UNIVERSITY OF CALIFORNIA

Los Angeles

Optimal Bipartite Network Clustering

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Zhixin Zhou

2018

ABSTRACT OF THE DISSERTATION

Optimal Bipartite Network Clustering

by

Zhixin Zhou
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2018
Professor Arash Ali Amini, Chair

We consider the problem of 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. The optimal rate we obtain sharpens some of the existing results and generalizes others to a wide regime of average degree growth. As a special case, we recover the known exact recovery threshold in the $\log n$ regime of sparsity. To obtain the general consistency result, as part of the provable version of the algorithm, we introduce a block partitioning scheme that is also computationally attractive, allowing for distributed implementation of the algorithm without sacrificing optimality. The provable version of the algorithm is derived from a general blueprint for pseudo-likelihood biclustering algorithms that employ simple EM type updates. We show the effectiveness of this general class by numerical simulations.

The dissertation of Zhixin Zhou is approved.

Yingnian Wu

Jingyi Li

Qing Zhou

Arash Ali Amini, Committee Chair

University of California, Los Angeles

2018

To my family, especially my grandmother . . .

who have been patiently waiting for my degree in the past six years.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Arash A. Amini for his supervision and encouragement during the journey of my research. I am extremely fortunate to have him as my advisor. Besides, I would also like to express my gratitude to Professor Jingyi J. Li, Professor Yingnian Wu, and Professor Qing Zhou for their insightful comments and serving on my committee.

I would like to thank Professor Michael J. Klass and Professor Peter J. Bickel at Berkeley for their enlightening discussion.

I would like to thank the staffs of our department, especially Glenda Jones. I would also like to thank my friends at UCLA, especially Spencer Frei, Shihao Gu, Jiayin Guo, Terri A. Johnson, Wei Li, Jianfeng Lin, Seunghyun Min, James D. Molyneux, Junhyung Park, Yidan Sun, Medha Uppala, Lin Wang, Nan Xi, Fei Xie, Qian Xiao, Qiaoling Ye, Yunfeng Zhang, Kun Zhou, Jiaping Zhu and Wei Zhu.

Finally, I am very grateful for the encouragement from Mayu Matsuoka's dramas, movies and variety shows, as well as 9nine's music.

VITA

2008-2012	B.A. in Applied Mathematics, University of California, Berkeley, California.
2012-2015	M.A. in Mathematics, University of California, Los Angeles, California.
2015-present	Ph.D. Candidate in Statistics, University of California, Los Angeles, California.

CHAPTER 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 [FH16]. In particular, there has been much excitement and progress in understanding and analyzing the stochastic block model (SBM) and its variants. We refer to [Abb17] for a recent survey of the field. Much of this effort, especially on the theoretical side has been focused on the univariate (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 [ZRMZ07; LCJ14; WFL14; Roh15; RAL17] as well as other domains, such as topic modeling and text mining [Dhi01; Dhi03], as well as in biological applications [CC00; MTSCO10]. But much of this work either lacks theoretical investigations or has not considered the issue of statistical optimality.

In this paper, we provide a unified treatment of 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 [Har72; CC00; TSS02; GLMZ16]. On the other hand, in recent years, many algorithms have been proposed for clustering in univariate SBMs, including global approaches such as spectral clustering [RCY11; Krz+13; LR13; Fis+13; Vu14; Mas14; YP14; BLM15; GLM17; PZ17] and convex relaxations via semidefinite programs (SDPs) [AL14; HWX16a; Ban15; GV16; MS16; RTJM16; ABKK17; PW17], as well as local methods such as belief propagation [DKMZ11], Bayesian MCMC [NS01] and variational Bayes [CDP12; BCCZ13], greedy profile likelihood [BC09; ZLZ+12] and pseudo-likelihood [ACBL+13] maximization, 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 [BC09], but the underlying optimization problem is computationally infeasible and the approach only applicable to networks of limited size. Pseudo-likelihood ideas were used in [ACBL+13] to derive EM type updates to maximize a surrogate to the likelihood of the SBM based on a block compression which is computed using initial labels obtained by spectral clustering.

The pseudo-likelihood approach belongs to the general class of algorithms based on "Good Initialization followed by Fast Local Updates" (GI-FLU) which has been a staple of recent developments in devising optimal clustering algorithms, as was pointed out by [GMZZ17]. The GI-FLU strategies often use a spectral initialization and due to their often cheap local updates are scalable to very large networks. In cases where they are accompanied by optimality guarantees they seem to occupy the ideal sweet spot in the computational versus statistical trade-off. We build on these ideas and especially the approach in [ACBL+13] to extend the algorithms to the bipartite settings. Moreover, we provide modifications to the general blueprint suggested by [ACBL+13]—in addition to the natural modification required for the bipartite setup—which allows us to demonstrate optimality of the procedure under the full generality of the bipartite SBM.

In the univariate setting, there has been interesting recent advancements in understanding

optimal recovery in what we refer to as the semi-sparse regime, where the (expected) average network degree d_{av} is allowed to grow to infinity but rather slowly, as the number of nodes n increases to infinity. In a series of papers [MNS15; ABH16; HWX16a; HWX16b] the thresholds for optimal exact recovery, also known as strong consistency, were established in the context of simple planted partition models, and SDPs where shown to achieve the optimal threshold. In [AS15], the problem of strong consistency was considered for a general SBM and the optimal threshold for strong consistency was established. In subsequent work [ZZ+16; GMZZ17; GMZZ16], 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 where established (with some slack in the exponent). To achieve the more relaxed consistency results, [GMZZ17] limited the model to what we refer to as strongly assortative SBM (see [AL18] for a definition).

Our work from a theoretical perspective is mostly inspired by the insightful work of [AS15; GMZZ17]. We extend these ideas by presenting results that are 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. Relative to [GMZZ17], our results are shaper, removing the $o(1)$ term in the exponents of the rates, and hold for the full generality of the SBM (i.e., no assortativity assumption needed). Compared to [AS15], our work greatly relaxes the assumptions on degree growth. It is well-known that for strong consistency one needs at least $d_{\text{av}} = \Omega(\log n)$, i.e., average degree should grow at least logarithmically, and this is the regime considered in [AS15]. In our work, d_{av} could grow to infinity arbitrarily slowly or at the other extreme as fast as $d_{\text{av}} = O(\sqrt{n})$. Thus, our results establish optimal rates of weak consistency below the sequare-root regime $d_{\text{av}} = O(\sqrt{n})$ and above the sparse regime $d_{\text{av}} = O(1)$, and in particular, between the $\log n$ and $O(1)$ regimes, where weak consistency is possible but not the strong consistency. In addition, in contrast to [AS15] (and some results in [GMZZ17]), we allow all the parameters of the model, including the number of communities, the spread of mean parameters (ω) and the community balance parameter (β) to grow subject to two compact conditions (namely (A3) and (A4)); these conditions encapsulate in a simple way how much cumulative growth these auxiliary parameters can exhibit relative to the information growth of the model ($I = I_{\min} \wedge I^{\text{col}}_{\min}$) for the optimal rate to still be achieved by the algorithm we present.

We make more detailed comparisons with the work of [AS15; GMZZ17] in Chapter 3.

Contributions. Establishing these results require a fair amount of technical and algorithmic novelty over the previous work. Here, we highlight some of these features and point out to the relevant parts of the paper for details:

1. We introduce an efficient block (or graph) partitioning scheme for the provable version of the algorithm, Algorithm 3, which for example replaces the edge splitting idea of [AS15] (cf. Chapter 3.1 and Remark 10 for the shortcomings of edge splitting in our setting). The block partitioning is mainly introduced to generate enough independence for the technical arguments to go through. However, the idea turns out to be computationally appealing as well. The computational bottleneck of GI-FLU approaches discussed earlier is often the spectral initialization. The subsequent (often likelihood based) local updates are usually quite fast, $O(n)$, computations. Our block partitioning scheme allows one to break the costly initial step into the application of spectral methods—as well as likelihood ratio classifiers—on smaller subblocks, without losing optimality. If done in parallel, spectral clustering on subblocks will be in fact computationally cheaper than performing a spectral decomposition of the entire matrix. Although, from a theoretical perspective breaking into $Q = 4$ blocks is enough (in the bipartite settings), our results allow for the number of blocks Q to even grow slowly to infinity. Thus, the provable version of our algorithm has computational appeal, esp. in *distributed settings* and for very large networks where it is prohibitive to perform eigendecomposition of the entire adjacency matrix. The algorithm is naturally parallelizable since it proceeds in stages and in each stage the operations on the underlying subblocks can be performed in parallel; see Chapter 6 for details.

2. Our algorithms being an extension of [ACBL+13], are modifications of a natural EM algorithm on mixtures of Poisson vectors, hence very familiar from a statistical perspective. In other words, they are not tailor-made to the community detection problem and are derived from fairly general well-known principals. Although other (optimal) algorithms in the literature are more or less preforming similar operations, the link to EM algorithms and mixture modeling is quite clear in our work. We provide in Chapter 4.2 the general blueprint of the algorithms based on the pseudo-likelihood idea and block compression (Algorithm 1). We then show how a provable version can be constructed by combining with the block par-

titioning ideas in Chapter 6. It is worth noting that although we can provide no guarantees for the general algorithms of Chapter 4.2, empirically they perform very well, as illustrated in Chapter 9.

3. In order to get the sharper rate, analyzing a single step of an EM type algorithm is not enough, and thus we analyze the second step as well. We will show that the first step gets us from a good (but crude) initial rate γ_1 to the fast rate $\approx \exp(-I/Q)$ which is in the vicinity of the optimal rate, and then repeating the iteration once more, with the more accurate labels obtained in the first step—hence more accurate parameter estimates—gets us to the optimal rate $\approx \exp(-I)$.

4. Among the technical contributions, are a uniform consistency result (Lemma 5) for the likelihood ratio classifier (LRC) over a subset of the parameters close to the truth, sharp approximations for the Poisson-binomial distributions (Chapter 10.4), and extension (and elucidation) of a novel technique of [AS15] in deriving error exponents to general exponential families (cf. Chapter 10.3). The uniform consistency result for LRCs lets us tolerate some degree of dependence among the statistics from iteration to iteration, allowing the subblock partioning idea to go through. That is, we can run LRCs on the same blocks used in estimating their parameters; see Sections 6 and 7.1 for details.

5. The bipartite clustering setup (as opposed to the symmetric case) allows us to introduce an oracle version of the problem which helps in understanding the nature of the optimal rates observed in community detection and their relation to classical hypothesis testing and mixuture modeling. That is, we try to answer the curious question of why or how the Chernoff exponent of a (simple binary) hypothesis testing problem seems to control the misclassification rate in community detection and network clustering. The oracle also provides a lower bound on the performance of any algorithm. See Chapter 3 and Proposition 1 for details.

