\,11} &:= \eta_{kr}\big(\delta_2;\ \omega, \beta, m, \Lambda\big)\\ &= 21\delta_2\,\omega L\|\Lambda\|_\infty + \frac{5\beta L^2\|\Lambda\|_\infty^2}{m} + \log\Big(11\omega\Big(\frac{1}{\varepsilon_{kr} - 2\omega(1+\varepsilon_{kr})\,\delta_2} + 1\Big)\Big) - \frac{1}{2}\log\Lambda_{\min}\\ &=: T_1 + T_2 + T_3 + T_4,\end{aligned} \tag{81}$$

where we have called the four summands above $T_1, \dots, T_4$ in the order they appear. We have $\delta_2 = 24C_{\gamma_2} + 6\tau_2 \le 48\beta^2\gamma_2 + 6\tau_2$ by (69) and the assumption $\gamma_2 \le \frac{1}{2}$. Then,

$$T_1 \;\le\; 21(48)\beta^2\omega L\|\Lambda\|_\infty\,\gamma_2 \;+\; 21(6)\omega L\|\Lambda\|_\infty\,\tau_2 \;=:\; T_{11} + T_{12}.$$

For any $b > 0$, by (80)

$$T_{11} = O\big(\beta^2\omega L\|\Lambda\|_\infty\gamma_2\big) = o\Big(\frac{\beta^3\omega(K\vee L)^3\|\Lambda\|_\infty}{[(I_{\min}\wedge I_{\min}^{\text{col}})/Q]^b}\Big). \tag{82}$$

Recall that we have $\beta\omega(K\vee L)\|\Lambda\|_\infty = o([(I_{\min}\wedge I_{\min}^{\text{col}})/Q]^a)$ for some $a > 0$ by (A4′). Taking $b = 3a$ in (82), we obtain $T_{11} = o(1)$. Letting $\tau_2 = (\omega L\|\Lambda\|_\infty)^{-1}$, we have $T_{12} = O(1)$, hence, $T_1 = O(1)$. Recalling the probability bound in Lemma 13, we have by (A3)

$$\begin{aligned}Qp_3(\tau_2) &= QKL\exp\Big(-\frac{n\Lambda_{\min}\,h_1(\tau_2)}{32Q^2\beta K}\Big)\\ &\le QKL\exp\Big(-\frac{n}{256Q^2\beta KL^2\omega^3\|\Lambda\|_\infty}\Big) = o(1)\end{aligned} \tag{83}$$

where we have used $h_1(\tau) \geq \tau^2/8$ for $\tau \leq 1$ and $\|\Lambda\|_\infty/\Lambda_{\min} \leq \omega$. Using (A3) again, $T_2 = 5\beta L^2 \|\Lambda\|_\infty^2/m = O(1)$.

Now let us consider the third piece T_3 in (81). Recall that $J_{kr} = L\|\Lambda\|_\infty/I_{kr}$. By Lemma 20 in Appendix F.3.1, $\varepsilon_{kr} \geq 2\big(J_{kr}^{-1} \wedge 1\big)$. In bounding T_1, we have shown $\delta_2 \omega L \|\Lambda\|_\infty = O(1)$, hence $2\omega\delta_2 = O((L\|\Lambda\|_\infty)^{-1})$. Since $I_{kr} \to \infty$ and $J_{kr} \geq 1/2$ (see Remark 3), $(L\|\Lambda\|_\infty)^{-1} = o\big(J_{kr}^{-1} \wedge 2\big)$. Therefore, $2\omega\delta_2 = o(\varepsilon_{kr} \wedge 1)$. As a result, $2\omega(1+\varepsilon_{kr})\delta_2 = o(\varepsilon_{kr})$, hence

$$e^{T_3} := 11\omega\Big(1 + \frac{1}{\varepsilon_{kr} - 2\omega(1+\varepsilon_{kr})\delta_2}\Big) = O\Big(\omega\Big(1 + \frac{1}{\varepsilon_{kr}}\Big)\Big). \tag{84}$$

Finally, we let $v = \sqrt{\log \Lambda_{\min}}$. Since $\Lambda_{\min} \to \infty$, $e^{-v} = o(1)$. Applying Lemma 13, combined with $T_1 + T_2 = O(1)$, and (84), then for $\zeta = 1/\sqrt{\log \Lambda_{\min}} = o(1)$,

$$\begin{aligned}\mathrm{Mis}_k\left(\widehat{y}_{\mathrm{top}},\, y_{\mathrm{top}}\right) &= O\Big(\omega \sum_{r \neq k} \Big(1 + \frac{1}{\varepsilon_{kr}}\Big) \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min} + \sqrt{\log \Lambda_{\min}}\Big)\Big) \\ &= O\Big(\omega \sum_{r \neq k} \Big(1 + \frac{1}{\varepsilon_{kr}}\Big) \exp\Big(-I_{kr} - \Big(\frac{1}{2} - \zeta\Big) \log \Lambda_{\min}\Big)\Big)\end{aligned}$$

fails w.p. $\leq 2(e^{-v} + 2Q\, p_3(\tau_2)) + \mathbb{P}(\mathfrak{P}^c) + Q(2e^{-u \wedge u^{\mathrm{col}}} + Q(p_3(\tau_1^{\mathrm{row}}) + p_3'(\tau_1^{\mathrm{col}}))) = o(1)$. When we swap A_{top} and A_{bottom} and repeat the algorithm, the same misclassification rate holds. The proof of Theorem 3 is complete. □

Appendix D. Proof of Other Main Results

D.1. Proof of Theorem 1

We proceed by stating a few lemmas. The proofs are deferred to Appendix G.2.

Lemma 14. *$\sum_{\ell \in [L]} (\lambda_{r\ell} - \lambda_{k\ell})^2 \geq 2\Lambda_{\min} I_{kr}$. As a consequence, $\Lambda_\wedge^2 \geq 2\Lambda_{\min} I_{min}$.*

Combining Lemma 14 with Theorem 4, and noting that $\|\Lambda\|_\infty/\Lambda_\wedge^2 \leq \omega/(2I_{\min})$ as a consequence of the lemma, we obtain the following guarantee for spectral clustering in terms of the information matrix (I_{kr}):

Corollary 5. *Consider the spectral algorithm given in Algorithm 4, assume that for a sufficiently small $C_1 > 0$,*

$$\frac{\beta^2 \omega K L (K \wedge L)\, \alpha}{2 I_{min}} \leq C_1 (1+\kappa)^{-2}. \tag{85}$$

Then the algorithm outputs estimated row labels $\tilde{y}$ satisfying w.h.p.

$$\mathrm{Mis}(\tilde{y}, y) \;\leq\; \frac{(1+\kappa)^2 \omega \beta L (K \wedge L) \alpha}{2 C_1 I_{min}}.$$

We next modify Corollary 5 to be applicable on sub-blocks:

Lemma 15 (Spectral clustering on subblocks)**.** *Suppose* (A3) *holds, and we assume for a sufficiently small $C_1 > 0$,*

$$\frac{6Q\beta^2 \omega^2 K L (K \wedge L)\, \alpha}{I_{min}} \leq C_1 (1+\kappa)^{-2}. \tag{86}$$

Using Algorithm 4 in Step 3 of Algorithm 3, w.h.p., the misclassification rate of $\tilde{y}^{(q)}$ satisfies

$$\operatorname{Mis}(\tilde{y}^{(q)}, y^{(q)}) \leq \frac{3Q(1+\kappa)^2\omega^2\beta L(K \wedge L)\alpha}{C_1 I_{min}} \quad \forall q \in [Q].$$

A similar result holds for misclassification rate of the spectral clustering for column labels, with appropriate modifications.

of Theorem 1. Assumption (A4) implies (86), eventually as $I_{\min} \to \infty$. Letting γ_1^{row} and γ_1^{col} be bounds on the misclassification rates of the spectral clustering algorithms in steps 3 and 4, we can take, by Lemma 15 (and its column counterpart), w.h.p.

$$\gamma_1^{\text{row}} = O\Big(\frac{Q\omega\beta L(K \wedge L)\alpha}{I_{\min}}\Big), \quad \gamma_1^{\text{col}} = O\Big(\frac{Q\omega\beta K(K \wedge L)\alpha^{-1}}{I_{\min}^{\text{col}}}\Big).$$

That is, by the end of step 4, w.h.p., $\operatorname{Mis}(\tilde{y}^{(q)}, y^{(q)}) \leq \gamma_1^{\text{row}}$ and $\operatorname{Mis}(\tilde{z}^{(q)}, z^{(q)}) \leq \gamma_1^{\text{col}}$ for all $q \in [Q]$. Since the matching step increases the misclassification rate by at most a factor of 2, the same bounds hold for the overall initial labels at step 6. Taking $\gamma_1 = \gamma_1^{\text{row}} \vee \gamma_1^{\text{col}}$, we observe that in order to satisfy condition (40) of Theorem 3, it is enough to have

$$\frac{Q\omega\beta(K \vee L)^2(\alpha \vee \alpha^{-1})}{I_{\min} \wedge I_{\min}^{\text{col}}} = o\Big(\frac{1}{\beta^2\omega}\frac{I_{\min}}{L\|\Lambda\|_\infty} \wedge \frac{I_{\min}^{\text{col}}}{K\|\Gamma\|_\infty}\Big)$$

which holds if we require the stronger condition

$$\frac{Q\omega\beta(K \vee L)^2(\alpha \vee \alpha^{-1})}{I_{\min} \wedge I_{\min}^{\text{col}}} = o\Big(\frac{1}{\beta^2\omega}\frac{I_{\min} \wedge I_{\min}^{\text{col}}}{(K \vee L)(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty)}\Big).$$

But this latter condition is satisfied by assumption (A4). Thus, the assumptions of Theorem 3 hold with high probability, and so is its result. The proof is complete. □

D.2. Proof of Corollary 1

Proof of Corollary 1. From the proof of Theorem 1, we have that

$$\operatorname{Mis}_k\,(\widehat{y},\, y) = O\Big(\sum_{r \neq k} \omega\Big(1 + \frac{1}{\varepsilon_{kr}}\Big)\exp\Big(-I_{kr} - \frac{1}{2}\log\Lambda_{\min} + v\Big)\Big) \tag{87}$$

fails with probability at most $2e^{-v} + o(1)$. First, we show that

$$\chi_r := \omega\Big(1 + \frac{1}{\varepsilon_{kr}}\Big)\Lambda_{\min}^{-1/2} = o(1), \quad \text{uniformly in } r. \tag{88}$$

By Lemma 20 in Appendix F.3.1, $\varepsilon_{kr} \geq 2\big(J_{kr}^{-1} \wedge 1\big)$. Hence,

$$1 + \frac{1}{\varepsilon_{kr}} \;\leq\; 1 + \frac{1}{2}(J_{kr} \vee 1) \;\leq\; \frac{3}{2}(J_{kr} \vee 1) \;\leq\; 3J_{kr}$$

using $2J_{kr} \geq 1$ (see Remark 3). Thus, to show (88), it is enough to show $\omega J_{kr}/\sqrt{\Lambda_{\min}} = o(1)$. Using $\omega^{-1}\|\Lambda\|_\infty \leq \Lambda_{\min}$, we have

$$\omega J_{kr}\,\Lambda_{\min}^{-1/2} = \frac{\|\Lambda\|_\infty}{I_{kr}} L\,\omega\Lambda_{\min}^{-1/2} \leq \frac{L\omega^{3/2}\|\Lambda\|_\infty^{1/2}}{I_{kr}} \leq \frac{L\omega^{3/2}\|\Lambda\|_\infty^{1/2}}{I_{\min}} = o(1)$$

where the last equality is by $\omega^3 L^2 \|\Lambda\|_\infty = o(I_{\min}^2)$ which is implied by (A4). Thus, we have (88), i.e., $\chi := \max_r \chi_r = o(1)$, as desired. Now, let $2v = -\log \chi$. It follows that $e^{-v} = \sqrt{\chi} = o(1)$, and we have

$$\mathrm{Mis}_k\,(\widehat{y},\, y) = O\Big(\chi \sum_{r \neq k} \exp\big(-I_{kr} + v\big)\Big) = O\Big(\sqrt{\chi} \sum_{r \neq k} \exp\big(-I_{kr}\big)\Big) = o\Big(\sum_{r \neq k} \exp\big(-I_{kr}\big)\Big)$$

completing the proof. □

D.3. Proof of Example 1

Without loss of generality assume $a > b$ so that $\varepsilon_{kr} = a/b - 1$. Also, $\Lambda_{\min} \geq b/(\beta K)$. By (87), which holds in the general case, we have that

$$\begin{aligned}\mathrm{Mis}_k\,(\widehat{y},\, y) &= O\Big(\sum_{r \neq k} \omega\Big(\frac{b}{a}\Big) \exp\Big(-I_{kr} - \frac{1}{2}\log\Big(\frac{b}{\beta K}\Big) + v\Big)\Big) \\ &= O\Big(\sqrt{\beta}\omega K^{3/2} b^{-1/2} \exp\Big(-\frac{(\sqrt{a}-\sqrt{b})^2}{\beta K} + v\Big)\Big)\end{aligned}$$

fails with probability at most $2e^{-v} + o(1)$. Assumption $\beta\omega^2 K^3 = o(b)$ implies

$$\chi := \sqrt{\beta}\omega K^{3/2} b^{-1/2} = o(1).$$

Letting $2v = -\log \chi$, the rest of the proof follows similar to that of Corollary 1.

Appendix E. Proof of Theorem 2

The distribution of A depends on (y, z, P), so that the expectation in (16) is, in fact, $\mathbb{E} = \mathbb{E}_{(y,z,P)}$. Throughout the proof, we restrict the parameter space to a fixed z^* and P^* such that the corresponding row mean matrix, Λ^*, satisfies $I_{\min}(\Lambda^*) = I^*$. From now on, instead of writing $\mathbb{P}_{(y,z^*,P^*)}$ we simply write $\mathbb{P}_y$ for the distribution on A, and similarly for the expectations.

Let us write the optimal sum of errors in the test of $\mathrm{Poi}(\Lambda_{k*}^*)$ against $\mathrm{Poi}(\Lambda_{r*}^*)$ as follows:

$$\begin{aligned}P_{\mathrm{Err},+}^* &:= \inf\big[\mathbb{P}(\text{Type I error}) + \mathbb{P}(\text{Type II error})\big] \\ &\geq \exp\big(-I^* - \frac{L}{2}(\log \Lambda_{\min}^* + C')\big),\end{aligned}$$

where the inequality is by combining Proposition 1 and Lemma 26 (Appendix G.7), and noting that the assumption of that lemma is satisfied due to (A3). Note that $\log \Lambda_{\min}^*$ is not determined in $\mathcal{S}$; however, we can always replace it by $2 \log I^*$ by (A4).

The reduced parameter space is the set of all (y, z^*, P^*) such that y belongs to

$$\mathcal{T} := \big\{y \in \{0,1\}^{n \times K} :\ y \ \text{ satisfy (A2).}\big\}$$

Let $\alpha := \big(8\beta K \vee \frac{\beta}{\beta - 1}\big)$ and choose $S \subset [n]$ and $\tilde{y} \in \mathcal{T}$ such that

$$\big|\big\{i \in S^c : \tilde{y}_i = k\big\}\big| = n_0 := \big\lceil \frac{n}{K}\big(1 - \frac{1}{\alpha}\big)\big\rceil, \quad \forall k \in [K].$$

We then define the further restricted parameter space

$$\mathcal{T}' := \big\{ y = (y_S, \tilde{y}_{S^c}) : \; y_S \in \{k, r\}^{|S|} \big\}.$$

Since $1 - \frac{1}{\alpha} \geq \frac{1}{\beta}$, each label in $\mathcal{T}'$ has at least $n/\beta K$ labels in each community, hence $\mathcal{T}' \subset \mathcal{T}$. The direct misclassification rate, dMis (cf. Section 2.3), between any two labels in $\mathcal{T}'$ is at most

$$\varepsilon := \frac{|S|}{n} = 1 - \frac{|S^c|}{n} = 1 - \frac{Kn_0}{n} \in \Big[\frac{1}{2\alpha}, \frac{1}{\alpha}\Big]$$

using $x \leq \lceil x \rceil \leq x + 1$, and $K/n \leq 1/(2\alpha)$, which holds for large n, for the lower bound.

In particular, the dMis between any two elements in $\mathcal{T}'$ is at most $1/(8\beta K)$ (i.e., at most $n/(8\beta K)$ labels are different). It follows from Lemma 8 that the optimal matching permutation between any two label vectors in $\mathcal{T}'$ is identity, hence their misclassification rate is the same as their dMis.