The rest of the paper is organized as follows. We introduce the model and the biclustering oracle in Chapter 2, and then present our main results in Chapter 3. The general algorithms based on the pseduo-likelihood idea and spectral clustering are presented in Chapter 4 and Chapter 5. A provable version will be proposed in Chapter 6. The proofs of the consistency results will appear in Sections 7 and 8. In Chapter 9, we demonstrate the numerical

performance of the methods.

CHAPTER 2

Network biclustering

We start by introducing the network biclustering problem based a stochastic block modeling, and set up some notation that will be used throughout the paper. We then discuss how a biclustering oracle with side information can optimally recover the labels. These ideas will be the basis of the algorithms discussed in this paper.

2.1 Bipartite block model

We will be working with a bipartite network which can be represented by a biadjacency matrix $A \in \{0,1\}^{n\times m}$, where for simplicity we assume that 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,

$$\mathbb{E}[A_{ij}] = y_i^T P z_j = P_{k\ell}, \quad \text{if } y_i = k,\ z_j = \ell. \tag{2.1}$$

We denote this model compactly as $A \sim \mathrm{SBM}(y, z, P)$. It is helpful to consider the Poisson version of the model as well 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 mean given in (2.1). 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 sets of 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 $n(z) = (n_1(z), \ldots, n_L(z)) \in \mathbb{N}^L$, be the number of nodes in each of the communities of side 2. That is, $n_\ell(z) = \sum_{j=1}^{M} 1\{z_j = \ell\} = \sum_{j=1}^{M} z_{j\ell}$. A similar notation, namely $n(y) \in \mathbb{N}^K$, denotes the community sizes of side 2. The *row mean parameters* are defined as

$$\Lambda = (\lambda_{k\ell}) = (P_{k\ell}\, n_\ell(z)) = P\,\mathrm{diag}(n(z)) \in \mathbb{R}^{K\times L} \tag{2.2}$$

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 in a similar fashion, namely,

$$\Gamma^T = \big(n_k(y)P_{k\ell}\big) = \mathrm{diag}(n(y))P \in \mathbb{R}^{K\times L}. \tag{2.3}$$

Note the transpose in the above definition, i.e., $\Gamma \in \mathbb{R}^{L\times K}$, and 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 Chapter 3.

2.2 Biclustering oracle with side information

The key idea behind the algorithms discussed in this paper, as well as our consistency arguments is the following simple 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 $b^*_{i*} = (b^*_{i1}, \ldots, b^*_{iL})$ 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. Then, b^*_{i*} has the distribution of a vector of independent

Poisson variables. More precisely,

$$b^*_{i*} \sim \mathbb{Q}_k := \prod_{\ell=1}^{L} \mathrm{Poi}(\lambda_{k\ell}), \quad \text{if,} \quad y_i = k, \tag{2.4}$$

where $\lambda_{k\ell}$ are the row mean parameters defined in (2.2). Note that the distributions $\mathbb{Q}_k$, $k = 1, \ldots, K$ are known under our simplifying assumptions. The problem of determining the row labels thus reduces to deciding from which of these K known distributions it comes from. 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 the heart of the algorithms discussed in Sections 4 and 6. 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*. Note that the performance of this oracle gives a lower bound on the performance of any biclustering algorithm in our model. 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 $I^{\mathrm{col}}_{\ell r}$—defined in (3.1) and (3.2)—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 as well, hence nothing 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 SBM.

2.3 Notation on misclassification rates

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

$$\mathrm{Mis}(\widehat{y}, y) := \min_{\sigma \in \Pi_n} \frac{1}{n} \sum_{i=1}^{n} 1\{\sigma(\widehat{y}_i) \neq y_i\}. \tag{2.5}$$

Letting σ^* be a minimizer in (2.5), the misclassification rate over cluster k is

$$\mathrm{Mis}_k(\widehat{y}, y) := \frac{1}{n_k(y)} \sum_{i: y_i = k} 1\{\sigma^*(\widehat{y}_i) \neq y_i\} = \frac{|i : \sigma^*(\widehat{y}_i) \neq k,\ y_i = k|}{n_k(y)}, \tag{2.6}$$

using the cardinality notation to be discussed shortly. Note that (2.6) is not symmetric in its arguments. We will also use the notation $\sigma^*(\widehat{y} \to y)$ to denote an optimal permutation in (2.5). When $\mathrm{Mis}(\widehat{y}, y)$ 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 $\mathrm{dMis}(\widehat{y}, y)$, which is obtained by setting the permutation in (2.5) to the identity. With $\sigma^* = \sigma^*(\widehat{y} \to y)$, we have $\mathrm{Mis}(\widehat{y}, y) = \mathrm{dMis}(\sigma^*(\widehat{y}), y)$. We note that

$$\mathrm{Mis}(\widehat{y}, y) \ = \sum_{k \in K} \pi_k(y)\ \mathrm{Mis}_k(\widehat{y}, y) \ \leq \max_{k \in K} \mathrm{Mis}_k(\widehat{y}, y), \tag{2.7}$$

as well as $\max_{k \in K} \mathrm{Mis}_k(\widehat{y}, y) \leq \mathrm{Mis}(\widehat{y}, y) / \min_{k'} \pi_{k'}(y)$. We can similarly define the misclassification rate of an estimate $\widehat{z}$ relative to z. Our goal is to derive efficient algorithms to obtain $\widehat{y}$ 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 all parameters, including m, are functions of n. 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 exist 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 the 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 nothing that we use *community* and *cluster* interchangeably in this paper, although some authors prefer to use community for the assorative 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.

CHAPTER 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 (2.2) and (2.3). Let $\Lambda_{\min}$ and $\|\Lambda\|_\infty$ be the minimum and maximum value of the entries of Λ, respectively, and similarly for Γ. We assume

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

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

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

for all $k \in [K]$ and $\ell \in [L]$. The following key quantity controls the misclassification rate:

$$I_{kr} := I_{kr}(\Lambda) := \sup_{s \in (0,1)} \sum_{\ell=1}^{L} (1-s)\lambda_{k\ell} + s\lambda_{r\ell} - \lambda_{k\ell}^{1-s}\lambda_{r\ell}^{s}, \tag{3.1}$$

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 (3.1) as $s \mapsto I_s$, with some abuse of notation assuming k and r are fixed, and we note that 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 $\{\mathbb{Q}_k\}$, (3.1) is the Chernoff exponent in testing the two hypothesis $\mathbb{Q}_k$ and $\mathbb{Q}_r$ [Che52]. The difference with the classical setting, in which the Chernoff exponent appears, is that we work in the regime where we are effectively testing based on a sample of size of 1 and instead of the sample size, we let $I_{kr} \to \infty$. We define

the column information matrix similarly

$$I^{\mathrm{col}}_{\ell\ell'} := I_{\ell\ell'}(\Gamma) = \sup_{s\in(0,1)} \sum_{k=1}^{K} (1-s)\Gamma_{\ell k} + s\Gamma_{\ell' k} - \Gamma_{\ell k}^{1-s}\Gamma_{\ell' k}^{s}, \tag{3.2}$$

for all $\ell, \ell' \in [L]$. Another set of key quantities in our analysis are:

$$\varepsilon_{kr} := \max_{\ell\in[L]} \Big(\frac{\lambda_{k\ell}}{\lambda_{r\ell}} \vee \frac{\lambda_{r\ell}}{\lambda_{k\ell}} \Big) - 1, \quad \varepsilon_k := \min_{r\in[K]} \varepsilon_{kr}, \quad \text{and} \quad \varepsilon := \min_{k\in[K]} \varepsilon_k. \tag{3.3}$$

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 $\Lambda = [\lambda_{k*}; \lambda_{r*}] \in \mathbb{R}_+^{2\times K}$. 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*

$$\mathbb{P}(\textit{Type I error}) + \mathbb{P}(\textit{Type II error}) \begin{cases} \leq & C \exp\big(-I_{kr} - \frac{1}{2}\log\Lambda_{\min}\big), \\ \geq & \exp\big(-I_{kr} - \frac{L}{2}(\log\Lambda_{\min} + C')\big). \end{cases} \tag{3.4}$$

See Corollary 10 and Appendix Chapter A.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 6. 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 significance of the Chernoff exponent of the hypothesis test in controlling the rates is thus natural, given the full information about the $\{\mathbb{Q}_k\}$ 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 subblocks when partitioning, the details of which are discussed in Chapter 6. Under the following two assumptions:

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

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

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 (Chapter 2.1) satisfying* (A1)–(A4)*. Then, as* $I_{min} \wedge I^{col}_{min} \to \infty$ *and* $\Lambda_{\min} \to \infty$*, the row labels* $\widehat{y}$ *output by Algorithm 3 in Chapter 6 satisfies for some* $\zeta = o(1)$*,*

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = O\Big(\omega \sum_{r \neq k} \Big(1 + \frac{1}{\varepsilon_{kr}}\Big) \exp\Big(-I_{kr} - \Big(\frac{1}{2} - \zeta\Big) \log \Lambda_{\min}\Big)\Big) \tag{3.5}$$

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

One can replace the big O with the small o in (3.5) to obtain an equivalent result (due to the presence of $\zeta = o(1)$). Let us discuss the assumptions of Theorem 1. The only real assumptions are (A3) and (A4). The other two, namely (A1) and (A2) can be more or less thought of as definitions of ω and β. For example, (A2) only imposes the mild constraint that no cluster is empty. Similarly (A1) imposes the mild assumption that no entry of Λ or Γ is zero. The main constraints on ω and β are encoded in (A3) and (A4) in tandem with other parameters of the model.

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 $\sqrt{n}$. (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 Chapter 6.2.1, we present Theorem 7, for the likelihood-based portion of the algorithm, assuming that a good initialization is provided. Theorem 7 only requires a weakened version, (A4′), of assumption (A4).

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* $k \in [K]$*,*

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = o\Big(\sum_{r \neq k} \exp(-I_{kr})\Big). \tag{3.6}$$

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

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = O\Big(\sum_{r\neq k} \exp\Big(-I_{kr} - \Big(\frac{1}{2} - \zeta\Big)\log \Lambda_{\min}\Big)\Big). \tag{3.7}$$

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

$$\mathbb{E}\big[\,\mathrm{Mis}_k\left(\widehat{y},\, y\right)\big] = O\Big(\sum_{r\neq k} \exp\Big(-I_{kr} - \frac{1}{2}\log \Lambda_{\min}\Big)\Big), \tag{3.8}$$

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 6 for the details. This error rate coincides with (3.7), 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.

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

Example 1. Consider a simple planted partition model where

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

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 $k \neq \ell$. Applying (3.1) with $s = 1/2$,

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

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

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

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

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = o\Big(\exp\Big(-\frac{(\sqrt{a}-\sqrt{b})^2}{\beta K}\Big)\Big). \tag{3.9}$$

For the details of (3.9), see Chapter 8.4. In particular, if

$$\liminf_{n\to\infty} \frac{(\sqrt{a}-\sqrt{b})^2}{\beta K \log n} \;\geq\; 1, \tag{3.10}$$

we have $\mathrm{Mis}_k\left(\widehat{y},\, y\right) = o(1/n)$ w.h.p., that is, we have the exact recovery of the labels by Algorithm 3. (Whenever misclassification rate drops below $1/n$, it should be exactly zero.) Note that 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. Then,

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

where $\tilde{I}_{kr}$ is defined based on $\widetilde{\lambda}_{k\ell}$ and $\widetilde{\lambda}_{r\ell}$ as in (3.1). 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

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = o\Big(\exp\Big(-\min_{r\neq k} \tilde{I}_{kr} \log n\Big)\Big) = o\Big(n^{-\min_{r\neq k} \tilde{I}_{kr}}\Big). \tag{3.11}$$

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

3.1 Comparison with existing results

Let us now compare with [GMZZ17] and [AS15] whose results are closest to our work. 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 [GMZZ17]. They have shown that there is an algorithm with misclassification error

bounded by

$$\exp\left(-\frac{(1-o(1))(\sqrt{a}-\sqrt{b})^2}{\beta K}\right). \tag{3.12}$$

We have sharpened this rate to (3.9) under assumption (A3) (i.e., assuming the average degree grows slower than $O(\sqrt{n})$). Bound (3.12) 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 (3.10).

We note that the results in [GMZZ17] 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 [GMZZ17] 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 (3.12) or the one we gave in (3.9) 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 (3.1). In other words, Algorithm 3 that we present is *rate adaptive* over the class considered in [GMZZ17], 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 [GMZZ17]. 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 forms of the bounds. Another key is the result in Lemma 5(b) which provides a better error rate than the classical Chernoff bound, using a very innovative technique introduced in [AS15]. Moreover, we keep track of the balance parameter β in (A2) throughout, allowing it to go to infinity slowly. Last but not least, assortativity is a key assumption in [GMZZ17], while our result does not rely on it. Besides consistency, our provable algorithm is more computationally efficient in a practical sense. To obtain initial labels, we will apply spectral clustering on very few subgraphs (8 to be exact). However, the provable version of the algorithm in [GMZZ17] applies 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 [GMZZ17] costs roughly nC_{n-1}.

Our algorithm thus has a significant advantage in computational complexity when $n \to \infty$. To be fair, the algorithm introduced in [GMZZ17] 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 [GMZZ17]. In addition, [GMZZ17] have a result (their Theorem 5) for when ω grows arbitrarily fast which is not covered by our result. See the discussion below for comments on the symmetric case and dependence on ω.

The problem of exact recovery for a general SBM has been considered in [AS15], again for the case of a symmetric SBM, though the results are applicable to the bipartite setting (with $y = z$) as well. The model and scaling considered in [AS15] is the same as that of Example 2, and they show that exact recovery of all labels is possible if (and only if) $\min_{k,r:k\neq r} \tilde{I}_{kr} \geq 1$ which is the same result we obtain in Example 2 for Algorithm 3. Thus, our result contains that of [AS15] as a special case, namely in the $\log n$-degree regime with other parameters (such as K and the normalized connectivity matrix) kept constant. The results and algorithms of [AS15] do not apply to the general model in our paper; consider the following two points:

1. Only the regime $P \sim \log n/n$, i.e., the degree grows as fast as $\log n$, is investigated in [AS15], while we allow the degree to grow in the range from "arbitrarily slowly" up to "as fast as $O(\sqrt{n})$".

2. One needs independent versions of the adjacency matrix in different stages of the algorithm. To achieve this goal, *edge splitting* was introduced in [AS15]. 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 [AS15, 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 $\widetilde{P}_{\min} = \min_{k\ell} \widetilde{P}_{k\ell}$,

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

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 10.

3.2 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 [GMZZ17], 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 [GMZZ17, 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.

CHAPTER 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:

$$[\mathscr{L}(A, \tilde{y}, \tilde{z})]_{k\ell} = \frac{1}{n_k(\tilde{y})} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} 1\{\tilde{y}_i = k, \tilde{z}_j = \ell\}, \quad k \in [K],\ \ell \in [L]. \tag{4.1}$$

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

$$\hat{\Lambda} = (\hat{\lambda}_{k\ell}) := \mathscr{L}(A, \tilde{y}, \tilde{z}), \tag{4.2}$$

$$\Lambda(\tilde{y}, \tilde{z}) = (\lambda_{k\ell}(\tilde{y}, \tilde{z})) := \mathscr{L}(\mathbb{E}[A], \tilde{y}, \tilde{z}), \tag{4.3}$$

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

$$\Lambda(\tilde{y}, \tilde{z}) \,|_{\tilde{y}=y,\, \tilde{z}=z} = \Lambda \tag{4.4}$$

where Λ is the *true* (global) row mean parameter matrix defined in (2.2). In words, (4.4) 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, q' range over $[Q] = \{1, \ldots, Q\}$ creating a partition of A into Q^2 subblocks. We call

$$\Lambda^{(q',q)}(\tilde{y}, \tilde{z}) := (\lambda_{k\ell}^{(q',q)}(\tilde{y}, \tilde{z})) := \mathscr{L}(\mathbb{E}[A^{(q',q)}],\, \tilde{y}^{(q')},\, \tilde{z}^{(q)}), \tag{4.5}$$

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 Chapter 6. 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 the 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 Chapter 7.2. 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 $q \in [Q]$,

$$\lambda_{k\ell}^{(q',q)}(y, z) = P_{k\ell}\, n_\ell(z^{(q)}), \quad \forall q' \in [Q], \tag{4.6}$$

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 [ACBL+13] to the bipartite setup. The pseudo-likelihood

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

algorithm (PLA) is effectively an EM algorithm applied to the approximate mixture of Poissons obtained from the block compression of the adjacency matrix A. 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 Chapter 5 once we introduce the provable version of the algorithm. 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:

$$\mathcal{B}(A; \tilde{z}) := \boldsymbol{b}(\tilde{z}) := \big(b_{i\ell}(\tilde{z})\big) \in \mathbb{Z}_+^{n \times L}, \quad b_{i\ell}(\tilde{z}) := \sum_{j=1}^{m} A_{ij} 1\{\tilde{z}_j = \ell\}. \tag{4.7}$$

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

$$\mathbb{E}[b_{i\ell}(\tilde{z})] = \lambda_{k\ell}(y, \tilde{z}) = \lambda_{k\ell}(y, \tilde{z})|_{\tilde{z}=z}, \quad \text{if } y_i = k, \tag{4.8}$$

where $\lambda_{k\ell}(y, \tilde{z})$ are the (global) row mean parameters defined in (4.3).

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* $\lambda_{k\ell}(y, \tilde{z})$:

$$\mathcal{L}(\boldsymbol{b}; \tilde{y}) := \hat{\Lambda} := [\hat{\lambda}_{k\ell}] \in \mathbb{R}_+^{K \times L}, \quad \hat{\lambda}_{k\ell} := \frac{1}{n_k(\tilde{y})} \sum_{i=1}^{n} b_{i\ell} 1\{\tilde{y}_i = k\}. \tag{4.9}$$

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 (4.9) are consistent with those of (4.2) 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

$$\pi(\tilde{y}) := (\pi_k(\tilde{y})), \quad \pi_k(\tilde{y}) := \frac{1}{n} \sum_{i=1}^{n} 1\{\tilde{y}_i = k\} \tag{4.10}$$

for the estimate of (row) class priors based on $\tilde{y}$. We note that the operation $\mathcal{B}$ 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 (4.7) to soft row labels, it is enough to replace $1\{z_j = \ell\}$ with $z_{j\ell}$. Extending (4.9) to soft column labels $\tilde{y}$ is done similarly.

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

$$\mathcal{F}(\boldsymbol{b}, \hat{\Lambda}, \tilde{\pi}) := (\hat{\pi}_{ik}) \in [0,1]^{n \times K}, \quad \hat{\pi}_{ik} := \frac{\tilde{\pi}_k \prod_{\ell=1}^{L} \varphi(b_{i\ell}, \hat{\lambda}_{k\ell})}{\sum_{k'=1}^{K} \tilde{\pi}_{k'} \prod_{\ell=1}^{L} \varphi(b_{i\ell}, \hat{\lambda}_{k'\ell})} \tag{4.11}$$

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 $\tilde{\pi}$ in this step; similarly, we only use a block compression which is based on 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 which can be considered our new estimates of y. We can convert $(\hat{\pi}_{ik})$ to hard labels if needed.

Algorithm 1 summarizes the general blueprint of PLA, which proceeds by iterating the three operators (4.7), (4.9) and (4.11). Optional conversion from soft to hard labels is

Algorithm 2 Simplified pseudo-likelihood clustering

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

performed by MAP assignment per row. With option 2 in step 6, the inner loop on lines 4–9 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 Chapter 9. In practice, we recommend to keep soft labels throughout, and only run the inner loop for a few iterations (maybe even one if the computational cost is of significance).

Remark 3 (PL naming)**.** We have borrowed the name pseudo-likelihood (PL) from [ACBL+13] based on which the algorithms in this paper are derived. In [ACBL+13], the setup is that of the symmetric SBM, and in order to treat the full likelihood as the product of independent (over nodes $i = 1, \ldots, n$) 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 [ACBL+13] 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.

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 Chapter 6—can be equivalently described as a *likelihood ratio classifier* (LRC). Let us write

the joint Poisson likelihood (up to a constant) as:

$$\Phi(x,\lambda) = \prod_{\ell=1}^{L} \varphi(x_\ell, \lambda_\ell) = \prod_{\ell=1}^{L} \exp(x_\ell \log \lambda_\ell - \lambda_\ell), \quad x \in \mathbb{R}^L,\ \lambda \in \mathbb{R}_+^L, \tag{4.12}$$

and the corresponding likelihood ratio as:

$$\Psi(x; \lambda \mid \lambda') = \log \frac{\Phi(x,\lambda)}{\Phi(x,\lambda')} = \sum_{\ell=1}^{L} x_\ell \log \frac{\lambda_\ell}{\lambda'_\ell} + \lambda'_\ell - \lambda_\ell, \quad x \in \mathbb{R}^L,\ \lambda, \lambda' \in \mathbb{R}_+^L. \tag{4.13}$$

Recalling the column compression (4.7), the *likelihood ratio classifier*, based on initial row labels $\tilde{z}$ and an estimate $\tilde{\Lambda}$ of the row mean parameter matrix, is

$$[\mathrm{LR}(A, \tilde{\Lambda}, \tilde{z})]_i \ \in \ \operatorname*{argmax}_{r \in [K]} \ \log \Phi(b_{i*}(\tilde{z}), \tilde{\lambda}_{r*}), \quad i \in [n]. \tag{4.14}$$

which 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 $\widehat{y} = \mathrm{LR}(A, \tilde{\Lambda}, \tilde{z})]$.