We next argue that $\widehat{y}$ can be restricted to $\mathcal{T}'$ as well: First suppose that under the optimal permutation π, $\widehat{y}$ has at most εn different labels on indices S^c compared to $\tilde{y}$, that is, $|\{i \in S^c : \pi(\widehat{y}_i) \neq \tilde{y}_i\}| \leq \varepsilon n$. It follows that the misclassification rate between $\widehat{y}$ and any $y \in \mathcal{T}'$ is at most $2\varepsilon \leq 1/(4\beta K)$. Thus, by Lemma 8, π is the optimal permutation for matching $\widehat{y}$ to any $y \in \mathcal{T}'$. Redefining $\widehat{y}_i := \pi^{-1}(\tilde{y}_i)$ for $i \in S^c$ then gives a uniformly better strategy over $\mathcal{T}'$. The new $\widehat{y}$ equals $\tilde{y}$ on S^c up to a permutation, so we can restrict $\widehat{y}$ to $\mathcal{T}'$. On the other hand, if $|\{i \in S^c : \pi(\widehat{y}_i) \neq \tilde{y}_i\}| > \varepsilon n$, then setting $\widehat{y}$ to be any fixed vector from $\mathcal{T}'$ (or randomly choosing from $\mathcal{T}'$) gives a better strategy.

The minimax risk is lower-bounded by

$$\begin{aligned}
\inf_{\widehat{y}} \sup_{y \in \mathcal{T}'} \mathbb{E}_y[\,\mathrm{Mis}(\widehat{y}, y)\,] &= \inf_{\widehat{y} \in \mathcal{T}'} \sup_{y \in \mathcal{T}'} \mathbb{E}_y[\,\mathrm{dMis}(\widehat{y}, y)\,] \\
&= \inf_{\widehat{y} \in \mathcal{T}'} \sup_{y \in \mathcal{T}'} \frac{1}{n} \sum_{i \in S} \mathbb{P}_y(\widehat{y}_i \neq y_i) \\
&\geq \varepsilon \cdot \inf_{\widehat{y} \in \mathcal{T}'} \operatorname*{avg}_{y \in \mathcal{T}'} \frac{1}{|S|} \sum_{i \in S} \mathbb{P}_y\big(\widehat{y}_i \neq y_i\big) \\
&= \varepsilon \cdot \inf_{\widehat{y} \in \mathcal{T}'} \operatorname*{avg}_{i \in S} \operatorname*{avg}_{y \in \mathcal{T}'} \mathbb{P}_y\big(\widehat{y}_i \neq y_i\big)
\end{aligned} \tag{89}$$

where we have used $n = |S|/\varepsilon$ and max $\geq$ avg. Let us now focus on a single term in the sum over S, say $i = 1$. For simplicity, let $S \setminus 1 = S \setminus \{1\}$. Let $\mathcal{T}'_u = \{(u, y_{S\setminus 1}, \tilde{y}_{S^c}) : \; y_{S\setminus 1} \in \{k, r\}^{|S|-1}\}$. Then, $\mathcal{T}'$ is the disjoint union of $\mathcal{T}'_k$ and $\mathcal{T}'_r$ and we have

$$\begin{aligned}
\operatorname*{avg}_{y \in \mathcal{T}'} \mathbb{P}_y(\widehat{y}_1 \neq y_1) &= \frac{1}{|\mathcal{T}'|} \sum_{y \in \mathcal{T}'} \mathbb{P}_y(\widehat{y}_1 \neq y_1) \\
&= \frac{1}{|\mathcal{T}'|} \sum_{y_{S\setminus 1}} \Big[\mathbb{P}_{(k,\, y_{S\setminus 1},\, \tilde{y}_{S^c})}\big(\widehat{y}_1 \neq k\big) + \mathbb{P}_{(r,\, y_{S\setminus 1},\, \tilde{y}_{S^c})}\big(\widehat{y}_1 \neq r\big) \Big] \\
&\geq \frac{1}{|\mathcal{T}'|} \sum_{y_{S\setminus 1}} P^*_{\mathrm{Err},+} = \frac{1}{2} P^*_{\mathrm{Err},+}
\end{aligned}$$

where the second equality is by decomposing the sum as $\sum_{y_{S\setminus 1}} \sum_{y_1}$, and the last equality by noting that the sum over $y_{S\setminus 1}$ is over $\{k, r\}^{|S|-1}$ whose cardinality is $|\mathcal{T}'|/2$. The same lower bound holds for all other $i \neq 1$ in (89) by symmetry. Hence, we conclude

$$\inf_{\widehat{y}} \sup_{y \in \mathcal{T}'} \mathbb{E}_y[\mathrm{Mis}(\widehat{y}, y)] \geq \frac{\varepsilon}{2} P^*_{\mathrm{Err},+} \geq \frac{1}{4\alpha} P^*_{\mathrm{Err},+}.$$

Recalling the definition of α and using the assumptions that $\beta > 1$ is constant and $K \leq \exp(cL)$ gives the desired result.

Appendix F. Proofs of the main lemmas

In this section, we give the proof of the three main lemmas of Appendix B.1. We we first give the proofs of Lemma 1 and 3 in Appendix F.1 and F.2. The proof of Lemma 2(b) is more technical and occupies the remainder of this section, including auxiliary results on the error exponents and Poisson-binomial approximations, in Appendix F.3 and F.4.

Throughout, we will use the following concentration inequality (Giné and Nickl, 2015, p. 118):

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

$$\mathbb{P}\big(S \geq vt\big) \leq \exp[-vh_c(t)], \quad t \geq 0, \quad \textit{where } h_c(t) := \frac{3}{4c} t \log\big(1 + \frac{2c}{3}t\big). \tag{90}$$

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

Note that $h_c(t) \asymp t^2$ as $t \to 0$ and $h_c(t) \asymp t \log t$ as $t \to \infty$.

F.1. Proof of Lemma 1

Let us define the confusion matrix as $R(\tilde{z}, z) \in [0,1]^{L \times L}$ with entries

$$R_{k\ell}(\tilde{z}, z) = \frac{1}{m} \sum_{j=1}^{m} 1\{\tilde{z}_j = k, z_j = \ell\} = \frac{|j : \tilde{z}_j = k, z_j = \ell|}{m}. \tag{91}$$

We can similarly define $R_{k\ell}(z, \tilde{z})$. It is easy verify that $R(\tilde{z}, z) = R(z, \tilde{z})^T$. By definition (23) of the (global) row mean parameters,

$$\lambda_{k\ell'}(y, \tilde{z}) = \sum_{j=1}^{m} \sum_{\ell=1}^{L} P_{k\ell} 1\{z_j = \ell, \tilde{z}_j = \ell'\} = m P_{k*} R_{*\ell'}(z, \tilde{z}). \tag{92}$$

To see (92), note that since we are using true labels y in the first argument of $\lambda_{k\ell'}(y, \tilde{z})$, the averaging $\frac{1}{n_k(y)} \sum_i 1\{y_i = k\}\big(\cdots\big)$ over i, in the definition, is vacuous. That is, for any i with $y_i = k$, we have $\lambda_{k\ell'}(y, \tilde{z}) = \sum_j \mathbb{E}[A_{ij}] 1\{\tilde{z}_j = \ell'\}$. We then further break this sum according to column labels $z_j = \ell$ to get (92).

Recall that $n(z)$ is the vector of sizes of clusters in z and $\pi(z) = n(z)/m$ is the corresponding proportions. To simplify, let

$$N(z) := \mathrm{diag}(n(z)), \quad \Pi(z) := \mathrm{diag}(\pi(z)).$$

We have $mI_L = N(z)\Pi(z)^{-1}$ where I_L is the $L \times L$ identity matrix, hence

$$mP_{k*}R_{*\ell'}(z,\tilde{z}) = P_{k*}\,N(z)\,\Pi(z)^{-1}R_{*\ell'}(z,\tilde{z}) = \lambda_{k*}(y,z)\,\Pi(z)^{-1}R_{*\ell'}(z,\tilde{z})$$

using (2). Let use define

$$U(z,\tilde{z}) := \Pi(z)^{-1}R(z,\tilde{z}).$$

Since $\pi(z)$ contains the row sums of $R_{*\ell'}(z,\tilde{z})$, $U(z,\tilde{z})$ is the row-normalized confusion matrix, i.e. $U = (R_{k\ell}/R_{k+})$. We have

$$\lambda_{k\ell'}(y,\tilde{z}) = \lambda_{k*}(y,z)\,U_{*\ell'}(z,\tilde{z}), \tag{93}$$

and its matrix version $\Lambda(y,\tilde{z}) = \Lambda(y,z)\,U(z,\tilde{z})$. We can similarly define $U(\tilde{y},y) = \Pi(\tilde{y})^{-1}R(\tilde{y},y)$. Recalling definition (23), and some algebra gives

$$\begin{aligned}\lambda_{k'\ell'}(\tilde{y},\tilde{z}) &= \frac{1}{n_{k'}(\tilde{y})}\sum_{i=1}^{n}\sum_{k\in[K]}\lambda_{k\ell'}(y,\tilde{z})1\{y_i = k, \tilde{y}_i = k'\} \\ &= \frac{1}{n_{k'}(\tilde{y})}\sum_{k\in[K]}\lambda_{k\ell'}(y,\tilde{z})\,|i : y_i = k, \tilde{y}_i = k'|,\end{aligned}$$

where to get the first equality one further breaks the sums over $\sum_k 1\{y_i = k\}$ and use the expression for $\lambda_{k\ell'}(y,\tilde{z})$ in the comments after (92). Using the definition of the confusion matrix in (91), adapted to row labels, and the definition of U, we have

$$\lambda_{k'\ell'}(\tilde{y},\tilde{z}) = \frac{1}{\pi_{k'}(\tilde{y})}R_{k'*}(\tilde{y},y)\,\lambda_{*\ell'}(y,\tilde{z}) = U_{k'*}(\tilde{y},y)\,\lambda_{*\ell'}(y,\tilde{z}), \tag{94}$$

or compactly $\Lambda(\tilde{y},\tilde{z}) = U(\tilde{y},y)\Lambda(y,\tilde{z})$. We also define a column-normalized confusion matrix,

$$V(z,\tilde{z}) := R(z,\tilde{z})\Pi(\tilde{z})^{-1}.$$

Lemma 16. *(A2) and (B3) imply*

$$\max_k\,[1 - U_{kk}(\tilde{y},y)] \;\le\; \gamma, \tag{B3.1}$$

$$\max_\ell\,[1 - U_{\ell\ell}(z,\tilde{z})] \;\le\; \gamma, \quad \textit{and} \tag{B3.2}$$

$$\max_\ell\,[1 - V_{\ell\ell}(z,\tilde{z})] \;\le\; \gamma. \tag{B3.3}$$

Proof. Without loss of generality, assume that the optimal permutation matching $\tilde{y}$ to y is identity, and similarly for $\tilde{z}$ to z. By definition, $1 - U_{kk}(\tilde{y},y)$ is the misclassification rate withing the kth community of $\tilde{y}$, hence

$$1 - U_{kk}(\tilde{y},y) = \frac{|i : \tilde{y}_i = k,\, y_i \ne k|/n}{|i : \tilde{y}_i = k|/n} = \frac{|i : \tilde{y}_i = k,\, y_i \ne k|/n}{|i : \tilde{y}_i = k,\, y_i \ne k|/n + |i : \tilde{y}_i = y_i = k|/n}.$$

Recall that we can write (see Section 2.3)

$$\mathrm{Mis}(\tilde{y},y) = \frac{1}{n}|i : \tilde{y}_i \ne y_i| = \frac{1}{n}\sum_k |i : \tilde{y}_i = k,\, y_i \ne k| = \frac{1}{n}\sum_k |i : \tilde{y}_i \ne k,\, y_i = k|. \tag{95}$$

Then, (B3) and the second equality in (95) implies $|i : \tilde{y}_i = k, y_i \neq k|/n \leq \gamma/(\beta K)$, while the third equality in (95) gives $|i : \tilde{y}_i = y_i = k|/n \geq \pi_k(y) - \gamma/(\beta K)$. Letting $f(x) = x/(x+1)$,

$$1 - U_{kk}(\tilde{y}, y) = f\Big(\frac{|i : \tilde{y}_i = k,\, y_i \neq k|/n}{|i : \tilde{y}_i = y_i = k|/n}\Big) \leq \frac{\gamma/(\beta K)}{\gamma/(\beta K) + \pi_k(y) - \gamma/(\beta K)} = \frac{\gamma}{\pi_k(y)\beta K} \leq \gamma$$

where the first inequality is by monotonicity of f, and the last by (A2). This proves (B3.1).

Similarly, $1 - U_{\ell\ell}(z, \tilde{z})$ is the misclassification rate within the ℓth community of z, i.e., $\mathrm{Mis}_\ell(\tilde{z}, z)$, hence

$$1 - U_{\ell\ell}(z, \tilde{z}) = \frac{|j : z_j = \ell, \tilde{z}_j \neq \ell|/m}{\pi_\ell(z)} = \frac{\gamma}{\pi_\ell(z)\beta L} \leq \gamma,$$

proving (B3.2). The same bound holds for $1 - U_{\ell\ell}(\tilde{z}, z)$ by an argument similar to that used for $U_{kk}(\tilde{y}, y)$. To prove (B3.3), we observe

$$U(\tilde{z}, z) = \Pi(\tilde{z})^{-1}R(\tilde{z}, z) = \Pi(\tilde{z})^{-1}R(z, \tilde{z})^T = [R(z, \tilde{z})\Pi(\tilde{z})^{-1}]^T = V(z, \tilde{z})^T$$

hence $1 - V_{\ell\ell}(z, \tilde{z}) = 1 - U_{\ell\ell}(\tilde{z}, z) \leq \gamma$. All statements are true for any $k \in [K]$ and $\ell \in [L]$. □

F.1.1. Proof of Lemma 1(a)

For the lower bound, by (93) and (B3.2),

$$\begin{aligned}\lambda_{k\ell'}(y, \tilde{z}) = \lambda_{k*}(y, z)U_{*\ell'}(z, \tilde{z}) \;&\geq\; \lambda_{k\ell'}(y, z)U_{\ell'\ell'}(z, \tilde{z}) \\ &\geq\; (1-\gamma)\lambda_{k\ell'}(y, z) \geq \lambda_{k\ell'}(y, z) - C_\gamma\|\Lambda\|_\infty\end{aligned}$$

where the last inequality is by $\gamma \leq C_\gamma$ and $\lambda_{k\ell'}(y, z) \leq \|\Lambda\|_\infty$. For the upper bound, we write

$$\lambda_{k*}(y, z)U_{*\ell'}(z, \tilde{z}) = \lambda_{k\ell'}(y, z)U_{\ell'\ell'}(z, \tilde{z}) + \sum_{\ell \neq \ell'} \lambda_{k\ell}(y, z)U_{\ell\ell'}(z, \tilde{z}).$$

The first term obviously satisfies $\lambda_{k\ell'}(y, z)\, U_{\ell'\ell'}(z, \tilde{z}) \leq \lambda_{k\ell'}(y, z)$, hence

$$\lambda_{k\ell'}(y, \tilde{z}) - \lambda_{k\ell'}(y, z) \leq \sum_{\ell \neq \ell'} \lambda_{k\ell}(y, z)U_{\ell\ell'}(z, \tilde{z}). \tag{96}$$

By (B3.3), for every $\ell' \in [L]$,

$$\pi_{\ell'}(z) \;\geq\; \frac{1}{m}|j : z_j = \tilde{z}_j = \ell'| \;=\; \pi_{\ell'}(\tilde{z})V_{\ell'\ell'}(z, \tilde{z}) \;\geq\; (1-\gamma)\pi_{\ell'}(\tilde{z}). \tag{97}$$

By (A2), for every ℓ' and ℓ, we have $\pi_{\ell'}(z) \leq \beta^2\pi_\ell(z)$, hence

$$U_{\ell\ell'}(z, \tilde{z}) = \frac{1}{\pi_\ell(z)}R_{\ell\ell'}(z, \tilde{z}) = \frac{\pi_{\ell'}(\tilde{z})}{\pi_\ell(z)}V_{\ell\ell'}(z, \tilde{z}) \leq \frac{\beta^2}{1-\gamma}V_{\ell\ell'}(z, \tilde{z}).$$

Combining with (96)

$$\lambda_{k\ell'}(y, \tilde{z}) - \lambda_{k\ell'}(y, z) \;\leq\; \frac{\beta^2}{1-\gamma}\sum_{\ell \neq \ell'} \lambda_{k\ell}(y, z)V_{\ell\ell'}(z, \tilde{z}) \;\leq\; \frac{\beta^2\gamma}{1-\gamma}\|\lambda_{k*}(y, z)\|_\infty \tag{98}$$

where the last inequality is by (B3.3) and that V is column normalized. This proves the upper bound, and completes the proof of $\|\Lambda(y, \tilde{z}) - \Lambda\|_\infty \leq C_\gamma\|\Lambda\|_\infty$. Since we assume $C_\gamma \leq 1$, it follows that $\|\Lambda(y, \tilde{z})\|_\infty \leq 2\|\Lambda\|_\infty$.