CHAPTER 5

Spectral clustering

5.1 Notation

In this chapter, it will be convenient to consider another set of notations.

Orthogonal matrices. We write $\mathbb{S}^n$ for the set of symmetric $n \times n$ matrices, and $\mathbb{O}^{n\times k}$ for the set of $n \times k$ matrices with orthonormal columns. The condition $k \leq n$ is implicit in defining $\mathbb{O}^{n\times k}$. The case $\mathbb{O}^{n\times n}$ is the set of orthogonal matrices, though with some abuse of terminology we also refer to matrices in $\mathbb{O}^{n\times k}$ as orthogonal even if $k < n$. Thus, $Z \in \mathbb{O}^{n\times k}$ iff $Z^T Z = I_k$. We also note that $Z \in \mathbb{O}^{n\times k_1}$ and $U \in \mathbb{O}^{k_1\times k}$ implies $ZU \in \mathbb{O}^{n\times k}$. The following holds:

$$\|Ux\|_2 = \|x\|_2, \quad \forall x \in \mathbb{R}^k,\ U \in \mathbb{O}^{k_1\times k}, \tag{5.1}$$

for any $k_1 \geq k$. On the other hand,

$$\|U^T x\|_2 \leq \|x\|_2, \quad \forall x \in \mathbb{R}^{k_1},\ U \in \mathbb{O}^{k_1\times k}, \tag{5.2}$$

where equality holds for all $x \in \mathbb{R}^{k_1}$, iff $k_1 = k$. To see this latter inequality, let $u_1, \ldots, u_k \in \mathbb{R}^{k_1}$ be the columns of U, constituting an orthonormal sequence which can be completed to an orthonormal basis by adding say $u_{k+1}, \ldots, u_{k_1}$. Then, $\|U^T x\|_2^2 = \sum_{j=1}^{k} \langle u_j, x\rangle^2 \leq \sum_{j=1}^{k_1} \langle u_j, x\rangle^2 = \|x\|_2^2$.

Memebership matrices and misclassification. We let $\mathbb{H}^{n\times k}$ denote the set of *hard* cluster labels: $\{0,1\}$-valued $n \times k$ matrices where each row has exactly a single 1. A matrix $Z \in \mathbb{H}^{n\times k}$ is also called a membership matrix, where row i is interpreted as the membership

of node i to one of k clusters (or communities). Here we implicitly assume that we have a network on nodes in $[n] = \{1, \dots, n\}$, and there is a latent partition of $[n]$ into k clusters. In this sense, $Z_{ik} = 1$ iff node i belongs to cluster k. Given, two membership matrices $Z, Z' \in \mathbb{H}^{n \times k}$, we can consider the average misclassification rate between them, which we denote as $\overline{\mathrm{Mis}}(Z, Z')$: Letting z_i^T and $(z_i')^T$ denote the ith row of Z and Z' respectively, we have

$$\overline{\mathrm{Mis}}(Z, Z') := \min_Q \frac{1}{n} \sum_{i=1}^n 1\{z_i \neq Q z_i'\} \tag{5.3}$$

where the minimum is taken over $k \times k$ permutations matrices Q. We also let $\mathrm{Mis}_r(Z, Z')$ be the misclassification rate between the two, over the rth cluster of Z, that is, $\mathrm{Mis}_r(Z, Z') = \frac{1}{n_r} \sum_{i:\, z_i = r} 1\{z_i \neq Q^* z_i'\}$ where $n_r = \sum_{i=1}^n 1\{z_i = r\}$ is the size of the rth cluster of Z, and Q^* is the optimal permutation in (5.3). Note that in contrast to $\overline{\mathrm{Mis}}$, Mis_r is not symmetric in its two arguments. We also write $\mathrm{Mis}_\infty := \max_r \mathrm{Mis}_r$. These definitions can be extended to misclassification rates between k-means matrices introduced in Section 5.3.4.

5.2 Stochastic Block Model

Stochastic Block Model (SBM) with bi-adjacency matrix $A \in \{0,1\}^{n_1 \times n_2}$. We assume throughout that $n_2 \geq n_1$, without loss of generality. We have membership matrices $Z_r \in \mathbb{H}^{n_r \times k_r}$ for each of the two sides $r = 1, 2$, where $k_r \leq n_r$ denotes the number communities on side r. Each element of A is an independent draw from a Bernoulli variable, and

$$P := \mathbb{E}[A] = Z_1 B Z_2^T, \quad B = \frac{\Psi}{\sqrt{n_1 n_2}} \tag{5.4}$$

where $B \in [0,1]^{k_1 \times k_2}$ is the connectivity— or the edge probability—matrix, and Ψ is its rescaled version. We also use the notation

$$A \sim \mathrm{Ber}(P) \iff A_{ij} \sim \mathrm{Ber}(P_{ij}), \text{ independent across } (i,j) \in [n_1] \times [n_2]. \tag{5.5}$$

Classical SBM which we refer to as *symmetric SBM* in this paper corresponds to the following modifications:

(a) A is assumed to be symmetric: Only the upper diagonal elements are drawn independently and the bottom half is filled symmetrically. For simplicity, we allow for self-loops, i.e. draw the diagonal elements from the same model. This will have negligible effect in the arguments.

(b) $n_1 = n_2 = n,\ k_1 = k_2 = k,\ Z_1 = Z_2 = Z.$

(c) B is assumed symmetric.

We note that (5.4) still holds over all the elements. Directed SBM is also a special case, where (b) is assumed but not (a) or (c). That is, A is not assumed to be symmetric and all the entries are independently drawn, while B may or may not be symmetric.

We refer to P as the *mean matrix* and note that it is of rank at most $k := \min\{k_1, k_2\}$. Often $k \ll n_1, n_2$, that is P is a low-rank matrix which is the key in why spectral clustering works well for SBMs. Let us write P in a form which is more suitable for understanding its spectral properties. We let $N_r = \mathrm{diag}(n_{r1}, \ldots, n_{rk_r})$ for $r = 1, 2$ where n_{rj} is the size of the jth cluster of Z_r; that is, N_r is a diagonal matrix whose diagonal elements are the sizes of the clusters on side r. We note that $Z_r^T Z_r = N_r$. (To see this, let z_{ri}^T be the ith row of Z_r and note that $Z_r^T Z_r = \sum_{i=1}^n z_{ri} z_{ri}^T$. Since $z_{ri} \in \mathbb{H}^{1\times k_r}$, each $z_{ri} z_{ri}^T$ is a diagonal matrix with a single 1 on the diagonal at the position determined by the cluster assignment of node i on side r.)

Letting $\bar{Z}_r = Z_r N_r^{-1/2}$, we observe that $\bar{Z}_r$ is an orthogonal matrix $\bar{Z}_r^T \bar{Z}_r = I_{k_r}$. In other words, $\bar{Z}_r \in \mathbb{O}^{n_r \times k_r}$, and we can write

$$P = \bar{Z}_1 \bar{B} \bar{Z}_2^T, \quad \text{for} \quad \bar{B} := N_1^{1/2} B N_2^{1/2} = \bar{N}_1^{1/2} \Psi \bar{N}_2^{1/2},$$

where we have further normalized N_r to get the cluster proportions:

$$\bar{N}_r := N_r / n_r = \mathrm{diag}(\pi_{r1}, \ldots, \pi_{rk_r}), \quad \pi_{rj} = n_{rj}/n_r. \tag{5.6}$$

For sparse graphs, we expect $\bar{N}_r$ and Ψ to remain stable as $n_r \to \infty$, hence $\bar{B}$ remains stable;

see Remark 4 below. Let $\bar{B}$ have the following reduced SVD:

$$\bar{B} = U_\psi \Sigma V_\psi^T \tag{5.7}$$

where $U_\psi \in \mathbb{O}^{k_1 \times k}, V_\psi \in \mathbb{O}^{k_2 \times k}$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$, and we recall that $k = \min\{k_1, k_2\}$. It then follows that the mean matrix P has the following reduced SVD

$$P = (\bar{Z}_1 U_\psi)\, \Sigma\, (\bar{Z}_2 V_\psi)^T \tag{5.8}$$

where $\bar{Z}_1 U_\psi \in \mathbb{O}^{n_1 \times k}$ and $\bar{Z}_2 V_\psi \in \mathbb{O}^{n_2 \times k}$. When dealing with the symmetric SBM, we will drop the subscript r from all the relevant quantities; for example, we write $N = N_1 = N_2$, $\bar{Z} = \bar{Z}_1 = \bar{Z}_2$, $\pi_j = \pi_{1j} = \pi_{2j}$, and so on.

Remark 4 (Scaling and sparsity)**.** Let us comment on the normalization in (5.4). As can be seen from the above discussion leading to (5.7) and (5.8), this normalization is natural for studying spectral clustering. In the symmetric case, where $n_1 = n_2 = n$, the normalization reduces to $B = \Psi/n$, which is often assumed when studying sparse SBMs by requiring that either $\|\Psi\|_\infty$ is $O(1)$ or grows slowly with n. To see why this implies a sparse network, note that the expected average degree of the symmetric SBM (under this scaling) is

$$\frac{1}{n} 1^T P 1 = \frac{1}{n} 1^T N B N 1 = 1^T \bar{N} \Psi \bar{N} 1 = \sum_{i,j=1}^{k} \pi_i \pi_j \Psi_{ij} =: d_{\mathrm{av}}$$

using $1_n^T Z = 1_k^T N$. (Here and elsewhere, 1 is the vector of all ones of an appropriate dimension; we write 1_n if we want to emphasize the dimension n.) Thus, the growth of the average expected degree, d_{av}, is the same as Ψ, and as long as Ψ is $O(1)$ or grows very slowly with n, the network is sparse. Alternatively, we can view the expected density of the network (the expected number of edges divided by the total number of possible edges) as a measure of sparsity. For the symmetric case, the expected density is $(\frac{1}{2} 1^T P 1)/\binom{n}{2} \sim d_{\mathrm{av}}/n$ and is $O(n^{-1})$ if $d_{\mathrm{av}} = O(1)$. Similar observations hold in the general bipartite case if we let $n = \sqrt{n_1 n_2}$, the geometric mean of the dimensions. The expected density of the bipartite

network under the scaling of (5.4) is

$$\frac{1^T P 1}{n^2} = \frac{1}{n^2} 1^T N_1 B N_2 1 = \frac{1}{n} 1^T \bar{N}_1 \Psi \bar{N}_2 1 = \frac{d_{\text{av}}}{n}, \quad (n = \sqrt{n_1 n_2})$$

where $d_{\text{av}} := 1^T \bar{N}_1 \Psi \bar{N}_2 1 = \sum_{i,j} \pi_{1i} \pi_{2j} \Psi_{ij}$ can be thought of as the analog of the expected average degree in the bipartite case. As long as $\|\Psi\|_\infty$ grows slowly relative to $n = \sqrt{n_1 n_2}$, the bipartite network is sparse.