F.1.2. Proof of Lemma 1(b)

Recalling (94), we have

$$\lambda_{k'\ell'}(\tilde{y},\tilde{z}) \;=\; U_{k'*}(\tilde{y},y)\lambda_{*\ell'}(y,\tilde{z}) \;=\; U_{k'k'}(\tilde{y},y)\lambda_{k'\ell'}(y,\tilde{z}) + \sum_{k\neq k'} U_{k'k}(\tilde{y},y)\lambda_{k\ell'}(y,\tilde{z}).$$

By (B3.1), the first term is bounded as

$$(1-\gamma)\lambda_{k'\ell'}(y,\tilde{z}) \;\le\; U_{k'k'}(\tilde{y},y)\,\lambda_{k'\ell'}(y,\tilde{z}) \;\le\; \lambda_{k'\ell'}(y,\tilde{z})$$

and the second term as

$$0 \;\le\; \sum_{k\neq k'} U_{k'k}(\tilde{y},y)\lambda_{k\ell'}(y,\tilde{z}) \;\le\; \gamma\|\lambda_{*\ell'}(y,\tilde{z})\|_\infty$$

recalling that U is row normalized hence $\sum_{k\neq k'} U_{k'k} = 1 - U_{k'k'} \le \gamma$, by (B3.1). Combining the two bounds, we have

$$\begin{aligned}\lambda_{k'\ell'}(\tilde{y},\tilde{z}) - \lambda_{k'\ell'}(y,\tilde{z}) \;&\in\; \big[-\gamma\lambda_{k'\ell'}(y,\tilde{z}),\, 0\big] + \big[0, \gamma\|\lambda_{*\ell'}(y,\tilde{z})\|_\infty\big] \\ &\subseteq \|\lambda_{*\ell'}(y,\tilde{z})\|_\infty\big[-\gamma,\gamma\big]\end{aligned}$$

showing that $\|\Lambda(\tilde{y},\tilde{z}) - \Lambda(y,\tilde{z})\|_\infty \le \gamma\|\Lambda(y,\tilde{z})\|_\infty$. Combining with $\|\Lambda(y,\tilde{z})\|_\infty \le 2\|\Lambda\|_\infty$ from part (a) of the lemma, we have the first assertion of part (b). The second assertion follows from $\gamma \le 1/2$ and part (a) by triangle inequality. (Note that assumption $6C_\gamma\omega \le 1$ in fact implies $\gamma \le 1/6$ since $\beta,\omega \ge 1$ and $\gamma \le C_\gamma$.)

F.1.3. Proof of Lemma 1(c)

Recalling definitions of $\hat{\lambda}_{k\ell}$ and $\lambda_{k\ell}(\tilde{y},\tilde{z})$ from (22) and (23), we have

$$n_k(\tilde{y})\big[\hat{\lambda}_{k\ell} - \lambda_{k\ell}(\tilde{y},\tilde{z})\big] = \sum_{i=1}^n\sum_{j=1}^m \big(A_{ij} - \mathbb{E}[A_{ij}]\big)\,1\{\tilde{y}_i = k, \tilde{z}_j = \ell\}$$

which is of the form $S = \sum_{ij} X_{ij}$ with independent centered terms $X_{ij} = A_{ij} - \mathbb{E}[A_{ij}]$ with $|X_i| \le 1$ and $\sum_{ij}\mathbb{E}X_{ij}^2 = \sum_{ij}\operatorname{var}(A_{ij}) \le \sum_{ij}\mathbb{E}A_{ij} = n_k(\tilde{y})\lambda_{k\ell}(\tilde{y},\tilde{z})$. Note that the sums in these expressions run over $\{(i,j) : \tilde{y}_i = k,\, \tilde{z}_j = \ell\}$. Applying the two-sided version of Proposition 2, with $v = n_k(\tilde{y})\lambda_{k\ell}(\tilde{y},\tilde{z})$, $t = \tau$ and $c = 1$, we have

$$\begin{aligned}\mathbb{P}\big(\big|\hat{\lambda}_{k\ell} - \lambda_{k\ell}(\tilde{y},\tilde{z})\big| > \lambda_{k\ell}(\tilde{y},\tilde{z})\,\tau\big) &= \mathbb{P}\big(\,n_k(\tilde{y})\big|\hat{\lambda}_{k\ell} - \lambda_{k\ell}(\tilde{y},\tilde{z})\big| > n_k(\tilde{y})\lambda_{k\ell}(\tilde{y},\tilde{z})\,\tau\,\big) \\ &\le 2\exp\big(-n_k(\tilde{y})\lambda_{k\ell}(\tilde{y},\tilde{z})h_1(\tau)\,\big).\end{aligned}$$

Applying union bound over $(k,\ell) \in [K]\times[L]$, and using part (b) of this lemma, we have $\|\hat{\Lambda} - \Lambda(\tilde{y},\tilde{z})\|_\infty \le \tau\|\Lambda(\tilde{y},\tilde{z})\|_\infty \le 4\tau\|\Lambda\|_\infty$ with probability at least

$$1 - 2KL\big(-\min_k n_k(\tilde{y})\,\min_{k,\ell}\lambda_{k\ell}(\tilde{y},\tilde{z})\,h_1(\tau)\,\big).$$

We have $n_k(\tilde{y}) \ge n\pi_k(y)(1-\gamma) \ge n(\beta K)^{-1}/2$ using (B3.1), (A2) and $\gamma \le 1/2$; see (97). Similarly, since $\|\Lambda(\tilde{y},\tilde{z}) - \Lambda\|_\infty \le 3C_\gamma\|\Lambda\|_\infty$, we have

$$\min_{k,\ell}\lambda_{k\ell}(\tilde{y},\tilde{z}) \;\ge\; \Lambda_{\min} - 3C_\gamma\|\Lambda\|_\infty \;\ge\; \Lambda_{\min}(1 - 3C_\gamma\omega) \;\ge\; \Lambda_{\min}/2.$$

F.2. Proof of Lemma 3

Let $b_{i*} = b_{i*}(\tilde{z})$. Recall (46), (47) and (48), and let

$$\widehat{S}_k = S_k(\boldsymbol{b}, \hat{\Lambda}), \quad \widehat{Z}_{ik} = Z_{ik}(b_{i*}, \hat{\Lambda}), \quad \widehat{Y}_{ikr} = Z_{ik}(b_{i*}, \hat{\Lambda}).$$

For any event $\mathcal{A}$ and random variable X, let us write $\mathbb{E}[X; \mathcal{A}] := \mathbb{E}[X 1_{\mathcal{A}}]$. Consider the following event: $\mathcal{A} := \{\hat{\Lambda} \in \mathscr{B}_\Lambda(\delta)\}$. Pick some $i \in [N]$ with $y_i = k$. Then,

$$\begin{aligned}
\mathbb{E}\big[\widehat{S}_k; \mathcal{A}\big] = \mathbb{E}\big[\widehat{Z}_{ik}; \mathcal{A}\big] &= \mathbb{P}\Big(\bigcup_{r \neq k} \{\widehat{Y}_{ikr} \geq 0\} \cap \mathcal{A}\Big) \\
&\leq \sum_{r \neq k} \mathbb{P}\big(Y_{ikr}(b_{i*}(\tilde{z}), \hat{\Lambda}) \geq 0,\ \hat{\Lambda} \in \mathscr{B}_\Lambda(\delta)\big) \\
&\leq \sum_{r \neq k} \mathbb{P}\big(\exists \tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta),\ Y_{ikr}(b_{i*}(\tilde{z}), \tilde{\Lambda}) \geq 0\big) \\
&\leq \sum_{r \neq k} \exp\big(-I_{kr} + \eta'\big) =: p_k
\end{aligned}$$

where the last inequality follows from Lemma 2 with η_{kr} defined there. Using Markov inequality

$$\begin{aligned}
\mathbb{P}\big(\widehat{S}_k \geq t p_k\big) &\leq \mathbb{P}\big(\{\widehat{S}_k \geq t p_k\} \cap \mathcal{A}\big) + \mathbb{P}(\mathcal{A}^c) \\
&\leq \frac{\mathbb{E}[\widehat{S}_k; \mathcal{A}]}{t p_k} + \mathbb{P}(\mathcal{A}^c) \leq \frac{1}{t} + \mathbb{P}(\mathcal{A}^c).
\end{aligned}$$

for any $t > 0$. The version of Markov inequality used follows from (pointwise) inequality: $1_{\{X \geq u\}} 1_{\mathcal{A}} \leq (X 1_{\mathcal{A}})/u$. Taking $t = e^u$ complete the proof.

F.3. Error exponents

We start by obtaining a bound on the error exponent (i.e., the negative logarithm of the probability of error) for binary hypothesis testing in an exponential family. This result is a generalization of the result that appears in Abbe and Sandon (2015), and is proved by the same technique. The result (and the technique inspired by Abbe and Sandon (2015)) is interesting since it provides a bound different than the classical Chernoff bound on the error exponent (Chernoff, 1952); see also Verdú (1986) and Theorem 11.9.1 of Cover and Thomas (2006). This leads for example to a sharper control for the case of Poisson hypothesis testing. We start with the result for a general exponential family and then in Appendix F.3.1 specialize to the case of interest in this paper, the Poisson family.

General exponential family. Let $\pi(t; \gamma)$ denote the density of a 1-dimensional standard exponential family w.r.t. to some measure ν on $\mathbb{R}$:

$$\pi(t; \gamma) = h(t) \exp\big(\gamma t - A(\gamma)\big). \tag{99}$$

We consider distributions on $\mathbb{R}^L$ that are products of these distributions, having density:

$$p(x; \theta) = \prod_{\ell=1}^{L} \pi(x_\ell, \theta_\ell), \quad x = (x_\ell) \in \mathbb{R}^L, \quad \theta = (\theta_\ell) \in \mathbb{R}^L \tag{100}$$

with respect to $\mu = \nu^{\otimes L}$ (L-fold product measure whose coordinate measures are all ν).

Proposition 3. *Let $p_r(x) := p(x;\theta_r)$, $r = 0,1$ be two exponential family densities on $\mathbb{R}^L$ (relative to $\mu = \nu^{\otimes L}$) as defined in (99) and (100). Assume that ν is either the Lebesgue measure on $\mathbb{R}$ or the counting measure on $\mathbb{Z}$, and that $\theta_0 \neq \theta_1$. For $s \in (0,1)$, let*

$$\theta_{s\ell} = (1-s)\theta_{0\ell} + s\theta_{1\ell}, \quad \textit{and} \quad I_{s\ell} = \big[(1-s)A(\theta_{0\ell}) + sA(\theta_{1\ell})\big] - A(\theta_{s\ell}), \tag{101}$$

as well as $I_s := \sum_{\ell=1}^{L} I_{s\ell}$, $T := \{\ell : \theta_{0\ell} \neq \theta_{1\ell}\}$ and

$$C(\alpha) := \int e^{-\alpha|t|} d\nu(t) = \begin{cases} \frac{2}{\alpha} & \nu \textit{ is Lebesgue,} \\ \frac{1+e^{-\alpha}}{1-e^{-\alpha}} \leq \frac{2}{1-e^{-\alpha}} & \nu \textit{ is counting.} \end{cases} \tag{102}$$

Consider testing p_0 against p_1 using the likelihood ratio test based on a single observation. Let p_r be the probability of error under p_r for $r = 0,1$. Then, the sum of the error probabilities is bounded as

$$P_{e,0} + P_{e,1} \;\leq\; \inf_{\ell \in T} \inf_{s \in (0,1)} \Big[e^{-I_s} \|\pi(\,\cdot\,;\theta_{s\ell})\|_\infty \, C\Big(\min(s, 1-s)|\theta_{0\ell} - \theta_{1\ell}| \Big) \Big]. \tag{103}$$

Remark 7. The proof goes through for any translation invariant measure ν (e.g., a Haar measure) with an appropriate constant $C(\alpha)$. It also goes through if we replace t in (99) with a general sufficient statistic $\phi(t)$, as long as (1) ϕ is surjective from the support of the exponential family to $\mathbb{R}$ and (2) $C(\alpha) = \int e^{-\alpha|\phi(t)|} d\nu(t) < \infty$ for all $\alpha > 0$ and (3) ϕ has a measurable inverse.

Remark 8. Let s^* be the maximizer of $s \mapsto I_s$. Then, noting that $\alpha \mapsto C(\alpha)$ is decreasing, Proposition 3 implies

$$P_{e,0} + P_{e,1} \leq \exp\Big(-I_{s^*} + \log \|\pi(\,\cdot\,;\theta_{s^*\ell})\|_\infty + \log C(\alpha^*) \Big), \quad \text{where,} \tag{104}$$

$$\alpha^* = \min(s^*, 1-s^*) \max_{\ell \in [L]} |\theta_{0\ell} - \theta_{1\ell}|. \tag{105}$$

The bound is an improvement over the Chernoff bound if $\log \|\pi(\,\cdot\,;\theta_{s^*\ell})\|_\infty$ is negative and $\log C(\alpha^*)$ is controlled. This is the case for the Poisson distribution as we show in the sequel.

F.3.1. Poisson case

The Poisson case corresponds to (99) with $\gamma = \log \lambda$, $h(t) = (1/t!)1\{t \geq 0\}$, $\nu =$ the counting measure and $A(\log \lambda) = \lambda$. Letting $\theta_{s\ell} = \log \lambda_{s\ell}$ for all $s \in [0,1]$, we have from (101)

$$\lambda_{s\ell} = \lambda_{0\ell}^{1-s}\,\lambda_{1\ell}^{s}, \quad I_{s\ell} = \big[(1-s)\lambda_{0\ell} + s\lambda_{1\ell}\big] - \lambda_{s\ell}.$$

We also note that $|\theta_{0\ell} - \theta_{1\ell}| = |\log(\lambda_{0\ell}/\lambda_{1\ell})|$. Let us define

$$s^* = \operatorname*{argmax}_{s \in (0,1)} I_s, \quad \text{and,} \quad I^* = \max_{s \in (0,1)} I_s, \quad \text{where} \quad I_s = \sum_{\ell=1}^{L} I_{s\ell}$$

We will assume

$$\lambda_{0\ell}/\lambda_{1\ell} \in [1/\omega, \omega], \ \forall \ell \in [L], \quad \text{for some } \omega > 1. \tag{106}$$

The following lemma shows that s^* stays away from the boundary:

Lemma 17. *Assuming* (106)*, we have* $s^* \in [\frac{1}{2\omega}, 1 - \frac{1}{2\omega}]$*.*

Proof of this lemma and subsequent results appear in Appendix G.4. From (105), we have $\alpha^* = \min(s^*, 1-s^*)\max_\ell |\log(\lambda_{0\ell}/\lambda_{1\ell})|$ in the Poisson case. Defining

$$\varepsilon_{01} := \varepsilon_{01}(\Lambda) := \max_{\ell \in [L]} \Big(\frac{\lambda_{0\ell}}{\lambda_{1\ell}} \vee \frac{\lambda_{1\ell}}{\lambda_{0\ell}}\Big) - 1, \quad \alpha_{01} := \frac{1}{2\omega}\log(1+\varepsilon_{01}) \tag{107}$$

we note that $\alpha^* = \min(s^*, 1-s^*)\log(1+\varepsilon_{01})$, hence Lemma 17 implies $\alpha^* \geq \alpha_{01}$, that is, $C(\alpha^*) \leq C(\alpha_{01})$ in (104), where $C(\cdot)$ has the form given in (102) for the counting measure, i.e.,

$$C(\alpha_{01}) = \frac{1+e^{-\alpha_{01}}}{1-e^{-\alpha_{01}}} \leq \frac{2}{1-e^{-\alpha_{01}}} \overset{(a)}{\leq} \Big(\frac{4}{\varepsilon_{01}} + 3\Big)\omega \tag{108}$$

where the last inequality is by the following lemma:

Lemma 18. *Inequality (a) in* (108) *holds.*

Next we bound the maximum of the density:

Lemma 19 (Hodges and Le Cam (1960))**.** *Let* $\pi(t; \log\lambda) = e^{-t}(\lambda^t/t!)1\{t \geq 0\}$ *be the desnity of the Poisson family. Then, for all* $\lambda > 0$*,*

$$\|\pi(\,\cdot\,; \log\lambda)\|_\infty \leq \Big(1 + \frac{1}{12\lambda}\Big)\frac{1}{\sqrt{2\pi\lambda}}.$$

In particular $\|\pi(\,\cdot\,;\log\lambda)\|_\infty \leq \exp\big(-\frac{1}{2}\log\lambda\big)$ *for* $\lambda \geq 0.056$*.*

Combining Lemmas 17, 18 and 19, we have the following corollary which gives the following overall bound on the error exponent:

Corollary 6. *Consider testing two Poisson vector models with mean vectors given by the rows of* $\Lambda = [\lambda_{0*}\,;\,\lambda_{1*}] \in \mathbb{R}_+^{2\times L}$*, satisfying* (106)*. Let* $\Lambda_{\min} = \min_{r\ell}\lambda_{r\ell}$*. Then, the sum of the error probabilities for the likelihood ratio test is bounded as*

$$P_{e,0} + P_{e,1} \;\leq\; \omega\Big(\frac{4}{\varepsilon_{01}} + 3\Big)\exp\Big(-I^* - \frac{1}{2}\log\Lambda_{\min}\Big). \tag{109}$$

We also have the following general lower bound on ε_{01} in terms of the information I^*:

Lemma 20. *Let* $\Lambda = [\lambda_{0*}\,;\,\lambda_{1*}] \in \mathbb{R}_+^{2\times L}$*. There exists* $\ell \in [L]$ *such that*

$$\Big|\log\frac{\lambda_{0\ell}}{\lambda_{1\ell}}\Big| \geq \frac{1}{2}\log\Big(1 + \frac{8I^*}{L\|\Lambda\|_\infty}\Big),$$

which implies $\varepsilon_{01} \geq \min\big(\frac{2I^*}{L\|\Lambda\|_\infty}, 2\big)$*.*

Although the bound in Lemma 20 holds without any further assumption, it is not always tight. The difference in our two sets of results, namely (13) and (14) is due to using the sharper bound (109) versus replacing ε_{01} with its universal lower bound.