5.3 Analysis steps

Throughout, we focus on recovering the row clusters. Everything that we discuss goes through, with obvious modifications, for recovering the column clusters. Recalling the decomposition (5.8), the idea of spectral clustering in the context of SBMs is that $\bar{Z}_1 U_\psi$ has enough information for recovering the clusters and it can be obtained by computing a reduced SVD of P. In particular, applying a k-means type clustering on the rows of $\bar{Z}_1 U_\psi$ should recover the cluster labels. On the other hand, the actual random adjacency matrix, A, is concentrated around the mean matrix P, after proper regularization if need be. We denote this potentially *regularized version as* A_{re}. Then, by the spectral perturbation theory, if we compute a reduced SVD of $A_{\text{re}} = \widehat{Z}_1 \hat{\Sigma} \widehat{Z}_2^T$ where $\widehat{Z}_r \in \mathbb{O}^{n_r \times k}$, $r = 1, 2$ and $\hat{\Sigma}$ is diagonal, we can conclude that $\widehat{Z}_1$ concentrates around $\bar{Z}_1 U_\psi$. Hence, applying a stable (i.e., continuous) k-means type algorithm on $\widehat{Z}_1$ should be able to recover the labels with a small error.

5.3.1 Analysis sketch

Let us sketch the argument above in more details. A typical approach in proving consistency of spectral clustering consists of the following steps:

1. We replace A with a properly regularized version A_{re}. We provide the details for one such regularization in Theorem 3 (Section 5.3.3). However, the only property we require of the regularized version is that it concentrates, with high probability, around

the mean of A, at the following rate (assuming $n_2 \geq n_1$):

$$\|A_{\text{re}} - \mathbb{E}A\| \leq C\sqrt{a}, \quad \text{where} \quad a \geq \sqrt{\frac{n_2}{n_1}}\|\Psi\|_\infty. \tag{5.9}$$

Here and throughout $\|\cdot\|$ is the $\ell_2 \to \ell_2$ operator norm and $\|\Psi\|_\infty = \max_{ij} \Psi_{ij}$.

2. We pass from A_{re} and $P = \mathbb{E}[A]$ to their (symmetrically) dilated versions $A_{\text{re}}^\dagger$ and $P^\dagger$. The symmetric dilation operator will be given in (5.12) (Section 5.3.2) and allows us to use spectral perturbation bounds for symmetric matrices. A typical final result of this step is a bound of the form:

$$\|\widehat{Z}_1 - \bar{Z}_1 U_\psi Q\|_F \leq \frac{C_2}{\sigma_k}\sqrt{ka}, \quad \text{w.h.p.} \tag{5.10}$$

for some $Q \in \mathbb{O}^{k\times k}$. We recall that $\|\cdot\|_F$ is the Frobenius norm. Here, σ_k is the smallest nonzero singular value of $\bar{B}$ as given in (5.7). The form of (5.10) will be different if instead of $\widehat{Z}_1$ one considers other objects as the end result of this step; see Section 5.4 (e.g., (5.31)) for instances of such variations. The appearance of Q is inevitable and is a consequence of the necessity of *properly aligning* the bases of spectral subspaces, before they can be compared in Frobenius norm (cf. Lemma 2). Nevertheless, the growing stack of orthogonal matrices on the RHS of $\bar{Z}_1$ has little effect on the performance of row-wise k-means, as we discuss shortly.

3. The final step is to analyze the effect of applying a k-means algorithm to $\widehat{Z}_1$. Here, we introduce the concept of a *k-means matrix*, one whose rows take at most k distinct values. (See Section 5.3.4 for details). A k-means algorithm $\mathscr{K}$ takes a matrix $\hat{X} \in \mathbb{R}^{n\times d}$ and outputs a k-means matrix $\mathscr{K}(\hat{X}) \in \mathbb{R}^{n\times d}$. Our focus will be on k-means algorithms with the following property: If $X^* \in \mathbb{R}^{n\times d}$ is a k-means matrix, then for some constant $c > 0$,

$$\|\hat{X} - X^*\|_F^2 \leq \varepsilon^2 \implies \overline{\text{Mis}}(\mathscr{K}(\hat{X}), X^*) \leq c\,\varepsilon^2/(n\delta^2). \tag{5.11}$$

Here, $\overline{\text{Mis}}$ is the average misclassification rate between two k-means matrices. As will become more clear in Section 5.3.4, k-means matrices encode both the cluster label information and cluster center information, and these two pieces can be recovered from

them in a lossless fashion. Thus, it makes sense to talk about misclassification rate between k-means matrices, by interpreting it as a statement about their underlying label information. $\delta^2 = \delta^2(X^*)$ in (5.11) is the minimum center separation of X^* (cf. Definition 2). In Section 5.3.4, we will discuss k-means algorithms that satisfy (5.11).

Applying (5.11) with $\hat{X} = \widehat{Z}_1$, $X^* = \bar{Z}_1 U_\psi Q$ and $\varepsilon^2 = C_2^2\, ka/\sigma_k^2$, and combining with (5.10) leads to a misclassification rate for the spectral clustering algorithm (cf. Theorem 2).

The preceding three steps of the analysis follow the three steps of a general spectral clustering algorithm, which we refer to as *regularization*, *spectral truncation* and *kmeans* steps, respectively. Recalling the definition of cluster proportions, let us assume for some $\beta_r \geq 1$,

$$\max_{(t,s):\, t\neq s} \frac{2}{\pi_{rt}^{-1} + \pi_{rs}^{-1}} \leq \frac{\beta_r}{k_r}, \quad r = 1, 2. \tag{A1}$$

The LHS is the maximum *harmonic mean* of pairs of distinct cluster proportions. For balanced clusters, we have $\pi_{rt} = 1/k_r$ for all $t \in [k_r]$ and we can take $\beta_r = 1$. In general, β_r measures the deviation of the clusters (on side r) from balancedness. The following is a prototypical consistency theorem for a spectral clustering algorithm:

Theorem 2 (Prototype SC consistency)**.** *Consider a spectral algorithm with a kmeans step satisfying* (5.11)*, and the "usual" spectral truncation step, applied to a regularized bi-adjacency matrix A_{re} satisfying concentration bound* (5.9)*. Let $\mathscr{K}(\widehat{Z}_1)$ be the resulting estimate for membership matrix Z_1, and assume $k_1 = k =: \min\{k_1, k_2\}$. Then, under the SBM model of Section 5.2 and assuming* (A1)*, w.h.p.,*

$$\overline{\mathrm{Mis}}(\mathscr{K}(\widehat{Z}_1), \bar{Z}_1) \lesssim \beta_1 \Big(\frac{a}{\sigma_k^2}\Big).$$

Here, and in the sequel, "with high probability", abbreviated w.h.p., means with probability at least $1 - n^{-c_1}$ for some universal constant $c_1 > 0$. The notation $f \lesssim g$ means $f \leq c_2\, g$ where $c_2 > 0$ is a universal constant. In addition, $f \asymp g$ means $f \lesssim g$ and $g \lesssim f$.

Proof. The only remaining calculation is that of the minimum center separation of $X^* =$

$\bar{Z}_1 O \in \mathbb{R}^{n_1 \times k}$, where $\bar{Z}_1 \in \mathbb{O}^{n_1 \times k_1}$ and $O := U_\psi Q \in \mathbb{O}^{k_1 \times k}$. We have

$$\delta^2 = \delta^2(\bar{Z}_1 O) = \delta^2(\bar{Z}_1) = \min_{t \neq s} \|n_{1t}^{-1/2} e_t - n_{1s}^{-1/2} e_s\|_2^2 = \min_{t \neq s} (n_{1t}^{-1} + n_{1s}^{-1})$$

where $e_s \in \mathbb{R}^{k_1}$ is the sth standard basis vector. The second equality uses invariance of δ^2 to right-multiplication by a square orthogonal matrix. This is a consequence of $\|u^T O - v^T O\|_2 = \|u - v\|_2$ for $u, v \in \mathbb{R}^{k_1}$ and $O \in \mathbb{O}^{k_1 \times k}$ when $k_1 = k$; see (5.1). The third equality is from the definition $\bar{Z}_1 = Z_1 N_1^{-1/2}$. Using (A1),

$$(n_1 \delta^2)^{-1} \leq \max_{t \neq s} (\pi_{1t}^{-1} + \pi_{1s}^{-1})^{-1} \leq \frac{\beta_1}{2k_1}.$$

We obtain, with $\varepsilon^2 = C_2^2 ka/\sigma_k^2$,

$$\overline{\text{Mis}}(\mathscr{K}(\widehat{Z}_1), \bar{Z}_1) = \overline{\text{Mis}}(\mathscr{K}(\widehat{Z}_1), \bar{Z}_1 O) \lesssim \frac{\varepsilon^2}{n_1 \delta^2} \lesssim \beta_1 \frac{k}{k_1} \frac{a}{\sigma_k^2}$$

which gives the result under the assumption $k_1 = k$. □

For (5.11) to hold for a $k$means algorithm, one usually requires some additional constraints on $\varepsilon^2/(n\delta^2)$, ensuring for example that this quantity is small. We will restate Theorem 2 with such conditions explicitly once we consider the details of some k-means algorithms. For now Theorem 2 should be thought of as a general blueprint, with specific variations obtained in Section 5.4 for various spectral clustering algorithms.

Remark 5. To see that Theorem 2 is a consistency result, consider the typical case where $\beta_1 \asymp 1$, and $\sigma_k \asymp a$, so that $\overline{\text{Mis}}(\mathscr{K}(\widehat{Z}_1), \bar{Z}_1) = O(a^{-1})$. Then, as long as $a \to \infty$, i.e., average degree of the network grows with n, assuming $n_1 \asymp n_2 \asymp n$ (for some n), we have $\overline{\text{Mis}}(\mathscr{K}(\widehat{Z}_1), \bar{Z}_1) = o(1)$, i.e., the average misclassification rate vanishes with high probability. More specific examples are given in Section 5.4.

Remark 6. Condition (A1) is more relaxed that what is commonly assumed in the literature (though the proof is the same). Stating the condition as a harmonic mean allows one to have similar results as the balanced case when one cluster is large, while others remain more or less balanced. For example, let $\pi_{r1} = 1 - c$ for some constant $c \in (0, 1)$, say $c = 0.4$, and let

$\pi_{rt} = c/(k_r - 1)$ for $t \neq 1$. Then, we have for $s \neq t$

$$\frac{2}{\pi_{rt}^{-1} + \pi_{rs}^{-1}} \leq 2\min\{\pi_{rt}, \pi_{rs}\} = \frac{2c}{k_r - 1} \leq \frac{4}{k_r}$$

assuming $k_r \geq 2$. Hence (A1) holds with $\beta_r = 4$. Note that as k_r is increased, all but one cluster get smaller.