Lemma 21 (Perturbation of ε_{01})**.** *Suppose* $\Lambda' \in \mathscr{B}_\Lambda(\delta)$*, and let* $\varepsilon'_{01} = \varepsilon_{01}(\Lambda')$ *and* $\varepsilon_{01} = \varepsilon_{01}(\Lambda)$ *as in* (107)*, and assume* (106)*. Then* $\varepsilon'_{01} \geq \varepsilon_{01} - 2\omega(1+\varepsilon_{01})\delta$*.*

F.4. Approximation results for Lemma 2(b)

Let us collect some approximation lemmas that will be used in the proof of Lemma 2(b). The proofs can be found in Appendix G.4. We write pmf for the probability mass functions. We recall that a Poisson-binomial variable with parameter $(p_1, \dots, p_n)$ is one that can be written as $\sum_{i=1}^n X_i$ where $X_i \sim \mathrm{Ber}(p_i)$, independent over $i = 1, \dots, n$. We write pmf for the probability mass function.

Lemma 22 (Poisson-binomial approximation)**.** *Let $\varphi(x;\lambda)$ be the pmf of a Poisson variable with mean λ, and let $\widetilde{\varphi}(x,p)$ be the pmf of a Poisson-binomial variable with parameters $p = (p_1, \dots, p_n)$ where $\sum_{j=1}^n p_j = \lambda$. Let $p^* := \max_{j \in [n]} p_j$. Then,*

$$\frac{\widetilde{\varphi}(x;p)}{\varphi(x;\lambda)} \le e^{xp^*}, \quad \forall x \in \mathbb{Z}_+.$$

This result immediately extends to the comparison between vector versions of the two distributions:

Corollary 7 (Poisson-binomial approximation)**.** *Let $p^{(\ell)} = (p_1^{(\ell)}, \dots, p_{n_\ell}^{(\ell)}) \in [0,1]^{n_\ell}$ be a vector of probabilities for each $\ell \in [L]$ and let $\lambda^{(\ell)} = \sum_{i=1}^{n_\ell} p_i^{(\ell)} \in \mathbb{R}_+$. Let*

$$\widetilde{\Phi}(x, (p^{(1)}, \dots, p^{(L)})) := \prod_{\ell=1}^{L} \widetilde{\varphi}(x_\ell; p^{(\ell)}), \quad \textit{for each } x = (x_1, \dots, x_L) \in \mathbb{Z}_+^L \tag{110}$$

be the pmf of a vector Poisson-binomial variable, and $\Phi(x, (\lambda^{(1)}, \dots, \lambda^{(L)})) = \prod_{\ell=1}^{L} \varphi(x_\ell; \lambda^{(\ell)})$ be the corresponding vector Poisson pmf. Then, we have

$$\frac{\widetilde{\Phi}(x, (p^{(1)}, \dots, p^{(L)}))}{\Phi(x, (\lambda^{(1)}, \dots, \lambda^{(L)}))} \le \exp\Big(p^* \sum_{\ell=1}^{L} x_\ell\Big), \quad \forall x \in \mathbb{Z}_+^L,$$

where $p^* = \max\{p_i^{(\ell)} : i \in [n_\ell], \ell \in [L]\}$.

Lemma 23 (Poisson likelihood approximation)**.** *Suppose $\max(|\lambda_1 - \lambda|, |\lambda_2 - \lambda|) \le \rho \le \frac{1}{3}\lambda$, then for any $x \in \mathbb{Z}_+$, we have*

$$\frac{\phi(x;\lambda_1)}{\phi(x;\lambda_2)} \le \exp\Big(\frac{3\rho x}{\lambda} + 2\rho\Big).$$

Lemma 24 (Degree Truncation)**.** *Let $b_{i+} = \sum_{\ell \in [L]} b_{i\ell} = \sum_{j=1}^m A_{ij}$ be the degree of (row) node i. Then,*

$$\mathbb{P}(b_{i+} > 5L\|\Lambda\|_\infty) \le \exp(-3L\|\Lambda\|_\infty).$$

of Lemma 24. Let row node i belong to row cluster k, and let $b_{i+} = \sum_{\ell \in [L]} b_{i\ell} = \sum_{j=1}^m A_{ij}$ be its degree, with expectation $\lambda_{k+} := \sum_{\ell \in [L]} \lambda_{k\ell}$. By definition, we have $\lambda_{k+} \le L\|\Lambda\|_\infty$. We would like to find an upper bound on the probability

$$\mathbb{P}\big(\, b_{i+} > 5L\|\Lambda\|_\infty \,\big) \le \mathbb{P}\big(\, b_{i+} - \lambda_{k+} > 4L\|\Lambda\|_\infty \,\big)$$

We let $v = \lambda_{k+}$, $vt = 4L\|\Lambda\|_\infty$, so $t \ge 4$. By Proposition 2, we have

$$\mathbb{P}\big(b_{i+} - \lambda_{k+} > 4L\|\Lambda\|_\infty\big) \le \exp\Big[-\frac{3}{4}vt\log\Big(1 + \frac{2t}{3}\Big)\Big] \le \exp\Big(-\frac{3}{4}vt\Big) \le \exp(-3L\|\Lambda\|_\infty).$$

□

F.5. Proof of Lemma 2(b)

Fix $i \in [n]$ such that $y_i = k$, and $\tilde{z} \in [K]^n$ and let $b_{i*} = b_{i*}(\tilde{z})$. Throughout, let $\Lambda' = (\lambda'_{k\ell}) := \Lambda(y, \tilde{z})$ which belongs to $\mathscr{B}_\Lambda(\delta)$ by assumption. Denoting the kth row of Λ' as λ'_{k*}, we have $\mathbb{E}[b_{i*}] = \lambda_{k*}$. For $r \neq k \in [K]$, i such that $y_i = k$ and $\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta)$,

$$Y_{ikr}(b_{i*}, \tilde{\Lambda}) = \sum_{\ell=1}^{L} b_{i\ell} \log \frac{\tilde{\lambda}_{r\ell}}{\tilde{\lambda}_{k\ell}} + \tilde{\lambda}_{k\ell} - \tilde{\lambda}_{r\ell} \leq \sum_{\ell=1}^{L} \Big[b_{i\ell} \log \frac{\lambda_{r\ell} + \rho}{\lambda_{k\ell} - \rho} + \lambda_{k\ell} - \lambda_{r\ell} + 2\rho \Big] := Y^*$$

where $\rho := \delta \|\Lambda\|_\infty$ is the radius of $\mathscr{B}_\Lambda(\delta)$. Hence,

$$\mathbb{P}(\exists \tilde{\Lambda} \in \mathscr{B}_\Lambda, Y_{ikr}(\tilde{\Lambda}) \geq 0) \ \leq \ \mathbb{P}(Y^* \geq 0) \ = \ \mathbb{P}(b_{i*} \in F),$$

where we have defined (recalling the definition of Ψ from (33)):

$$F := \Big\{ x \in \mathbb{Z}_+^L : \ \Psi(x; \lambda_{r*} + \rho \mid \lambda_{k*} - \rho) \ \geq \ -2L\rho \Big\}.$$

Degree truncation. Let $b_{i+} = \sum_{\ell \in [L]} b_{i\ell} = \sum_{j=1}^m A_{ij}$ be the degree of (row) node i, and

$$E = \Big\{ x \in \mathbb{Z}_+^L : \sum_{\ell=1}^{L} x_\ell \leq 5L\|\Lambda\|_\infty \Big\}. \tag{111}$$

Using Lemma 24, we have $\mathbb{P}(b_{i*} \notin E) \leq \exp(-3L\|\Lambda\|_\infty)$, which is faster than the rate we want to establish. Hence, for the rest of the proof it is enough to work on $\{b_{i*} \in E\}$. We have the following two approximations on this event:

Poisson-binomial approximation. Recall that P is the connectivity matrix and we have,

$$\|P\|_\infty \leq \frac{\|\Lambda\|_\infty}{\min_i n_i(z)} \leq \frac{\beta L \|\Lambda\|_\infty}{m} \tag{112}$$

where the first inequality follows from definition of Λ in (2), and the second from assumption (A2). We note that $b_{i\ell} = b_{i\ell}(\tilde{z}) = \sum_{j=1}^m A_{ij} 1\{\tilde{z}_j = \ell\}$ as defined in (27), follows a Poisson-binomial distribution. In order the describe the parameters of this distribution, let us introduce the following notation

$$\mathrm{lab}_\ell(\tilde{z}) := (z_j : \ j \in [m] \text{ such that } \tilde{z}_j = \ell\},$$

that is, the vector of true labels associated with nodes in the ℓth cluster of $\tilde{z}$. Then, $P_{k,\mathrm{lab}_\ell(\tilde{z})} = (P_{k,z_j} : \ j \in [m] \text{ s.t. } \tilde{z}_j = \ell) \in \mathbb{R}^{n_\ell(\tilde{z})}$ is the probability vector associated with the Poisson-binomial distribution of $b_{i\ell}$. Also, let

$$\mathrm{lab}(\tilde{z}) := (\mathrm{lab}_1(\tilde{z}), \ldots, \mathrm{lab}_L(\tilde{z})), \quad \text{and} \quad P_{k,\mathrm{lab}(\tilde{z})} := (P_{k,\mathrm{lab}_1(\tilde{z})}, \ldots, P_{k,\mathrm{lab}_L(\tilde{z})}).$$

Then, we can say that $b_{i*} = b_{i*}(\tilde{z})$ is a product Poisson-binomial distribution with parameter $P_{k,\mathrm{lab}(\tilde{z})}$. In particular, b_{i*} has pmf $\widetilde{\Phi}(x; P_{k,\mathrm{lab}(\tilde{z})})$ as defined in (110). We also note that $\mathbb{E}[b_{i*}(\tilde{z})] = \lambda_{k*}(y, \tilde{z}) =: \lambda'_{k*}$. It follows from Corollary 7, noting that $\|P_{k,\mathrm{lab}(\tilde{z})}\|_\infty \leq \|P\|_\infty$ combined with (112),

$$\frac{\widetilde{\Phi}(x; P_{k,\mathrm{lab}(\tilde{z})})}{\Phi(x; \lambda'_{k*})} \leq \exp\Big(\|P_{k,\mathrm{lab}(\tilde{z})}\|_\infty \sum_{\ell=1}^{L} x_\ell \Big) \leq \exp\Big(\frac{5\beta L^2 \|\Lambda\|_\infty^2}{m} \Big) =: \zeta_1, \quad \forall x \in E.$$

Poisson likelihood approximation. Recall that $\rho = \delta\|\Lambda\|_\infty$. Since by assumption, $\omega\delta \le \frac{1}{3}$, we have $\rho \le \frac{\|\Lambda\|_\infty}{3\omega} \le \frac{1}{3}\Lambda_{\min}$. Recall that by assumption $\Lambda' = (\lambda'_{k\ell}) \in \mathscr{B}_\Lambda(\delta)$. By Lemma 23,

$$\begin{aligned}\frac{\Phi(x;\lambda'_{k*})}{\Phi(x;\lambda_{k*}-\rho)} &\le \prod_{\ell\in[L]} \exp\Big(\frac{3\rho x_\ell}{\lambda_{k\ell}} + 2\rho\Big) \le \exp\Big(2L\rho + \frac{15\rho}{\Lambda_{\min}} L\|\Lambda\|_\infty\Big)\\ &\le \exp\Big(17\omega L\rho\Big) =: \zeta_2, \quad \forall x \in E.\end{aligned}$$

With some abuse of notation, we treat Φ and $\widetilde{\Phi}$ are measures as well, thus, for example, $\Phi(E) = \sum_{x\in E}\Phi(x)$. Then, we have

$$\mathbb{P}(b_{i*} \in E\cap F) \;=\; \widetilde{\Phi}\big(E\cap F; P_{k*}\big) \;\le\; \zeta_1\,\Phi\big(E\cap F;\lambda'_{k*}\big) \;\le\; \zeta_1\zeta_2\,\Phi\big(E\cap F;\lambda_{k*}-\rho\big). \tag{113}$$

Thus, it is enough to bound $\Phi(F;\lambda_{k*}-\rho)$ which gives a further upper bound. This quantity is closely related to testing Poisson vector distributions with mean $\lambda_{k*}-\rho$ and $\lambda_{r*}-\rho$ against each other. Let us write $p_0(x) := \Phi(x;\lambda_{k*}-\rho)$ and $p_1(x) := \Phi(x;\lambda_{r*}+\rho)$ and note that $\Psi(\,\cdot\,;\lambda_{r*}+\rho \mid \lambda_{k*}-\rho) = \log(p_1/p_0)$. We have

$$\begin{aligned}\sum_{x\in F}\Phi(x;\lambda_{k*}-\rho) &= \sum_{x\,\in\,\mathbb{Z}_+^L} p_0(x)\; 1\Big\{\log\frac{p_1(x)}{p_0(x)} \ge -2L\rho\Big\}\\ &= \sum_{x\,\in\,\mathbb{Z}_+^L} p_0(x)\; 1\Big\{\frac{e^{2L\rho}p_1(x)}{p_0(x)} \ge 1\Big\}\\ &\le \sum_{x\,\in\,\mathbb{Z}_+^L} \min\Big(e^{2L\rho}p_1(x),\; p_0(x)\Big) \le e^{2L\rho}\sum_{x\,\in\,\mathbb{Z}_+^L}\min\big(p_1(x),\; p_0(x)\big).\end{aligned} \tag{114}$$

Let us define

$$I_s(\lambda_0 \mid \lambda_1) = \sum_{\ell=1}^{L}\big[s\lambda_{0\ell} + (1-s)\lambda_{1\ell}\big] - \lambda_{0\ell}^s\lambda_{1\ell}^{1-s}, \quad \lambda_0,\lambda_1 \in \mathbb{R}_+^L. \tag{115}$$

We can now apply Corollary 6. Since $\frac{\|\Lambda\|_\infty+\rho}{\Lambda_{\min}-\rho} \le \frac{\omega\Lambda_{\min}+\frac{1}{3}\Lambda_{\min}}{\frac{2}{3}\Lambda_{\min}} \le 2\omega$, we need to substitute ω in Corollary 6 by 2ω. It follows that

$$\begin{aligned}\sum_{x\,\in\,\mathbb{Z}_+^L}\min\big(p_1(x),\; p_0(x)\big) \;&\le\; \zeta_3\exp\Big(-I_s(\lambda_{r*}+\rho \mid \lambda_{k*}-\rho) - \frac{1}{2}\log(\Lambda_{\min}-\rho)\Big)\\ &\le\; \zeta_3\exp\Big(-I_s(\lambda_{k*} \mid \lambda_{r*}) + 2\omega L\rho - \frac{1}{2}\Big(\log\Lambda_{\min} + \log\frac{2}{3}\Big)\Big)\\ &\le\; 8\sqrt{\frac{3}{2}}\zeta_3\exp\Big(-I_s(\lambda_{k*} \mid \lambda_{r*}) + 2\omega L\rho - \frac{1}{2}\log\Lambda_{\min}\Big)\end{aligned} \tag{116}$$

where $\zeta_3 = \omega/(\varepsilon_{kr} - 2\omega(1+\varepsilon_{kr})\delta) + \omega$ from Lemma 21, and the second line follows from the following elementary inequality:

$$(a-\rho)^{1-s}(b+\rho)^s \le a^{1-s}\Big(b^s + \frac{s\rho}{b^{1-s}}\Big) \le a^{1-s}b^s + \rho\omega$$

assuming $a/b \le \omega$. Note that $I_{kr} = \sup_{s\in(0,1)} I_s(\lambda_{r*} \mid \lambda_{k*})$. Putting the pieces (113), (114) and (116) together (and taking supremum over s) we have

$$\mathbb{P}(b_{i*} \in E \cap F) \;\le\; 8\sqrt{\frac{3}{2}}\zeta_2\zeta_3\, e^{2L\rho+2\omega L\rho} \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min}\Big).$$

We note that

$$\log(\zeta_1\zeta_2\, e^{2L\rho+2\omega L\rho}) \;\le\; 17\omega L\rho + \frac{5\beta L^2\|\Lambda\|_\infty^2}{m} + 4\omega L\rho \;\le\; 21\omega L\rho + \frac{5\beta L^2\|\Lambda\|_\infty^2}{m} =: \log\zeta_4$$

It follows that

$$\mathbb{P}(b_{i*} \in E \cap F) \le \zeta_3\zeta_4 \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min}\Big).$$

Finally we have

$$\begin{aligned}\mathbb{P}(b_{i*} \in F) &\le \mathbb{P}(b_{i*} \in E \cap F) + \mathbb{P}(b_{i*} \in E^c)\\ &\le 8\sqrt{\frac{3}{2}}\zeta_3\zeta_4 \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min}\Big) + \exp\big(-3L\|\Lambda\|_\infty\big)\\ &\le 11\zeta_3\zeta_4 \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min}\Big),\end{aligned}$$

assuming that I_{kr} and $\Lambda_{\min}$ are sufficiently large. Noting that by the definition of η_{kr} in the statement of the theorem, $\eta_{kr} = \log(2\zeta_3\zeta_4)$, the proof is complete.

Appendix G. Remaining proofs

G.1. Proofs of Appendix B.2, B.3 and B.4

Proof of Lemma 4. We have $n_k(y^{(q)}) \sim \text{Hypergeometric}(n/Q, n_k(y), n)$. For any fixed $k \in [K]$ and $q' \in [Q]$, the concentration of hypergeometric distribution (Chvátal, 1979) gives $|\pi_k(y^{(q')}) - \pi_k(y)| \le \xi$ with probability at least $1 - 2\exp(-n\xi^2/Q)$. The same probability bound holds for $|\pi_\ell(z^{(q)}) - \pi_\ell(z)| \le \xi$, for any fixed $\ell \in [L]$ and $q \in [Q]$. Taking the union bound over k, ℓ, q, q' gives the desired result. □

Proof of Lemma 5. Recall the definition of the true local mean parameters in (26), and the corresponding global parameters in Section 4.1. We have

$$\begin{aligned}\lambda_{k\ell}^{(q)} - \frac{\lambda_{k\ell}}{Q} &= P_{k\ell}\Big(n_\ell(z^{(q)}) - \frac{n_\ell(z)}{Q}\Big)\\ &= \frac{P_{k\ell}\, n_\ell(z)}{Q}\Big(\frac{\pi_\ell(z^{(q)})}{\pi_\ell(z)} - 1\Big) = \frac{\lambda_{k\ell}}{Q}\Big(\frac{\pi_\ell(z^{(q)})}{\pi_\ell(z)} - 1\Big).\end{aligned}$$

Since $|\pi_\ell(z^{(q)}) - \pi_\ell(z)| \le \xi$ and $\pi_\ell(z) \ge 1/(\beta L)$ by assumptions (B4a) and (A2), the first inequality in (57) follows, from which we have the second inequality by (B4a). □

Proof of Lemma 6. From Lemma 5, we have $\|\Lambda^{(q)} - \Lambda/Q\|_\infty \le (\xi L\beta)\|\Lambda/Q\|_\infty$ and $\|\Lambda^{(q)}\|_\infty \le \frac{3}{2}\|\Lambda/Q\|_\infty$. Note that $\Lambda^{(q)}$ is the true (local) mean parameter matrix associated with subblock $A^{(q',q)}$, and this subblock has $n/(2Q)$ rows. We will apply Lemma 1 to the submatrix $A^{(q',q)}$ and sublabels $z^{(q')}$ and $y^{(q)}$. In order to do so, we have to verify conditions (A1), (A2)) and (B3) for the subblock. (Condition (B3) is satisfied by assumption.) By Lemma 5, we have

$$\frac{\|\Lambda^{(q)}\|_\infty}{\Lambda^{(q)}_{\min}} \le 3\frac{\|\Lambda\|_\infty}{\Lambda_{\min}} \le 3\omega.$$

By (55), the condition (A2) holds with β replaced with 2β. We also need to replace $\Lambda_{\min}$ in Lemma 5 with $\Lambda^{(q)}_{\min} \ge \Lambda_{\min}/(2Q)$, and C_γ with $4C_\gamma$ (more precisely, we are replacing $C_{\gamma,\beta}$ with $C_{\gamma,2\beta}$). Thus, assuming $6(4C_\gamma)(3\omega) \le 1$, we obtain

$$\mathbb{P}\Big(\|\hat{\Lambda}^{(q',q)} - \Lambda^{(q)}\|_\infty \le 4(4C_\gamma + \tau)\|\Lambda^{(q)}\|_\infty\Big) \ge 1 - 2p_1\Big(\tau;\, \frac{n}{2Q}, \frac{\Lambda_{\min}}{2Q}, 2\beta\Big) \tag{117}$$

where $p_1(\cdot)$ is as in (45). Since $\|\Lambda^{(q)}\|_\infty \le \frac{3}{2}\|\Lambda/Q\|_\infty$, $4(4C_\gamma{+}\tau)\|\Lambda^{(q)}\|_\infty \le (24C_\gamma{+}6\tau)\|\Lambda/Q\|_\infty$. Thus, on the event in (117), we have by triangle inequality

$$\begin{aligned}\|\hat{\Lambda}^{(q',q)} - \Lambda/Q\|_\infty &\le 4(4C_\gamma + \tau)\|\Lambda^{(q)}\|_\infty + (\xi L\beta)\|\Lambda/Q\|_\infty \\ &\le \Big[4(4C_\gamma + \tau)\frac{3}{2} + \xi L\beta\Big]\|\Lambda/Q\|_\infty,\end{aligned}$$

which is the desired result. □

Proof of Lemma 7. For the proof, it is enough to consider $\Lambda = [\lambda_0\,;\,\lambda_1] \in \mathbb{R}^{2\times L}_+$, where $\lambda_0, \lambda_1 \in \mathbb{R}^L_+$ are the two rows of Λ. Similarly, let $\tilde{\Lambda} = [\tilde{\lambda}_0\,;\,\tilde{\lambda}_1] \in \mathbb{R}^{2\times L}_+ \in \mathscr{B}_\Lambda(\delta)$. Let us define

$$I_s(\lambda_0 \mid \lambda_1) = \sum_{\ell=1}^{L} \big[(1-s)\lambda_{0\ell} + s\lambda_{1\ell}\big] - \lambda_{0\ell}^{1-s}\lambda_{1\ell}^{s}, \quad \lambda_0, \lambda_1 \in \mathbb{R}^L_+ \tag{118}$$

and $\alpha_\ell = \max\{|\lambda_{0\ell} - \tilde{\lambda}_{0\ell}|, |\lambda_{1\ell} - \tilde{\lambda}_{1\ell}|\}$. We have

$$\big|I_s(\lambda_0 \mid \lambda_1) - I_s(\tilde{\lambda}_0 \mid \tilde{\lambda}_1)\big| \le \sum_{\ell=1}^{L}\Big[\alpha_\ell + \big|\lambda_{0\ell}^{1-s}\lambda_{1\ell}^{s} - \tilde{\lambda}_{0\ell}^{1-s}\tilde{\lambda}_{1\ell}^{s}\big|\Big].$$

Consider the function $f(a,b) = a^{1-s}b^s$ for $a, b > 0$. Assuming $\max\{a/b, b/a\} \le \omega$, we have

$$\|\nabla f(a,b)\|_1 \le (1-s)(b/a)^s + s(a/b)^{1-s} \le (1-s)\omega^s + s\omega^{1-s} \le \omega$$

using $\omega \ge 1$. It follows that $|a^{1-s}b^s - u^{1-s}v^s| \le \omega\max\{|a-u|, |b-v|\}$ for $a, b, u, v > 0$. Thus, $\big|I_s(\lambda_0 \mid \lambda_1) - I_s(\tilde{\lambda}_0 \mid \tilde{\lambda}_1)\big| \le (1+\omega)\sum_\ell \alpha_\ell \le 2\omega L\delta\|\Lambda\|_\infty$ since $\max_\ell \alpha_\ell \le \delta\|\Lambda\|_\infty$ by assumption. Taking the supremum over s gives part (a). □

Proof of Lemma 8. Assume $\mathrm{dMis}(\tilde{y}, y) \le \alpha$ and let $n_k = |i : y_i = k|$ and $N_{kk'} = |i :\; y_i = k,\, \tilde{y}_i = k'|$. Then,

$$n\,\mathrm{dMis}(\tilde{y}, y) = \sum_k \sum_{k' \ne k} N_{kk'} \le \alpha n \le \frac{\alpha}{\pi_k} n_k =: \varepsilon n_k.$$

It follows that $\sum_{k' \ne k} N_{kk'} \le \varepsilon n_k$ for every k. We also obtain $N_{kk'} \le \varepsilon n_k$ for all k and k' such that $k \ne k'$. Since $\sum_{k'} N_{kk'} = n_k$, we have $N_{kk} \ge (1-\varepsilon)n_k$. Thus, as long as $\varepsilon < 1/2$, we have

$N_{kk} > N_{kk'}$ for all k and k' such that $k \neq k'$. That is, the diagonal of the confusion matrix is bigger than every element in the corresponding row. Now take $\sigma \neq \mathrm{id}$. Then, there exists k such that $k' := \sigma^{-1}(k) \neq k$.

$$N^{\sigma}_{kk} := |i: \; y_i = k, \, \sigma(\tilde{y}_i) = k| \; = \; |i: \; y_i = k, \, \tilde{y}_i = k'| = N_{kk'} < N_{kk}.$$

Then we have

$$n \, \mathrm{dMis}(\sigma(\tilde{y}), y) = \sum_k (n_k - N^{\sigma}_{kk}) > \sum_k (n_k - N_{kk}) = n \, \mathrm{dMis}(\tilde{y}, y).$$

showing that id is the unique optimal permutation and proving part (a). For part (b), we note that $|\{i : \tilde{y}_i = k\}| \geq N_{kk} \geq (1-\varepsilon) n_k > (1/2) n_k$ whenever $\varepsilon < 1/2$. □

Proof of Lemma 9. Assume that $\mathrm{Mis}(\tilde{y}, y) \leq \alpha$ and $\mathrm{Mis}(\tilde{y}', y) \leq \alpha$ where $\alpha < \frac{1}{4} \min_k \pi_k(\tilde{y})$. By definition of the optimal permutation, $\mathrm{dMis}(\sigma(\tilde{y}), y) \leq \alpha$ and $\mathrm{dMis}(\sigma'(\tilde{y}'), y) \leq \alpha$. Since dMis is a metric (being the sum of discrete metrics over the coordinates), we have

$$\mathrm{dMis}(\sigma^{-1} \circ \sigma'(\tilde{y}'), \tilde{y}) = \mathrm{dMis}(\sigma'(\tilde{y}'), \sigma(\tilde{y})) \leq 2\alpha < \frac{1}{2} \min_k \pi_k(\tilde{y})$$

where the first inequality is by the triangle inequality for dMis and the second by assumption. Applying Lemma 8 gives the desired result. □

Proof of Corollary 2. Take $q = 2$ for simplicity. Assume that (60) holds with constant 8 in place of 32, which is all we need for this lemma. We have

$$(n/2) \, \mathrm{dMis}(\sigma_{12}(\tilde{y}^{(1)}), y^{(1)}) \leq n \, \mathrm{dMis}(\sigma_{12}(\tilde{y}^{(1,2)}), y^{(1,2)})$$

by the definition of the dMis. It then follows that

$$\mathrm{dMis}(\sigma_{12}(\tilde{y}^{(1)}), y^{(1)}) < 2 \frac{1}{8\beta K} \leq \frac{1}{2} \min_k \pi_k(y^{(1)})$$

where the second inequality holds by the counterpart of (55) for row labels. Applying Lemma (8) we conclude that $\sigma_{12} = \sigma^*(\tilde{y}^{(1)} \to y^{(1)}) =: \sigma_1$. □

Proof of Corollary 3. Take $q = 2$ for simplicity. Let $\varepsilon = 1/(32\beta K)$. By assumption, we have $\mathrm{Mis}(\tilde{y}^{(1,2)}, y^{(1,2)}) < \varepsilon$ and $\mathrm{Mis}(\tilde{y}^{(2,3)}, y^{(2,3)}) < \varepsilon$. By Corollary 2, $\sigma_{12} = \sigma_2$ and $\sigma_{23} = \sigma'_2$. Then, the argument leading to (38) implies $\mathrm{Mis}(\tilde{y}^{(2)}, y^{(2)}) < 2\varepsilon$ and $\mathrm{Mis}(\tilde{y}'^{(2)}, y^{(2)}) < 2\varepsilon$. By assumption,

$$\mathrm{Mis}(\tilde{y}'^{(2)}, y^{(2)}) < 2\varepsilon = \frac{1}{16\beta K} \leq \frac{1}{8} \min_k \pi_k(y^{(2)}) \leq \frac{1}{4} \min_k \pi_k(\tilde{y}^{(2)})$$

where the second inequality holds by the counterpart of (55) for row labels, and the third inequality follows from the second inequality and Lemma 8(b). It thus follows from Lemma 9 that $\sigma_2^{-1} \circ \sigma'_2 = \sigma^*(\tilde{y}'^{(2)} \to \tilde{y}^{(2)})$ which is the desired result. □

G.2. Proofs of Appendix C.1

Proof of Lemma 14. Let us define $I := I_{kr}$ and

$$I_s = \sum_{\ell=1}^{L} (1-s)\lambda_{k\ell} + s\lambda_{r\ell} - \lambda_{k\ell}^{1-s}\lambda_{r\ell}^{s}.$$

in this proof. For $s \in [0,1]$, $s \mapsto I_s$ is a concave function and $I_0 = I_1 = 0$. We have defined $I := I_{s^*} = \sup_{s\in[0,1]} I_s$. Suppose $s^* \geq \frac{1}{2}$, since $0\Big(1 - \frac{1}{2s^*}\Big) + \frac{s^*}{2s^*} = \frac{1}{2}$ and $\frac{1}{2s^*} \geq \frac{1}{2}$, by concavity,

$$I_{1/2} \geq \Big(1 - \frac{1}{2s^*}\Big) I_0 + \frac{1}{2s^*} I_{s^*} \geq \frac{1}{2} I_{s^*} = \frac{I}{2}$$

Similarly, suppose $s^* \leq \frac{1}{2}$, $I_{1/2} \geq I/2$ still holds, from which it follows that

$$\sum_{\ell\in[L]} (\lambda_{r\ell} - \lambda_{k\ell})^2 = \sum_{\ell\in[L]} (\sqrt{\lambda_{r\ell}} - \sqrt{\lambda_{k\ell}})^2 (\sqrt{\lambda_{r\ell}} + \sqrt{\lambda_{k\ell}})^2 \geq (I/2)(4\Lambda_{\min}) = 2\Lambda_{\min} I.$$