In this rest of this section, we will fill in the details of the above three-step plan, starting with Step 2.

5.3.2 Dilation and SV truncation

Let us define the symmetric dilation operator : $\mathbb{R}^{n_1 \times n_2} \to \mathbb{S}^{n_1+n_2}$ by

$$P^\dagger := \begin{pmatrix} 0 & P \\ P^T & 0 \end{pmatrix}. \tag{5.12}$$

This operator will be very useful in translating the results between the symmetric and non-symmetric cases. Let us collect some of its properties:

Lemma 1. *Let $P \in \mathbb{R}^{n_1 \times n_2}$ have a reduced SVD given by $P = U\Sigma V^T$ where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ is a $k \times k$ nonnegative diagonal matrix . Then,*

(a) $P^\dagger$ has a reduced EVD given by

$$P^\dagger = W \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix} W^T, \quad W = \frac{1}{\sqrt{2}} \begin{pmatrix} U & U \\ V & -V \end{pmatrix} \in \mathbb{O}^{(n_1+n_2)\times 2k}.$$

(b) $P \mapsto P^\dagger$ is a linear operator; it preserves the operator norm: $\|P^\dagger\| = \|P\|$.

(c) $\|P^\dagger\|_F = \sqrt{2}\|P\|_F$.

(d) The gap between k top (signed) eigenvalues of $P^\dagger$ and the rest of its spectrum is $2\sigma_k$.

Proof. Part (a) can be verified directly (e.g. $W^T W = I_{2k}$ follows from $U^T U = V^T V = I_k$) and part (d) follows by noting that $\sigma_j \geq 0$ for all j. Part (b) and (c) also follow directly from part (a), using unitary-invariance of the two norms. □

In addition, let us define a singular value (SV) truncation operator $\mathscr{T}_k : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ that takes a matrix A with SVD $A = \sum_i \sigma_i u_i v_i^T$ to the matrix

$$A^{(k)} := \mathscr{T}_k(A) := \sum_{i=1}^{k} \sigma_i u_i v_i^T. \tag{5.13}$$

In other words, $\mathscr{T}_k$ keeps the largest k singular values (and the corresponding singular vectors) and zeros out the rest. Recall that we order singular values in nonincreasing fashion $\sigma_1 \geq \sigma_2 \geq \cdots$. We also refer to (5.13) as the *k-truncated SVD* of A. Using the dilation and the Davis–Kahan (DK) theorem for symmetric matrices, we have:

Lemma 2. *Let $A_{re}^{(k)} = \widehat{Z}_1 \hat{\Sigma} \widehat{Z}_2^T$ be the k-truncated SVD of A_{re} and assume that the concentration bound (5.9) holds. Let $\bar{Z}_1 U_\psi$ be given by the reduced SVD of P in (5.8). Then, the deviation bound (5.10) holds for some $k \times k$ orthogonal matrix Q, and $C_2 = 2C$.*

Proof. Let $\bar{W}$ and $\widehat{W}$ be the W of Lemma 1(a) for $P^\dagger$ and $A_{\mathrm{re}}^\dagger$, respectively. Let us also write $\bar{W}_1$ and $\widehat{W}_1$ for the $(n_1 + n_2) \times k$ matrices obtained by taking the submatrices of $\bar{W}$ and $\widehat{W}$ on columns $1, \ldots, k$. We have

$$\bar{W}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \bar{Z}_1 U_\psi \\ \bar{Z}_2 V_\psi \end{pmatrix}, \quad \widehat{W}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} \widehat{Z}_1 \\ \widehat{Z}_2 \end{pmatrix}.$$

Note that $\bar{W}_1, \widehat{W}_1 \in \mathbb{O}^{(n_1+n_2)\times k}$. Let $\Pi_{\bar{W}_1}$ be the (orthogonal) projection operator, projecting onto $\mathrm{Im}(\bar{W}_1)$, i.e., the column span of $\bar{W}_1$, and similarly for $\Pi_{\widehat{W}_1}$. We have

$$\begin{aligned}
\|\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}\| &\leq \frac{2}{2\sigma_k} \|A_{\mathrm{re}}^\dagger - P^\dagger\| && \text{(Symmetric DK and Lemma 1(d))} \\
&= \frac{1}{\sigma_K} \|(A_{\mathrm{re}} - P)^\dagger\| && \text{(Linearity of dilation)} \\
&= \frac{1}{\sigma_K} \|A_{\mathrm{re}} - P\| && \text{(Lemma 1(b)).}
\end{aligned}$$

The next step is to translate the operator norm bound on spectral projections into a Frobenius bound. The key here is the bound on the rank of spectral deviations which leads to a

$\sqrt{k}$ scaling as opposed to $\sqrt{n_1+n_2}$, when translating from operator norm to Frobenius:

$$\begin{aligned}\min_{Q\in\mathbb{O}^{k\times k}} \|\widehat{W}_1 - \bar{W}_1 Q\|_F &\le \|\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}\|_F && \text{(By Lemma 28 in Appendix A.9)}\\ &\le \sqrt{2k}\,\|\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}\| && (\,\mathrm{rank}(\Pi_{\widehat{W}_1} - \Pi_{\bar{W}_1}) \le 2k\,)\\ &\le \frac{\sqrt{2k}}{\sigma_k}\|A_{\mathrm{re}} - P\|.\end{aligned}$$

Since $2\|\widehat{W}_1 - \bar{W}_1 Q\|_F^2 = \|\widehat{Z}_1 - \bar{Z}_1 U_\psi Q\|_F^2 + \|\widehat{Z}_2 - \bar{Z}_2 V_\psi Q\|_F^2$, we obtain the desired result after combining with (5.9). □

Remark 7 (Symmetric case)**.** When P is symmetric one can still use the dilation operator. In this case, since P itself is symmetric, it has an eigenvalue decomposition (EVD), say $P = U\Lambda U^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ is the diagonal matrix of the eigenvalues of P. Since these eigenvalues could be negative, there is a slight modification needed to go from the EVD to the SVD of P. Let s_i be the sign of λ_i and set $S = \mathrm{diag}(s_i, i = 1, \ldots, k)$. Then, it is not hard to see that with $V = US$ and $\Sigma = \Lambda S = \mathrm{diag}(|\lambda_i|, i = 1, \ldots, k)$, we obtain the SVD $P = U\Sigma V^T$. In other words, all the discussion in this section, and in particular Lemma 2 hold with $V = US$ and $\sigma_i = |\lambda_i|$. The special case of Lemmas 1 and 2 for the symmetric case appears in [LR15]. These observations combined with the fact that the concentration inequality discussed in Section 5.3.3 holds in the symmetric case leads to the following conclusion: All the results discussed in this paper apply to the symmetric SBM, for the version of the adjacency-based spectral clustering that *sorts the eigenvalues based on their absolute values.* This is the most common version of spectral algorithms in use. On the other hand, one gets a different behavior for the algorithm that considers the top k (signed) eigenvalues. It is also worth noting that we have borrowed the term "symmetric dilation" from [Tro15] where these ideas have been successfully used in translating matrix concentration inequalities to the symmetric case.

5.3.3 Concentration

Next we provide the details of Step 1, namely, the concentration of the regularized adjacency matrix. We will use the non-symmetric version of [LLV17, Theorem 1]. We have the following slight generalization to the rectangular case:

Theorem 3. *Assume $n_1 \leq n_2$ and let $A \in \{0,1\}^{n_1 \times n_2}$ have independent Bernoulli entries with mean $\mathbb{E}[A_{ij}] = p_{ij}$. Take $d \geq \max_{ij} n_2 p_{ij}$. Let A_{re} be obtained from A by an arbitrary reduction of entries, but so that the row sums of A are bounded by $2d$. Then, with probability at least $1 - n^{-c_1}$,*

$$\|A_{re} - \mathbb{E}A\| \leq c_2\sqrt{d}. \tag{5.14}$$

The regularization described in Theorem 3 could be achieved, for example, by setting the entries in any row of A for which the row sum is $> 2d$ to zero. More generally, let

$$\mathcal{I}_d := \Big\{ i : \sum_{j=1}^{n_2} A_{ij} > 2d \Big\} = \big\{ i : \|A_{i*}\|_1 > 2d \big\}$$

where $A_{i*} \in \mathbb{R}^{n_2}$ is the ith row of A. Choose $v_i \in \mathbb{R}^{n_2}_+, i \in \mathcal{I}$ to be any collection of vectors such that $\|v_i\|_1 \leq 2d$ for all $i \in \mathcal{I}$. Then, letting $(A_{\mathrm{re}})_{i*} = A_{i*}$ for all $i \notin \mathcal{I}_d$ and $(A_{\mathrm{re}})_{i*} = v_i$ for $i \in \mathcal{I}$ satisfies the regularization described in Theorem 3.

Theorem 3 follows directly from [LLV17, Theorem 1], by noting that we can pad A with rows of zero to get a square $n_2 \times n_2$ matrix to which the result of [LLV17] is applicable. Recalling the scaling of the connectivity matrix in (5.4), and applying Theorem 3 with $d = a \geq \sqrt{n_2/n_1}\|\Psi\|_\infty = n_2\|P\|_\infty$, we obtain the desired concentration bound (5.9) for the regularization described in Theorem 3.

Remark 8. Results of the form described in (3) hold for A itself without any regularization if one further assumes that $d \gtrsim \log n_2$; see for example [TM10; LR15; CX16] or [BVH+16] for the more general result with $d = \max_i \sum_j p_{ij}$. The general regularization for the adjacency matrix is to either remove the high degree nodes as in [CRV15] or reduce their effect as in [LLV17] and Theorem 3 above. The regularization for the normalized Laplacian is somewhat different since there the low degree nodes are problematic. The general approach is to either inflate all the edges by a small amount before forming the Laplacian as is done in [ACBL+13] and analyzed in [LLV17] (see also [JY13]), or to directly inflate just the degrees as in [CCT12]. Since similar concentration bounds hold, at least for the former approach based on the work of [LLV17], much of the results of this paper also apply to the Laplacian based spectral clustering. The details have been omitted for brevity.