Taking the minimum over k and r completes the proof. □

Proof of Lemma 15. Recall our choice of ξ in (75)—which will also be assumed in this proof—giving $\mathbb{P}(\mathfrak{P}^c) = o(1)$ as shown (78). By Lemma 5, we have $\Lambda^{(q)} \in \mathscr{B}_{\Lambda/Q}(\xi L \beta)$ for all $q \in [Q]$, which combined with Lemma 7 (applied with $\delta = \xi L \beta$) gives

$$|I_{kr}(\Lambda^{(q)}) - I_{kr}(\Lambda/Q)| \leq 2\omega(\xi L\beta) L \|\Lambda/Q\|_\infty \leq \frac{2\omega(L\beta)L\|\Lambda/Q\|_\infty}{\beta\omega(K\vee L)^2(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty)} \leq \frac{2}{Q},$$

using (75). Thus

$$I_{\min}(\Lambda^{(q)}) \geq I_{\min}(\Lambda/Q) - \frac{2}{Q} \geq \frac{I_{\min}}{2Q}$$

as $I_{\min} \to \infty$. We are now ready to apply Corollary 5 to the Algorithm 4 operating on subblocks in G_1^{col}. It remains to verify that assumption (86) translates to condition (85) for the subblocks. Indeed, we have to replace $I_{\min}$ with $I_{\min}(\Lambda^{(q)})$, ω with 3ω (by Lemma 5), β with 2β (by (55)), and α with $\frac{m/4}{n/2} = \frac{\alpha}{2}$ since the subblocks in G_1^{col} are of size $\frac{n}{2} \times \frac{m}{4}$. Therefore, by assumption (86),

$$\frac{(2\beta)^2(3\omega)KL(K\wedge L)\,(\alpha/2)}{2I_{\min}(\Lambda^{(q)})} \leq \frac{6Q\beta^2\omega^2 KL(K\wedge L)\,\alpha}{I_{\min}} \leq C_1(1+\kappa)^{-2}, \tag{119}$$

verifying condition (85) on the subblocks. Applying Corollary 5, we have the misclassification rate of $\tilde{y}^{(q)}$ satisfies

$$\text{Mis}(\tilde{y}^{(q)}, y^{(q)}) \;\leq\; \frac{(1+\kappa)^2(3\omega)(2\beta)L(K\wedge L)(\alpha/2)}{2C_1(I_{\min}/2Q)}$$

which is the desired result. □

G.3. Proofs of Appendix F.3

Proof of Proposition 3. **Step 1: Interpolation.** Assume without loss of generality that $\theta_{01} \neq \theta_{11}$ and fix some $s \in (0,1)$. It is enough to establish the bound for $\ell = 1$ and this particular s. Let $P_{e,+} := P_{e,0} + P_{e,1}$ the be sum of the error probabilities under the two hypothesis. Then,

$$P_{e,+} = \int p_0 1\{p_0 \le p_1\} d\mu + \int p_1 1\{p_1 < p_0\} d\mu \tag{120}$$

$$= \int \min(p_0, p_1) d\mu = \int p_0^{1-s}\, p_1^s \, \min(l^s, l^{s-1}) d\mu. \tag{121}$$

where $l = p_0/p_1$ is the likelihood ratio. Let $p_{r\ell} := \pi(\,\cdot\,; \theta_{r\ell})$ so that $p_r(x) = \prod_\ell p_{r\ell}(x_\ell)$. Similarly, let

$$p_s := \frac{p_0^{1-s}\, p_1^s}{\int p_0^{1-s}\, p_1^s d\mu}, \quad \text{and} \quad p_{s\ell} := \frac{p_{0\ell}^{1-s}\, p_{1\ell}^s}{\int p_{0\ell}^{1-s}\, p_{1\ell}^s d\nu} \tag{122}$$

It is easy to see that $p_s(x) = \prod_{\ell=1}^L p_{s\ell}(x_\ell)$ and each $p_{s\ell}$ is a probability density (w.r.t. ν). One can also verify that

$$\int p_{0\ell}^{1-s}\, p_{1\ell}^s d\nu = e^{-I_{s\ell}}, \quad \text{and} \quad p_{s\ell} = \pi(\,\cdot\,; \theta_{s\ell}),$$

hence $p_s = p(\,\cdot\,; \theta_s)$ using definition (100). That is, p_s defined in (122) belongs to the same exponential family, with parameter θ_s interpolating θ_0 and θ_1. We also note that $p_{0\ell}^{1-s}\, p_{1\ell}^s = e^{-I_{s\ell}} p_{s\ell}$, hence $p_0^{1-s}\, p_1^s = e^{-I_s} p_s$. Substituting into (120), we obtain

$$P_{e,+} = e^{-I_s} \int p_s \, \min(l^s, l^{s-1}) d\mu. \tag{123}$$

Step 2: Reduction to the single component case ($L = 1$). Using $p_{r\ell}(t) = \pi(t; \theta_{r\ell})$, we have $p_{r\ell}(t)/p_{r\ell}(t') = \exp(\theta_{r\ell}(t - t'))$, hence

$$\frac{p_{0\ell}(t)}{p_{0\ell}(t')} \frac{p_{1\ell}(t')}{p_{1\ell}(t)} = \exp\left[(\theta_{0\ell} - \theta_{1\ell})(t - t')\right]$$

Using $p_r(x) = \prod_\ell p_{r\ell}(x_\ell)$, the likelihood ratio can be written as

$$l(x) = \frac{p_0(x)}{p_1(x)} = \prod_\ell l_\ell(x_\ell), \quad \text{where} \quad l_\ell(x_\ell) = \frac{p_{0\ell}(x_\ell)}{p_{1\ell}(x_\ell)} = \exp\left[(\theta_{0\ell} - \theta_{1\ell})x - A(\theta_{0\ell}) + A(\theta_{1\ell})\right]. \tag{124}$$

As long as $\theta_{0\ell} \neq \theta_{1\ell}$, l_ℓ is well defined on $\mathbb{R}$ and maps onto $\mathbb{R}^{++}$. For any $(x_2, \dots, x_L)$, let $x_1^* = x_1^*(x_2, \dots, x_L)$ be the solution of the following equation:

$$l_1(x_1^*) \prod_{\ell=2}^L l_\ell(x_\ell) = 1$$

which always exists in $\mathbb{R}$ (and not necessarily on the support of the exponential family). Then, we have, setting $\delta = \theta_{01} - \theta_{11}$,

$$l(x) = \frac{l_1(x_1)}{l_1(x_1^*)} = \frac{p_{01}(x_1)}{p_{01}(x_1^*)} \frac{p_{11}(x_1^*)}{p_{11}(x_1)} = \exp\left[(\theta_{01} - \theta_{11})(x_1 - x_1^*)\right] = \exp[\delta(x_1 - x_1^*)].$$

It follows that

$$\min(l(x)^s, l(x)^{s-1}) \leq e^{-\min(s,1-s)|\delta(x_1 - x_1^*)|} = e^{-\alpha|x_1 - x_1^*|}$$

where we have defined $\alpha := |\delta| \min(s, 1-s)$. Recall that $p_s(x) = \prod_{\ell=1}^L p_{s\ell}(x_\ell)$ which we write compactly as $p_s = \prod_{\ell=1}^L p_{s\ell}$. Let us write $\mu = \mu^1 \times \mu^{2:L}$ as the product of underlying coordinate measures. By Fubini theorem, we first integrate over the first coordinate in (123):

$$e^{I_s} P_{e,+} = \int \prod_{\ell=2}^{L} p_{s\ell} \Big[\int p_{s1} \min(l^s, l^{s-1}) d\mu^1 \Big] d\mu^{2:L} \tag{125}$$

Let $J = J(x_2, \dots, x_L)$ denote the inner integral in (125) (in brackets). We have the bound

$$J \leq \int p_{s1}(x_1) e^{-\alpha|x_1 - x_1^*|} d\mu^1(x_1) \leq \|p_{s1}\|_\infty \int e^{-\alpha|x_1 - x_1^*|} d\mu^1(x_1).$$

Note that x_1^* is the only place where dependence on $(x_2, \dots, x_L)$ appears in the bound. Since μ^1 is either the Lebesgue or the counting measure, and both these measures are translation invariant, the bound is in fact independent of x_1^*. That is, we have $J(x_2, \dots, x_L) \leq C(\alpha)\|p_{s1}\|_\infty$ for all $(x_2, \dots, x_L)$. It follows that the same bound holds for $P_{e,+}$ by (125), that is,

$$e^{I_s} P_{e,+} \; \leq \; C(\alpha)\|p_{s1}\|_\infty \int \Big(\prod_{\ell=2}^{L} p_{s\ell} \Big) d\mu^{2:L} \; = \; C(\alpha)\|p_{s1}\|_\infty$$

since $\prod_{\ell=2}^L p_{s\ell}$ is a probability density w.r.t $\mu^{2:L}$. Since the choice of the coordinate $\ell = 1$ and s was arbitrary, the proof is complete.

□

G.4. Proofs of Appendix F.3.1

Proof of Lemma 17. Let $f(s) = \sum_{\ell=1}^L (s-1)\lambda_{k\ell} - s\lambda_{r\ell} + \lambda_{k\ell}^{1-s}\lambda_{r\ell}^s$, then $f(s)$ is a concave function of s on $\mathbb{R}_+$. Since $f(0) = f(1) = 0$, $s^* \in (0,1)$. First, we show the statement is true when $L = 1$. In this case, s^* satisfies

$$\lambda_{k1} - \lambda_{r1} + \lambda_{k1}^{1-s^*} \lambda_{r1}^{s^*} \log\left(\frac{\lambda_{r1}}{\lambda_{k1}}\right) = 0 \tag{126}$$

Let $x = \frac{\lambda_{r1}}{\lambda_{k1}}$. Now (126) is equivalent to

$$1 - x + x^{s^*} \log x = 0.$$

Hence

$$s^*(x) = \frac{\log((x-1)/\log x)}{\log x}.$$

We extend the domain of $s^*(x)$ to 1 by defining $s^*(1) = \frac{1}{2}$, then $s^*(x)$ is an continuous increasing function on $(0, \infty)$. Since $\frac{\lambda_{r1}}{\lambda_{k1}} \in [1/\omega, \omega]$, we have $s^* \in [s^*(1/\omega), s^*(\omega)] \subset (0,1)$. One can observe that $s^*(x) = 1 - s^*(1/x)$, we have $s^* \in [s^*(1/\omega), 1 - s^*(1/\omega)]$. One can also observe that $s^*(x) \geq \frac{x}{2}$ for $x \in [0,1]$, so $s^* \in [\frac{1}{2\omega}, 1 - \frac{1}{2\omega}]$.

Now suppose $L > 1$, let s_ℓ^* be the optimizer of $f_\ell(s) = (s-1)\lambda_{k\ell} - s\lambda_{r\ell} + \lambda_{k\ell}^{1-s}\lambda_{r\ell}^s$, we still have $s^* \in [\frac{1}{2\omega}, 1 - \frac{1}{2\omega}]$. The optimizer s^* of $f(s) = \sum_{\ell=1}^L f_\ell(s)$ satisfies $s^* \in [\frac{1}{2\omega}, 1 - \frac{1}{2\omega}]$ because $f_\ell(s)$ is concave for every $\ell \in [L]$. □

Proof of Lemma 18. We first note the following Laurent series:

$$\frac{1}{1-(1+x)^{-r}} = \frac{1}{rx} + \frac{r+1}{2r} + \frac{r^2-1}{12r}x - O(x^2), \quad \text{as } x \to 0$$

from which we get the inequality

$$\frac{1}{1-(1+x)^{-r}} \leq \frac{r^{-1}}{x} + \frac{1+r^{-1}}{2}, \quad \text{for } x > 0,\ r < 1.$$

Let $\varepsilon = \varepsilon_{01}$ and $\alpha = \alpha_{01}$. Applying this inequality with $r = 1/(2\omega)$ and $x = \varepsilon$, and recalling $\alpha = \frac{1}{2\omega}\log(1+\varepsilon)$, we have

$$C(\alpha) \leq \frac{2}{1-e^{-\alpha}} = \frac{2}{1-(1+\varepsilon)^{-1/(2\omega)}} = 2\frac{2\omega}{\varepsilon} + (1+2\omega).$$

Using $1 \leq \omega$ completes the proof. □

Proof of Lemma 19. We have

$$e^{\lambda}\|\pi(\,\cdot\,;\log\lambda)\|_\infty = \sup_{t\,\in\,\mathbb{Z}_+}\frac{\lambda^t}{t!} \leq \sup_{t\,\in\,\mathbb{R}_+}\frac{\lambda^t}{\sqrt{2\pi\lambda}(t/e)^t} = \frac{e^{\lambda}}{\sqrt{2\pi\lambda}},$$

where the first inequality is by Stirling's approximation and the last equality is by plugging in the maximizer $t = \lambda$. □

Proof of Lemma 20. There exists $\ell \in L$ such that

$$(1-s^*)\lambda_{k\ell} + s^*\lambda_{r\ell} - \lambda_{k\ell}^{1-s^*}\lambda_{r\ell}^{s^*} \geq \frac{I_{kr}}{L}$$

Without loss of generality, we assume $\lambda_{k\ell} < \lambda_{r\ell}$. Let s^* be the optimizer of I_{kr}. Dividing λ_{k1} both side, we have

$$1 - s^* + s^*\frac{\lambda_{r\ell}}{\lambda_{k\ell}} - \Big(\frac{\lambda_{r\ell}}{\lambda_{k\ell}}\Big)^{s^*} \geq \frac{I_{kr}}{L\lambda_{k\ell}} \geq \frac{I_{kr}}{L\|\Lambda\|_\infty}$$

Let us define$f(x) := 1 - s^* + s^*x - x^{s^*}$, then for $x > 1$,

$$f(x) \leq \frac{1}{2}(1-s^*)s^*(x-1)^2 \leq \frac{1}{8}(x-1)^2$$

Thus $f(x) \geq \frac{I_{kr}}{L\|\Lambda\|_\infty}$ implies $x \geq \sqrt{1+\frac{8I_{kr}}{L\|\Lambda\|_\infty}}$, or equivalently, $\log x \geq \frac{1}{2}\log\Big(1+\frac{8I_{kr}}{L\|\Lambda\|_\infty}\Big)$. □

Proof of Lemma 21. Without loss of generality, we assume $\lambda_{01}/\lambda_{11} = 1+\varepsilon_{01}$. Letting $\rho = \delta\|\Lambda\|_\infty$, we have $\|\Lambda' - \Lambda\|_\infty \leq \rho$ by definition. Let $f(x) = (\lambda_{01}-x)/(\lambda_{11}+x)$. Then $f(x)$ is convex on $(0,\infty)$ with derivative $f'(x) = -(\lambda_{01}+\lambda_{11})/(\lambda_{11}+x)^2$, hence

$$\begin{aligned}\frac{\lambda'_{01}}{\lambda'_{11}} \geq \frac{\lambda_{01}-\rho}{\lambda_{11}+\rho} &= f(\rho) \geq f(0) + \rho f'(0)\\ &= \frac{\lambda_{01}}{\lambda_{11}} - \frac{\lambda_{01}+\lambda_{11}}{\lambda_{11}^2}\rho = 1+\varepsilon_{01} - \frac{\lambda_{01}+\lambda_{11}}{\lambda_{11}^2}\rho.\end{aligned}$$

Combined with

$$\frac{\rho}{\lambda_{11}} = \frac{\delta\|\Lambda\|_\infty}{\lambda_{11}} \leq \omega\delta \quad \text{and} \quad \frac{\lambda_{01}+\lambda_{11}}{\lambda_{11}} = 2+\varepsilon_{01} \leq 2(1+\varepsilon_{01}) \tag{127}$$

we have $\lambda'_{01}/\lambda'_{11} \geq 1+\varepsilon_{01} - 2\omega(1+\varepsilon_{01})\delta$ which gives the desired result. □

Proof of Lemma 22. Let $X_j \sim \text{Poi}(p_j)$ independent over $j = 1, \dots, n$, so that $\sum_{j=1}^n X_j \sim \text{Poi}(\lambda)$. Fix $x \in \mathbb{Z}_+$ and let $\mathcal{S}(x) = \{S \subset [n] : |S| = x\}$. For any subset S of $[n]$ and vectors $\alpha, \beta \in \mathbb{R}^n_+$, let $\psi(\alpha, \beta, S) = \prod_{j\in S} \alpha_j \prod_{j \notin S} \beta_j$. We have

$$\begin{aligned}\varphi(x;\lambda) &= \mathbb{P}\Big(\sum_j X_j = x\Big) \\ &\geq \mathbb{P}\Big(\sum_j X_j = x,\ X_j \in \{0,1\},\ \forall j \in [n]\Big) \\ &= \sum_{S\,\in\,\mathcal{S}(x)} \Big[\prod_{j\in S} \mathbb{P}(X_j = 1) \prod_{j \notin S} \mathbb{P}(X_j = 0)\Big] = \sum_{S\,\in\,\mathcal{S}(x)} \psi\big((p_j e^{-p_j}), (e^{-p_j}), S\big).\end{aligned}$$

On the other hand $\widetilde{\varphi}(x;p) = \sum_{S\,\in\,\mathcal{S}(x)} \psi((p_j), (1-p_j), S)$. Thus,

$$\begin{aligned}\frac{\widetilde{\varphi}(x;p)}{\varphi(x;\lambda)} \leq \frac{\sum_{S\,\in\,\mathcal{S}(x)} \psi\big((p_j), (1-p_j), S\big)}{\sum_{S\,\in\,\mathcal{S}(x)} \psi\big((p_j e^{-p_j}), (e^{-p_j}), S\big)} &\leq \max_{S\in\mathcal{S}(x)} \frac{\psi\big((p_j), (1-p_j), S\big)}{\psi\big((p_j e^{-p_j}), (e^{-p_j}), S\big)} \\ &= \max_{S\in\mathcal{S}(x)} \psi\big((e^{p_j}), ((1-p_j)e^{p_j}), S\big)\end{aligned}$$

using $(\sum a_i)/(\sum_i b_i) \leq \max(a_i/b_i)$ which holds assuming the sums have equal number of terms, all of which positive. Using $(1-x)e^x \leq 1$, It follows that

$$\frac{\widetilde{\varphi}(x;p)}{\varphi(x;\lambda)} \leq \max_{S\in\mathcal{S}(x)} \psi\big((e^{p_j}), (1), S\big) = \max_{S\in\mathcal{S}(x)} \prod_{j\in S} e^{p_j} \leq e^{xp^*}.$$

□

Proof of Lemma 23. We have

$$\frac{\phi(x;\lambda_1)}{\phi(x;\lambda_2)} = \Big(\frac{\lambda_1}{\lambda_2}\Big)^x e^{\lambda_2 - \lambda_1} \leq \Big(\frac{\lambda+\rho}{\lambda-\rho}\Big)^x e^{2\rho} = \Big(1 + \frac{2\rho}{\lambda-\rho}\Big)^x e^{2\rho} \leq \exp\Big(\frac{2\rho x}{\lambda-\rho} + 2\rho\Big).$$

Since $\lambda - \rho \geq \frac{2}{3}\lambda$ by assumption, the result follows. □

Proof of Lemma 24. Let row node i belong to row cluster k, and let $b_{i+} = \sum_{\ell\in[L]} b_{i\ell} = \sum_{j=1}^m A_{ij}$ be its degree, with expectation $\lambda_{k+} := \sum_{\ell\in[L]} \lambda_{k\ell}$. By definition, we have $\lambda_{k+} \leq L\|\Lambda\|_\infty$. We would like to find an upper bound on the probability

$$\mathbb{P}\big(\, b_{i+} > 5L\|\Lambda\|_\infty \,\big) \leq \mathbb{P}\big(\, b_{i+} - \lambda_{k+} > 4L\|\Lambda\|_\infty \,\big)$$

We let $v = \lambda_{k+}$, $vt = 4L\|\Lambda\|_\infty$, so $t \geq 4$. By Proposition 2, we have

$$\mathbb{P}\big(b_{i+} - \lambda_{k+} > 4L\|\Lambda\|_\infty\big) \leq \exp\Big[-\frac{3}{4} vt \log\Big(1 + \frac{2t}{3}\Big)\Big] \leq \exp\Big(-\frac{3}{4}vt\Big) \leq \exp(-3L\|\Lambda\|_\infty).$$

□

G.5. Proof of Lemma 2(a)

Proof of Lemma 2(a). Fix $\tilde{z}$ and let $b_{i*} = b_{i*}(\tilde{z})$. For $r \neq k \in [K]$ and i such that $y_i = k$, and $\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta)$,

$$Y_{ikr}(b_{i*}, \tilde{\Lambda}) = \sum_{\ell=1}^{L} \Big[b_{i\ell} \log \frac{\tilde{\lambda}_{r\ell}}{\tilde{\lambda}_{k\ell}} + \tilde{\lambda}_{k\ell} - \tilde{\lambda}_{r\ell} \Big] \;\le\; \sum_{\ell=1}^{L} \Big[b_{i\ell} \log \frac{\lambda_{r\ell} + \rho}{\lambda_{k\ell} - \rho} + \lambda_{k\ell} - \lambda_{r\ell} + 2\rho \Big] \;:=\; Y^*$$

where $\rho := \delta \|\Lambda\|_\infty$ is the radius of $\mathscr{B}_\Lambda(\delta)$. Hence $\mathbb{P}(\exists \tilde{\Lambda} \in \mathscr{B}_\Lambda, Y_{ikr} \ge 0) \le \mathbb{P}(Y^* \ge 0)$. By Markov inequality, we have $\mathbb{P}(Y^* \ge 0) \le \mathbb{E}[e^{sY^*}]$ for any $s \ge 0$. To simplify the notation, let us write $v_\ell = s \log[(\lambda_{r\ell} + \rho)/(\lambda_{k\ell} - \rho)]$ and $w_\ell = s(\lambda_{k\ell} - \lambda_{r\ell} + 2\rho)$, so that $sY^* = \sum_{\ell=1}^{L} b_{i\ell}\, v_\ell + w_\ell$. By independence, we have

$$\log \mathbb{E}[e^{sY_*}] = \log \mathbb{E}\Big[\prod_{\ell=1}^{L} e^{b_{i\ell}\, v_\ell + w_\ell} \Big] = \sum_{\ell=1}^{L} \log \mathbb{E}[e^{b_{i\ell}\, v_\ell + w_\ell}] = \sum_{\ell=1}^{L} \log \big[e^{w_\ell} \mathbb{E} e^{b_{i\ell} v_\ell} \big].$$

Since the mgf of a Poisson-binomial variable is bounded above by that of a Poisson variable with the same mean,

$$\log \mathbb{E} e^{sY_{ikr}} \le \sum_{\ell=1}^{L} w_\ell + \psi\big(v_\ell, \lambda_{k\ell}(y, \tilde{z})\big)$$

where $\psi(t, \mu) = \mu(e^t - 1)$ is the log-mgf of a $\text{Poi}(\mu)$ random variable. Recalling the assumption $\Lambda(y, \tilde{z}) \in \mathscr{B}_\Lambda$, we have

$$\sum_{\ell=1}^{L} w_\ell + \psi\big(v_\ell, \lambda_{k\ell}(y, \tilde{z})\big) = \sum_{\ell=1}^{L} \Big[\lambda_{k\ell}(y, \tilde{z}) \left(\frac{\lambda_{r\ell} + \rho}{\lambda_{k\ell} - \rho} \right)^s - \lambda_{k\ell}(y, \tilde{z}) + s(\lambda_{k\ell} - \lambda_{r\ell} + 2\rho) \Big].$$

Since $\Lambda(y, \tilde{z}) \in \mathscr{B}_\Lambda(\delta)$, $\lambda_{k\ell}(y, \tilde{z}) \le \lambda_{k\ell} + \delta \|\Lambda\|_\infty = \lambda_{k\ell} + \rho$. Since $\lambda_{k\ell} - \rho = \lambda_{k\ell} - \delta \|\Lambda\|_\infty \ge \lambda_{k\ell} - \frac{\|\Lambda\|_\infty}{3\omega} \ge \lambda_{k\ell} - \frac{1}{3}\Lambda_{\min} = \frac{2}{3}\lambda_{k\ell}$,

$$\begin{aligned}
\lambda_{k\ell} \left(\frac{\lambda_{r\ell} + \rho}{\lambda_{k\ell} - \rho} \right)^s &= \lambda_{k\ell} \left(\frac{\lambda_{r\ell} - \rho\frac{\lambda_{r\ell}}{\lambda_{k\ell}} + (1 + \frac{\lambda_{r\ell}}{\lambda_{k\ell}})\rho}{\lambda_{k\ell} - \rho} \right)^s \\
&\le \lambda_{k\ell} \left[\Big(\frac{\lambda_{r\ell}}{\lambda_{k\ell}} \Big)^s + s \Big(\frac{\lambda_{r\ell}}{\lambda_{k\ell}} \Big)^{s-1} \Big(\frac{(1 + \frac{\lambda_{r\ell}}{\lambda_{k\ell}})\rho}{\lambda_{k\ell} - \rho} \Big) \right] \\
&\le \lambda_{k\ell} \Big(\frac{\lambda_{r\ell}}{\lambda_{k\ell}} \Big)^s + \frac{3}{2} s \cdot 2\omega\rho \\
&\le \lambda_{k\ell} \Big(\frac{\lambda_{r\ell}}{\lambda_{k\ell}} \Big)^s + 3\omega\rho.
\end{aligned}$$

Moreover, $\lambda_{r\ell} + \rho = \lambda_{r\ell} + \delta \|\Lambda\|_\infty = \lambda_{r\ell} + \omega\delta\Lambda_{\min} \le \lambda_{r\ell} + \frac{1}{3}\Lambda_{\min} \le \frac{4}{3}\lambda_{r\ell}$. Thus, we have

$$\rho \left(\frac{\lambda_{r\ell} + \rho}{\lambda_{k\ell} - \rho} \right)^s \le \rho \Big(\frac{\frac{4}{3}\lambda_{r\ell}}{\frac{2}{3}\lambda_{k\ell}} \Big)^s \le 2\omega\rho.$$

Taking infimum over $s > 0$, by Lemma 17, the maximizer s^* of I_{kr} is always bounded between 0 and 1, hence we have

$$\begin{aligned}\mathbb{P}\big(\exists\tilde{\Lambda} \in \mathscr{B}_\Lambda,\ Y_{ikr}(b_{i*}, \tilde{\Lambda}) \geq 0\big) &\leq P(Y_* \geq 0)\\ &\leq \exp\big(-I_{kr} + 8L\omega\delta\|\Lambda\|_\infty\big) = \exp\big(-(1-\eta')I_{kr}\big).\end{aligned}$$

□

G.6. Proofs of Section 3

Proof of Proposition 1. The upper bound has been provided by Corollary 6. Here we will show the lower bound, using the notation established in the proof of Proposition 3 and Appendix F.3.1. We rename λ_{k*} and λ_{r*}, and work with λ_{0*} and λ_{1*} instead, and we assume throughout that, $\lambda_{0\ell}, \lambda_{1\ell} \geq 1$ for all $\ell \in [L]$. We recall from (123) that

$$P_{e,+} = \int \min(p_0, p_1) d\mu = e^{-I_s} \int p_s \min(l^s, l^{s-1}) d\mu, \tag{128}$$

where p_s is defined in (122) and l in (124). Since μ is the counting measure, we have

$$P_{e,+} \geq \max_{x \in \mathbb{Z}_+^L} \min(p_0(x), p_1(x)) = \max_{x \in \mathbb{Z}_+^L} e^{-I_s} p_s(x) \min(l^s(x), l^{s-1}(x)). \tag{129}$$

Finding the maximizer x over $\mathbb{Z}_+$ gives the lower bound. First, let us extend the Poisson density as $\phi(t;\lambda) = \lambda^t e^{-\lambda}/\Gamma(t+1)$ to any $t \in \mathbb{R}_+$, so that l is well-defined on $\mathbb{R}_+^L$, given by

$$l(x) = \exp\Big(\sum_{\ell \in [L]} x \log\frac{\lambda_{0\ell}}{\lambda_{1\ell}} - \lambda_{0\ell} + \lambda_{1\ell}\Big),\ x \in \mathbb{R}_+^L.$$

Recall that $\lambda_{s\ell} = \lambda_{0\ell}^{1-s}\lambda_{1\ell}^s$, $\lambda_s = (\lambda_{s\ell})$ and $I_s = \sum_\ell [(1-s)\lambda_{0\ell} + s\lambda_{1\ell} - \lambda_{s\ell}]$ (cf. Appendix F.3.1). We note that

$$\frac{dI_s}{ds} = \sum_{\ell \in [L]} -\lambda_{0\ell} + \lambda_{1\ell} + \lambda_{s\ell} \log\Big(\frac{\lambda_{0\ell}}{\lambda_{1\ell}}\Big). = \log(l(\lambda_s))$$

The function $s \mapsto I_s$ is concave, smooth, nonconstant (by assumption) and we have $I_0 = I_1 = 0$. Hence, the unique maximizer s^* of $s \mapsto I_s$ belongs to $(0,1)$ and satisfies $dI_s/ds\big|_{s^*} = 0$, that is, $\log(l(\lambda_{s^*})) = 0$, or equivalently $p_0(\lambda_{s^*})/p_1(\lambda_{s^*}) = l(\lambda_{s^*}) = 1$. By the definition of p_s, we have

$$e^{-I_{s^*}} p_{s^*}(\lambda_{s^*}) = p_0^{1-s}(\lambda_{s^*}) p_1^s(\lambda_{s^*}) = p_0(\lambda_{s^*}) = p_1(\lambda_{s^*}).$$

We recall that p_s is the product of Poisson densities with parameters $\lambda_{s\ell}$. By a version of the Stirling's inequality for the Gamma functions (Jameson, 2015):

$$\Gamma(x+1) = x\Gamma(x) \leq (2\pi)^{1/2} x^{x+1/2} e^{-x} e^{1/(12x)}, \forall x > 0$$

hence $\Gamma(x+1) \leq C_0\, x^{x+1/2} e^{-x}$ for all $x \geq 1$, where $C_0 = (2\pi)^{1/2} e^{1/12}$. Then,

$$\phi(\lambda;\lambda) = \frac{\lambda^\lambda e^{-\lambda}}{\Gamma(\lambda+1)} \geq C_0^{-1} \lambda^{-1/2},$$

from which it follows that

$$p_{s^*}(\lambda_{s^*}) = \prod_{\ell \in L} \phi(\lambda_{s^*\ell}; \lambda_{s^*\ell}) \geq C_0^{-L} \prod_{\ell \in [L]} \lambda_{s^*\ell}^{-1/2}.$$

Thus, $e^{-I_{s^*}} C_0^{-L} \prod_\ell \lambda_{s^*}^{-1/2}$ is a lower bound on $P_{e,+}$ whenever $\lambda_{s^*} \in \mathbb{Z}_+^L$. In general, λ_{s^*} does not have integer coordinates. Instead, pick any $x \in \mathbb{Z}_+^L$ satisfying $\|x - \lambda_{s^*}\|_{\ell_\infty} \leq 1$.