5.3.4 k-means step

Let us now give the details of the third and final step of the analysis. We introduce some notations and concepts that help in the discussion of k-means (type) algorithms.

k-means matrices. Recall that $\mathbb{H}^{n\times k}$ denotes the set of *hard* (cluster) labels: $\{0,1\}$-valued $n\times k$ matrices where each row has exactly a single 1. Take $Z \in \mathbb{H}^{n\times k}$. A related notion is that of a *cluster* matrix $Y = ZZ^T \in \{0,1\}^n$ where each entry denotes whether the corresponding pair are in the same cluster. Relative to Z, Y loses the information about the ordering of the cluster labels. We define the class of *k-means matrices* as follows:

$$\begin{aligned}\mathbb{M}^k_{n,\,d} &:= \{X \in \mathbb{R}^{n\times d} :\ X \text{ has at most } k \text{ distinct rows}\} \\ &= \{ZR :\ Z \in \mathbb{H}^{n\times k} :\ R \in \mathbb{R}^{k\times d}\}.\end{aligned} \tag{5.15}$$

The rows of R, which we denote as r_i^T, play the role of cluster centers. Let us also denote the rows of X as x_i^T. The second equality in (5.15) is due to the following correspondence: Any matrix $X \in \mathbb{M}^k_{n,\,d}$ uniquely identifies a *cluster* matrix $Y \in \{0,1\}^{n\times n}$ via, $Y_{ij} = 1$ iff $x_i = x_j$. This in turn "uniquely" identifies a label matrix Z up to $k!$ permutation of the labels. From Z, we "uniquely" recover R, with the convention of setting rows of R for which there is no label equal to zero. (This could happen if X has fewer than k distinct rows.)

With these conventions, there is a one-to-one correspondence between $X \in \mathbb{M}^k_{n,\,d}$ and $(Z,R) \in \mathbb{H}^{n\times k}\times\mathbb{R}^{k\times d}$, up to label permutations. That is, (Z,R) and (ZQ,QR) are considered equivalent for any permutation matrix Q. The correspondence allows us to talk about a (relative) misclassification rate between two k-means matrices: If $X_1, X_2 \in \mathbb{M}^k_{n,\,d}$ with membership matrices $Z_1, Z_2 \in \mathbb{H}^{n\times k}$, respectively, we set

$$\overline{\text{Mis}}(X_1,X_2) := \overline{\text{Mis}}(Z_1,Z_2). \tag{5.16}$$

k-means as projection. Now consider a general $\hat{X} \in \mathbb{R}^{n\times d}$. The classical k-means problem can be thought of as projecting $\hat{X}$ onto $\mathbb{M}^k_{n,\,d}$, in the sense of finding a nearest member of $\mathbb{M}^k_{n,\,d}$ to $\hat{X}$ in Frobenius norm. Let us write $d_F(\cdot,\cdot)$ for the distance induced by the Frobenius norm, i.e., $d_F(\hat{X},X) = \|\hat{X} - X\|_F$. The k-means problem is that of solving the following

optimization:

$$d_F(\hat{X}, \mathbb{M}^k_{n,\,d}) := \min_{X \in \mathbb{M}^k_{n,\,d}} d_F(\hat{X}, X). \tag{5.17}$$

The arguments to follow go through for any distance on matrices that has a ℓ_2 decomposition over the rows:

$$d_F(\hat{X}, X)^2 = \sum_{i=1}^{n} d(\hat{x}_i, x_i)^2, \tag{5.18}$$

where x_i^T and $\hat{x}_i^T$ are the rows of X and $\hat{X}$ respectively, and $d(\hat{x}_i, x_i)$ is some distance over vectors in $\mathbb{R}^d$. For the case of the Frobenius norm: $d(\hat{x}_i, x_i) = \|\hat{x} - x_i\|_2$, the usual ℓ_2 distance. This is the primary case we are interested in, though the result should be understood for the general case of (5.18). Since solving the k-means problem (5.17) is NP-hard, one can look for approximate solutions:

Definition 1. A κ-approximate k-means solution for $\hat{X}$ is a matrix $\widetilde{X} \in \mathbb{M}^k_{n,\,d}$ that achieves κ times the optimal distance:

$$d_F(\hat{X}, \widetilde{X}) \leq \kappa\, d_F(\hat{X}, \mathbb{M}^k_{n,\,d}). \tag{5.19}$$

We write $\mathcal{P}_\kappa : \mathbb{R}^{n \times d} \mapsto \mathbb{M}^k_{n,\,d}$ for the (set-valued) function that maps matrices $\hat{X}$ to κ-approximate solutions $\widetilde{X}$.

An equivalent restatement of (5.19) is

$$d_F(\hat{X}, \widetilde{X}) \leq \kappa\, d_F(\hat{X}, X), \quad \forall X \in \mathbb{M}^k_{n,\,d}. \tag{5.20}$$

Note that $\mathcal{P}_\kappa(\hat{X}) = \{\widetilde{X} \in \mathbb{M}^k_{n,\,d} : \ \widetilde{X} \text{ satisfies (5.20)}\}$.

Our goal is to show that whenever $\hat{X}$ is close to some $X^* \in \mathbb{M}^k_{n,\,d}$, then any κ-approximate k-means solution based on it, namely $\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})$ will be close to X^* as well. This is done in two steps:

1. If the distance $d_F(X, \widetilde{X})$ between two k-means matrices $X, \widetilde{X} \in \mathbb{M}^k_{n,\,d}$ is small, then their relative misclassification rate $\overline{\text{Mis}}(X, \widetilde{X})$ is so.

2. If a general matrix $\hat{X} \in \mathbb{R}^{n\times d}$ is close to a k-means matrix $X \in \mathbb{M}^k_{n,\,d}$, then so is its κ-approximate k-means projection. More specifically,

$$d_F(\widetilde{X}, X) \leq (1+\kappa)\, d_F(\hat{X}, X), \quad \forall \widetilde{X} \in \mathcal{P}_\kappa(\hat{X}). \tag{5.21}$$

This immediately follows from the triangle inequality $d_F(\widetilde{X}, X) \leq d_F(\widetilde{X}, \hat{X}) + d_F(\hat{X}, X)$ and (5.20). We will write (5.21) compactly as

$$d_F(\mathcal{P}_\kappa(\hat{X}), X) \leq (1+\kappa)\, d_F(\hat{X}, X) \tag{5.22}$$

interpreting $d_F(\mathcal{P}_\kappa(\hat{X}), X)$ as $\max_{\widetilde{X}\in\mathcal{P}_\kappa(\hat{X})} d_F(\widetilde{X}, X)$.

Combining the two steps (taking $X = X^*$), we will have the result.

Let us now give the details of the first step above. For this result, we need the key notion of center separation. k-means matrices have more information that just a membership assignment. They also contain an encoding of the relative positions of the clusters, and hence the minimal pairwise distance between them, which is key in establishing a misclassification rate.

Definition 2 (Center separation)**.** For any $X \in \mathbb{M}^k_{n,\,d}$, let us denote its centers, i.e. distinct rows, as $\{q_r(X), r \in [k]\}$, and let

$$\delta_r(X) = \min_{\ell:\, \ell\neq r} d(q_\ell(X), q_r(X)), \quad \delta_\wedge(X) = \min_r \delta_r(X). \tag{5.23}$$

In addition, let $n_r(X)$ be the number of nodes in cluster r according to X, and $n_\wedge(X) = \min_r n_r(X)$, the minimum cluster size.

If X has $m < k$, the convention would be to let $q_k(X) = 0$ for $k = m+1, \ldots, k$. We usually do not work with these degenerate cases. Implicit in the above definition is an enumeration of the clusters of X. We note that definition of $\delta_r = \delta_r(X)$ in (5.23) implies

$$d(q_\ell(X), q_r(X)) \geq \max\{\delta_\ell, \delta_r\}, \quad \forall (r,\ell) :\, r \neq \ell. \tag{5.24}$$

We recall that $\mathrm{Mis}_r(X; \widetilde{X})$ is the misclassification rate over the rth cluster of X (Sec-

tion 5.1).

Proposition 2. *Let $X, \widetilde{X} \in \mathbb{M}^k_{n,\,d}$ be two k-means matrices, and write $n_r = n_r(X)$, $n_\wedge = n_\wedge(X)$ and $\delta_r = \delta_r(X)$. Assume that $d_F(X, \widetilde{X}) \le \varepsilon$ and*

(a) X has exactly k nonempty clusters, and

(b) $c_r^{-2}\,\varepsilon^2/(\delta_r^2 n_r) < 1$ for $r \in [k]$, and constants $c_r > 0$ such that $c_r + c_\ell \le 1$, $r \ne \ell$.

Then, $\widetilde{X}$ has exactly k clusters and

$$\mathrm{Mis}_r(X; \widetilde{X}) \le \frac{c_r^{-2}\,\varepsilon^2}{n_r\,\delta_r^2}, \quad \forall r \in [k]. \tag{5.25}$$

In particular, under the conditions of Proposition 2 with $c_r = 1/2$, we have

$$\mathrm{Mis}_\infty(X, \widetilde{X}) \le \frac{4\,\varepsilon^2}{\min_r n_r \delta_r^2} \le \frac{4\,\varepsilon^2}{n_\wedge \delta_\wedge^2}, \qquad \overline{\mathrm{Mis}}(X, \widetilde{X}) \le \frac{4\,\varepsilon^2}{n \delta_\wedge^2}.$$

where the second one follows from the identity $\overline{\mathrm{Mis}}(X, \widetilde{X}) = \sum_{r=1}^k (n_r/n)\,\mathrm{Mis}_r(X, \widetilde{X})$. The proof follows the argument in [LR15, Lemma 5.3] which is further attributed to [Jin15]. Combining Proposition 2 with (5.21), we obtain the following corollary:

Corollary 2. *Let $X^* \in \mathbb{M}^k_{n,\,d}$ be a k-means matrix, and write $n_r = n_r(X^*)$, $n_\wedge = n_\wedge(X^*)$ and $\delta_r = \delta_r(X^*)$. Assume that $\hat{X} \in \mathbb{R}^{n \times d}$ is such that $d_F(X^*, \hat{X}) \le \varepsilon$ and*

(a) X^* has exactly k nonempty clusters, and

(b) $c_r^{-2}\,(1+\kappa)^2\varepsilon^2/(\delta_r^2 n_r) < 1$ for $r \in [k]$, and constants $c_r > 0$ such that $c_r + c_\ell \le 1$, $r \ne \ell$.