Since $t \mapsto \phi(t; \lambda)$ is a quasi-concave function (i.e., upper-level sets are convex), we have $\phi(t;\lambda) \geq \min\{\phi(a;\lambda), \phi(b;\lambda)\}$ for every $t \in [a,b]$, hence, for every $t \in [a-1, a+1]$, we obtain using $\Gamma(x+1) = x\Gamma(x)$,

$$\phi(t;\lambda) \geq e^{-\lambda} \min\Big\{\frac{\lambda^{a-1}}{\Gamma(a)}, \frac{\lambda^{a+1}}{\Gamma(a+2)}\Big\} = \frac{e^{-\lambda}\lambda^a}{\Gamma(a+1)} \min\Big\{\frac{a}{\lambda}, \frac{\lambda}{a+1}\Big\},$$

that is,

$$\frac{\phi(t;\lambda)}{\phi(a;\lambda)} \geq \min\Big\{\frac{a}{\lambda}, \frac{\lambda}{a+1}\Big\}, \quad t \in [a-1, a+1].$$

Since $|x_\ell - \lambda_{s^*\ell}| \leq 1$,

$$p_{0\ell}(x_\ell) \;\geq\; p_{0\ell}(\lambda_{s^*\ell}) \min\Big\{\frac{\lambda_{s^*\ell}}{\lambda_{0\ell}}, \frac{\lambda_{0\ell}}{\lambda_{s^*\ell}+1}\Big\} \;\geq\; (2\omega)^{-1} p_{0\ell}(\lambda_{s^*\ell})$$

where we have used, for any $s \in [0,1]$,

$$\min\Big\{\frac{\lambda_{s\ell}}{\lambda_{0\ell}}, \frac{\lambda_{0\ell}}{\lambda_{s\ell}}\Big\} = \Big(\min\Big\{\frac{\lambda_{1\ell}}{\lambda_{0\ell}}, \frac{\lambda_{0\ell}}{\lambda_{1\ell}}\Big\}\Big)^s \geq \Big(\frac{1}{\omega}\Big)^s \geq \frac{1}{\omega}$$

and $\lambda_{s^*\ell}/(\lambda_{s^*\ell}+1) \geq 1/2$ since $\lambda_{s^*\ell} \geq 1$. Similarly $p_{1\ell}(x_\ell) \geq p_{1\ell}(\lambda_{s^*\ell})/(2\omega)$. Hence,

$$\sum_{x \in \mathbb{Z}_+^L} \min\{p_0(x), p_1(x)\} \geq \frac{\min\{p_0(\lambda_{s^*}), p_1(\lambda_{s^*})\}}{(2\omega)^L} = \frac{e^{-I_{s^*}} p_{s^*}(\lambda_{s^*})}{(2\omega)^L} \geq \frac{e^{-I_{s^*}}}{(2C_0\omega)^L} \prod_{\ell \in [L]} \lambda_{s^*}^{-1/2}$$

where we have used $\min\{p_0(x), p_1(x)\} = e^{-I_s} p_s(x) \min\{l(x)^s, l(x)^{s-1}\}$ and $l(\lambda_{s^*}) = 1$. Thus,

$$\begin{aligned} P_{e,+} &\geq \exp\Big(-I_{s^*} - L\log(2C_0\omega) - \frac{L}{2}\log\|\Lambda\|_\infty\Big) \\ &\geq \exp\Big(-I_{s^*} - L\log(2C_0\omega^{3/2}) - \frac{L}{2}\log\Lambda_{\min}\Big) \end{aligned} \tag{130}$$

using the assumption $\|\Lambda\|_\infty \leq \omega\Lambda_{\min}$. The proof is complete. □

G.7. Auxiliary lemmas for Theorem 2

The following lemmas are used in the proof of the minimax Theorem 2.

Lemma 25. *For a discrete probability distribution $\mathcal{L}$ on $\mathbb{Z}_+$, let us write* $\mathrm{pmf}(x;\mathcal{L}), x \in \mathbb{Z}_+$ *for the probability mass function of $\mathcal{L}$. There is a universal constant $c > 0$, such that for $\omega > 1$,*

$$\mathrm{pmf}\Big(x; \mathrm{Bin}\Big(n, \frac{\lambda}{n}\Big)\Big) \geq c\,\mathrm{pmf}\Big(x; \mathrm{Poi}(\lambda)\Big), \quad \forall x \leq 2\omega\lambda \leq \sqrt{n}/3. \tag{131}$$

Proof. Let $\widetilde{p}$ be the pmf of the Binomial distribution in (131). By the Stirling's approximation,

$$\begin{aligned}\frac{n!}{(n-x)!} &\geq \frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi(n-x)}[(n-x)/e]^{n-x}} \cdot \frac{e^{1/(12n+1)}}{e^{1/(12(n-x))}} \\ &\geq c_1\sqrt{\frac{n}{n-x}}n^n(n-x)^{-(n-x)}e^{-x}\end{aligned}$$

where we have used $x \leq 2n/3$ to bound the second factor by c_1 from below. Hence,

$$\begin{aligned}\widetilde{p}(x) = \frac{n!}{x!(n-x)!}\Big(\frac{\lambda}{n}\Big)^x\Big(1-\frac{\lambda}{n}\Big)^{n-x}\Big(\frac{n-\lambda}{n-x}\Big)^{n-x} &= \frac{\lambda^x}{x!}\frac{n!}{(n-x)!}n^{-n}(n-\lambda)^{n-x} \\ &\geq c_1\frac{\lambda^x e^{-x}}{x!}\sqrt{\frac{n}{n-x}}\Big(\frac{n-\lambda}{n-x}\Big)^{n-x}.\end{aligned}$$

We have $\sqrt{n/(n-x)} \geq 1$. Using the inequality $1+t \geq (1-t^2)e^t$ for $|t| \leq 1$,

$$\Big(\frac{n-\lambda}{n-x}\Big)^{n-x} = \Big(1-\frac{\lambda-x}{n-x}\Big)^{n-x} \geq \Big[1-\Big(\frac{\lambda-x}{n-x}\Big)^2\Big]^{n-x}e^{x-\lambda}.$$

Again by $0 \leq x \leq 2\omega\lambda \leq \sqrt{n}/3 \leq n/3$ and using $(1-x)^n \geq 1-nx$, we have

$$\Big[1-\Big(\frac{\lambda-x}{n-x}\Big)^2\Big]^{n-x} \geq 1-\frac{(\lambda-x)^2}{n-x} \geq 1-\frac{(2\omega\lambda)^2}{2n/3} \geq 1-\frac{n/9}{2n/3} = \frac{5}{6}.$$

It follows that $\widetilde{p}(x) \geq c_2\,\lambda^x e^{-\lambda}/x!$ which is the desired result. □

Lemma 26. *Let $\widetilde{p}_{k\ell}$ be the probability mass function of $\mathrm{Poi}(\lambda_{k\ell})$ and $p_{k\ell}$ that of $\mathrm{Bin}(n_\ell, \lambda_{k\ell}/n_\ell)$. Let $\widetilde{p}_k = \bigotimes_{\ell=1}^L \widetilde{p}_{k\ell}$ and $p_k = \bigotimes_{\ell=1}^L p_{k\ell}$ and similarly define $\widetilde{p}_r$ and p_r. Assume that*

$$\max(\lambda_{k\ell}, \lambda_{r\ell}) \;\leq\; \min\big(\omega\lambda_{k\ell},\, \omega\lambda_{r\ell},\, \sqrt{n_\ell}/3\big), \quad \forall \ell \in [L]$$

for some $\omega > 1$. Then, there is a universal constant $C > 0$ such that the sum of the type I and type II errors of the likelihood ratio test for p_k against p_r satisfies

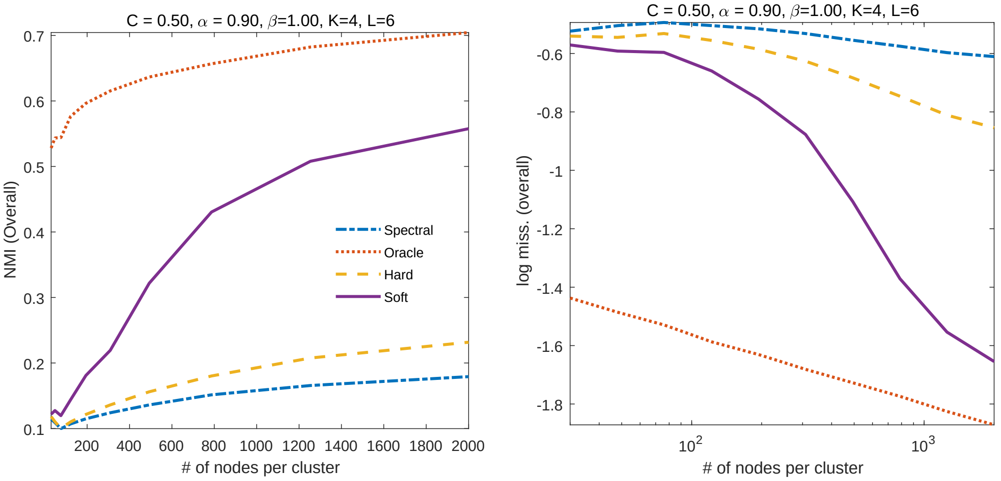
$$P_{e,+} \geq \exp\Big(-I_{s^*} - L\log(C\omega^{3/2}) - \frac{L}{2}\log\Lambda_{\min}\Big),$$

where I_{s^*} is the information between λ_{k*} and $\lambda_{\ell*}$ defined in (8) *and $\Lambda_{\min} = \min_\ell(\lambda_{k\ell}, \lambda_{r\ell})$.*

Proof. For $x \in \mathbb{Z}_+^L$ satisfying $\|x-\lambda_{s^*}\|_\infty \leq 1$, we have $x_\ell \leq \lambda_{s^*\ell}+1 \leq 2\omega\min(\lambda_{k\ell}, \lambda_{r\ell})$. Then by Lemma 25, $p_k(x) \geq c^L\widetilde{p}_k(x)$ and $p_r(x) \geq c^L\widetilde{p}_r(x)$. Therefore,

$$\begin{aligned}P_{e,+} \;&\geq\; \max_{x\in\mathbb{Z}_+^L}\min\big(p_k(x), p_r(x)\big) \;\geq\; c^L \max_{x:\,\|x-\lambda_{s^*}\|_\infty\leq 1}\min\big(\widetilde{p}_k(x), \widetilde{p}_r(x)\big) \\ &\geq c^L\exp\Big(-I_{s^*} - L\log(2C_0\omega^{3/2}) - \frac{L}{2}\log\Lambda_{\min}\Big) \\ &\geq \exp\Big(-I_{s^*} - L\log(C\omega^{3/2}) - \frac{L}{2}\log\Lambda_{\min}\Big),\end{aligned}$$

where the third inequality is by (130) and C and C_0 are positive universal constants. The proof is complete. □

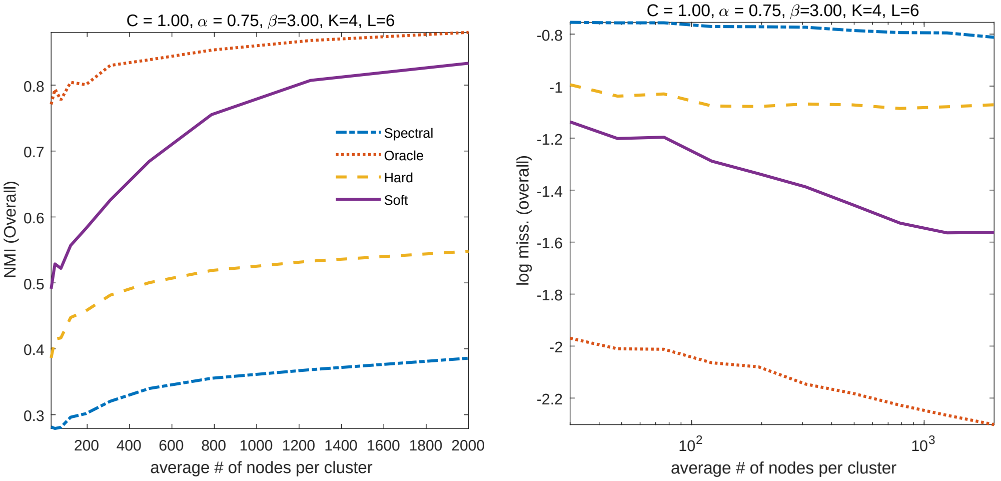


Appendix H. Extra Simulation Results

Here we present extra simulation results under the setup of Section 6. The following figure shows the overall NMI and log. error rate for different values of C and α: The next figure illustrates the results for unbalanced cluster sizes. To be specific,

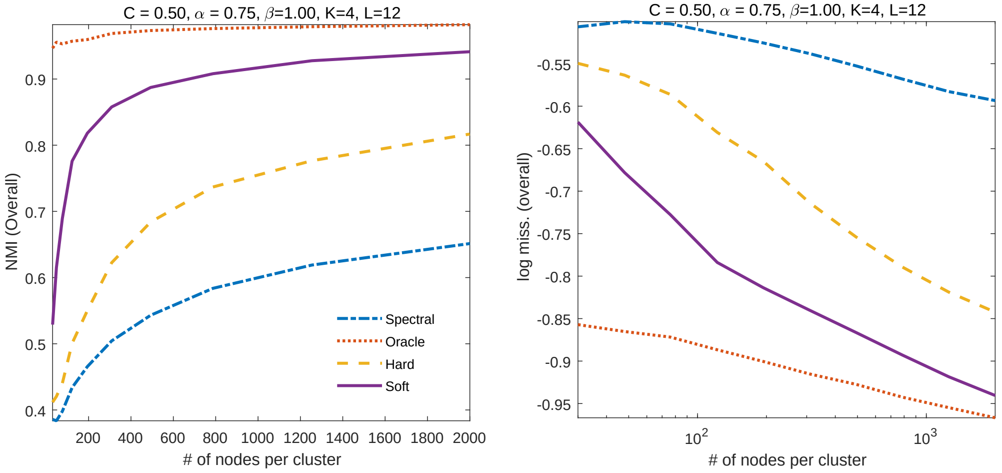
$$\pi(y) = \frac{(1,4,6,9)}{20} \quad \text{and} \quad \pi(z) = \frac{(1,3,4,6,7,9)}{30},$$

which implies $\beta \geq 3$ according to (A2): We also consider the setting where the number of the clusters of one side is significantly greater that of the other. We let $K = 4$, $L = 12$ and



$$B = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 1 & 2 & 3 \\ 7 & 8 & 9 & 10 & 11 & 12 & 1 & 2 & 3 & 4 & 5 & 6 \\ 10 & 11 & 12 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix}.$$

The simulation results are as follows:



References

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

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

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

Naman Agarwal, Afonso S Bandeira, Konstantinos Koiliaris, and Alexandra Kolla. Multisection in the stochastic block model using semidefinite programming. In *Compressed Sensing and its Applications*, pages 125–162. Springer, 2017.

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

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

Afonso S Bandeira. Random laplacian matrices and convex relaxations. *Foundations of Computational Mathematics*, pages 1–35, 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.

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

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

Rainer E Burkard and Eranda Cela. Linear assignment problems and extensions. In *Handbook of combinatorial optimization*, pages 75–149. Springer, 1999.

Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6: 1847–1899, 2012.

Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.

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

I Chien, Chung-Yi Lin, and I-Hsiang Wang. Community detection in hypergraphs: Optimal statistical limit and efficient algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 871–879, 2018.

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

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

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

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

Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. pages 269–274, 2001.

Inderjit S Dhillon. Information-Theoretic Co-clustering. pages 89–98, 2003.

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

Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

Chao Gao, Yu Lu, Zongming Ma, and Harrison H Zhou. Optimal estimation and completion of matrices with biclustering structures. *Journal of Machine Learning Research*, 17(161):1–29, 2016.

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

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

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

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

Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49(3):686–721, 2017.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016a.

Bruce Hajek, Yihong Wu, and Jiaming Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Transactions on Information Theory*, 62(10): 5918–5937, 2016b.

John A Hartigan. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129, 1972.

Joseph L Hodges and Lucien Le Cam. The poisson approximation to the poisson binomial distribution. *The Annals of Mathematical Statistics*, 31(3):737–740, 1960.

G. J. O. Jameson. A simple proof of stirling's formula for the gamma function. *The Mathematical Gazette*, 99(544):68, 2015.

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

Daniel B Larremore, Aaron Clauset, and Abigail Z Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

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

Sara C Madeira, Miguel C Teixeira, Isabel Sa-Correia, and Arlindo L Oliveira. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 7(1): 153–165, 2010.

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

Andrea Montanari and Subhabrata Sen. Semidefinite programs on sparse random graphs and their application to community detection. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 814–827. ACM, 2016.

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

Krzysztof Nowicki and Tom A B Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

Marianna Pensky, Teng Zhang, et al. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709, 2019.

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

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

Federico Ricci-Tersenghi, Adel Javanmard, and Andrea Montanari. Performance of a community detection algorithm based on semidefinite programming. In *Journal of Physics: Conference Series*, volume 699, page 012015. IOP Publishing, 2016.

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

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering for directed graphs: the stochastic co-blockmodel and spectral algorithm di-sim. *arXiv preprint arXiv:1204.2296*, 2012.

Amos Tanay, Roded Sharan, and Ron Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl_1):S136–S144, 2002.

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

Van Vu. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

Jason Wyse, Nial Friel, and Pierre Latouche. Inferring structure in bipartite networks using the latent blockmodel and exact icl. *Network Science*, 5(1):45–69, 2017.

Min Xu, Varun Jog, and Po-Ling Loh. Optimal rates for community estimation in the weighted stochastic block model. *arXiv preprint arXiv:1706.01175*, 2017.

Se-Young Yun and Alexandre Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014.

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.

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

Yunpeng Zhao, Elizaveta Levina, Ji Zhu, et al. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. *Physical review E*, 76(4):046115, 2007.

Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47):1–47, 2019.

Zhixin Zhou and Ping Li. Non-asymptotic chernoff lower bound and its application to community detection in stochastic block model. *arXiv preprint arXiv:1812.11269*, 2018.