Then, any $\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})$ has exactly k clusters and

$$\mathrm{Mis}_r(X^*; \mathcal{P}_\kappa(\hat{X})) \le \frac{c_r^{-2}\,(1+\kappa)^2\varepsilon^2}{n_r\,\delta_r^2}, \quad \forall r \in [k]. \tag{5.26}$$

As before, $\mathrm{Mis}_r(X^*, \mathcal{P}_\kappa(\hat{X}))$ should be interpreted as $\max_{\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})} \mathrm{Mis}_r(X^*, \widetilde{X})$, that is, the result hold for any κ-approximate $k$means solution for $\hat{X}$. In particular, under the

conditions of Corollary 2 with $c_r = 1/2$, we have

$$\mathrm{Mis}_\infty(X^*, \mathcal{P}_\kappa(\hat{X})) \le \frac{4(1+\kappa)^2\,\varepsilon^2}{\min_r n_r\delta_r^2} \le \frac{4\,(1+\kappa)^2\varepsilon^2}{n_\wedge \delta_\wedge^2}, \tag{5.27}$$

$$\overline{\mathrm{Mis}}(X^*, \mathcal{P}_\kappa(\hat{X})) \le \frac{4\,(1+\kappa)^2\varepsilon^2}{n\delta_\wedge^2}. \tag{5.28}$$

Proof of Corollary 2. Using (5.21), we have $d_F(X^*, \widetilde{X}) \le (1+\kappa)\varepsilon$ for any $\widetilde{X} \in \mathcal{P}_\kappa(\hat{X})$. We now apply Proposition 2 to X^* and $\widetilde{X}$, both k-means matrices, with $(1+\kappa)\varepsilon$ in place of ε. □

Proof of Proposition 2. Let $\mathcal{C}_r$ denote the rth cluster of X, having center $q_r = q_r(X)$. We have $|\mathcal{C}_r| = n_r$. Let x_i^T and $\widetilde{x}_i^T$ be the ith row of X and $\widetilde{X}$, respectively, and let

$$T_r := \{i \in \mathcal{C}_r :\ d(\widetilde{x}_i, q_r) < c_r\delta_r\} = \{i \in \mathcal{C}_r :\ d(\widetilde{x}_i, x_i) < c_r\delta_r\}$$

using $x_i = q_r$ for all $i \in \mathcal{C}_r$ which holds by definition. Let $S_r = \mathcal{C}_r \setminus T_r$. Then,

$$|S_r| c_r^2\delta_r^2 \ \le \sum_{i\,\in\, S_r} d(\widetilde{x}_i, x_i)^2 \le \varepsilon^2 \implies \frac{|S_r|}{|\mathcal{C}_r|} \le \frac{c_r^{-2}\varepsilon^2}{n_r\delta_r^2} < 1. \tag{5.29}$$

where we have used assumption (b). It follows that S_r is a proper subset of $\mathcal{C}_r$, that is, T_r is nonempty for all $r \in [k]$.

Next, we argue that if two elements belong to different $T_r, r \in [k]$, they have different labels according to $\widetilde{X}$. That is, $i \in T_r$, $j \in T_\ell$ for $r \ne \ell$ implies $\widetilde{x}_i \ne \widetilde{x}_j$. Assume otherwise, that is, $\widetilde{x}_i = \widetilde{x}_j$. Then, by triangle inequality and $c_r + c_\ell \le 1$,

$$d(q_k, q_\ell) \le d(q_k, \widetilde{x}_i) + d(q_\ell, \widetilde{x}_j) < c_r\delta_r + c_\ell\delta_\ell \le \max\{\delta_r, \delta_\ell\}$$

contradicting (5.24). This shows that $\widetilde{X}$ has at least k labels, since all T_r are nonempty, hence exactly k labels, since $\widetilde{X} \in \mathbb{M}_{n,\,d}^k$ by assumption.

Finally, we argue that if two elements belong to the same T_r, they have the same label according to $\widetilde{X}$. This immediately follows from the previous step since otherwise there will be at least $k+1$ labels. Thus, we have shown that, for all $r \in [k]$, the labels in each T_r are in the same cluster according to both X and $\widetilde{X}$, that is, they are correctly classified. The

Algorithm 3 SC-1

1: Apply degree-reduction regularization of Theorem 3 to A to obtain $A_{\rm re}$.
2: Obtain the k-truncated SVD of $A_{\rm re}$ as $A_{re}^{(k)} = \widehat{Z}_1 \hat{\Sigma} \widehat{Z}_2^T$. See (5.13).
3: Obtain a κ-approximate $k$means solution for input $\widehat{Z}_1$, that is, $\mathcal{P}_\kappa(\widehat{Z}_1)$.

misclassification rate over cluster $\mathcal{C}_r$ is then $\leq |S_r|/|\mathcal{C}_r|$ which establishes the result in view of (5.29). □

5.4 Consistency results

We now state our various consistency results. We start with a refinement of Theorem 2 for the specific algorithm SC-1 given in Algorithm 3.

Theorem 4. *Consider the spectral algorithm* SC-1 *given in Algorithm 3. Assume* $k_1 = k =: \min\{k_1, k_2\}$, *and for a sufficiently small* $C > 0$,

$$ka\,\sigma_k^{-2} \leq C(1+\kappa)^{-2}.$$

Then, under the SBM model of Section 5.2, w.h.p.,

$$\overline{\rm Mis}(\mathcal{P}_\kappa(\widehat{Z}_1), \bar{Z}_1) \;\lesssim\; (1+\kappa)^2 \beta_1 \Big(\frac{a}{\sigma_k^2}\Big).$$

where β_1 *is given in* (A1) *and* a *is defined in* (5.9).

Proof. Going through the three-step plan of analysis in Section 5.3, we observe that (5.9) holds for $A_{\rm re}$ by Theorem 3, and (5.10) holds by Lemma 2. We only need to verify conditions of Corollary 2, so that κ-approximate $k$means operator $\mathcal{P}_\kappa$ satisfies bound (5.11) of the $k$means step. As in the proof of Theorem 2, $X^* = \bar{Z}_1 O \in \mathbb{R}^{n_1 \times k}$, where $\bar{Z}_1 \in \mathbb{O}^{n_1 \times k_1}$ and $O := U_\psi Q \in \mathbb{O}^{k_1 \times k}$. Clearly, X^* has exactly k distinct rows (recalling $k = k_1$). Furthermore, using the calculation in the proof of Theorem 2,

$$n_{1t}\,\delta_t^2 = n_{1t} \min_{s:\, s \neq t} (n_{1t}^{-1} + n_{1s}^{-1}) = \min_{s:\, s \neq t} (1 + \frac{n_{1t}}{n_{1s}}) \geq 1.$$

Algorithm 4 SC-RR

1: Apply degree-reduction regularization of Theorem 3 to A to obtain $A_{\rm re}$.
2: Obtain the best rank k approximation of $A_{\rm re}$, that is, $A_{re}^{(k)} = \mathscr{T}_k(A_{\rm re})$. See (5.13).
3: Output $\in \mathcal{P}_\kappa(A_{re}^{(k)})$, i.e., a κ-approximate $k$means solution for input $A_{re}^{(k)}$.

Recalling that $\varepsilon^2 = C_2^2 ka/\sigma_k^2$, as long as

$$4(1+\kappa)^2\varepsilon^2 = 4C_2^2(1+\kappa)^2\, ka/\sigma_k^2 < 1 \le n_{1t}\,\delta_t^2$$

condition (b) of Corollary 2 holds and $\mathcal{P}_\kappa$ satisfies (5.11) with $c = 4(1+\kappa)^2$ as in (5.27). The rest of the proof follows as in Theorem 2. □

One can take κ to be a fixed small constant say 1.5, since there are κ-approximate $k$means algorithms for any $\kappa > 1$. In that case, $(1+\kappa)^2$ can be absorbed into other constants, and the bound in Theorem 4 is qualitatively similar to Theorem 2.

Reduced-rank SC. We now consider a variant of SC suggested in [YP14; GLMZ16], where one uses the entire rank k approximation of $A_{\rm re}$, and not just the singular vector matrix $\widehat{Z}_1$, as the input to the k-means step. The approach, which we call reduced-rank SC, or SC-RR, is detailed in Algorithm 4. Recall the SV truncation operator $\mathscr{T}_k$ given in (5.13). It is well-known that $\mathscr{T}_k$ maps every matrix to its best rank-k approximation in Frobenius norm, i.e.,

$$\mathscr{T}_k(A_{\rm re}) = \min\left\{\|R - A_{\rm re}\|_F :\ \mathrm{rank}(R) \le k\right\}$$

with the approximation error satisfying

$$\|\mathscr{T}_k(A_{\rm re}) - A_{\rm re}\| = \sigma_{k+1}(A_{\rm re}). \tag{5.30}$$

SC-RR uses this best rank-k approximation as a denoised version of $A_{\rm re}$ and runs a k-means algorithm on its rows. To analyze SC-RR, we need to replace bound (5.10) in Step 2 with an appropriate modification. The following lemma replaces Lemma 2 and provides the necessary bound in this case.

Lemma 3. *Let $A_{re}^{(k)} = \mathscr{T}_k(A_{re})$ be the k-truncated SVD of A_{re} and assume that the concentration bound (5.9) holds. Then,*

$$\|A_{re}^{(k)} - P\|_F \leq C\sqrt{8\,ka}. \tag{5.31}$$

Proof. Throughout the proof, let $\|\cdot\| = \|\cdot\|$ be the operator norm. Recall that $P = \mathbb{E}A$ is the mean matrix itself, and let $\Delta_{\text{re}} := A_{\text{re}} - P$. By Weyl's theorem on the perturbation of singular values, $|\sigma_i(A_{\text{re}}) - \sigma_i(P)| \leq \|\Delta_{\text{re}}\|$ for all i. Since $\sigma_{k+1}(P) = 0$ (see (5.8)), we have $\sigma_{k+1}(A_{\text{re}}) \leq \|\Delta_{\text{re}}\|$, hence

$$\begin{aligned}
\|A_{re}^{(k)} - P\| &\leq \|A_{re}^{(k)} - A_{\text{re}}\| + \|\Delta_{\text{re}}\| && \text{(triangle inequality)} \\
&= \sigma_{k+1}(A_{\text{re}}) + \|\Delta_{\text{re}}\| && \text{(by (5.30))} \\
&\leq 2\|\Delta_{\text{re}}\| && \text{(Weyl's theorem).}
\end{aligned}$$

Thus, in terms of the operator norm, we lose at most a constant in going from A_{re} to $A_{re}^{(k)}$. However, we gain a lot in Frobenious norm deviation. Since A_{re} is full-rank in general, the best bound on Δ_{re} based on its operator norm is $\|\Delta_{\text{re}}\|_F \leq \sqrt{n_\wedge}\,\|\Delta_{\text{re}}\|$ where $n_\wedge = \min\{n_1, n_2\}$. On the other hand, since $A_{re}^{(k)} - P$ is of rank $\leq 2k$, we get

$$\|A_{re}^{(k)} - P\|_F \leq \sqrt{2k}\|A_{re}^{(k)} - P\| \leq 2\sqrt{2k}\|\Delta_{\text{re}}\|.$$

Combining with (5.9), that is, $\|\Delta_{\text{re}}\| \leq C\sqrt{a}$, we have the result. □

Comparing with (5.10), we observe that (5.31) provides an improvement by removing the dependence on the singular value gap σ_k. However, we note that in terms of the relative error, i.e., $\|A_{re}^{(k)} - P\|_F / \|P\|_F$ this may or may not be an improvement. There are cases where $\|P\|_F \approx \sqrt{k}\sigma_k$, in which case the relative error predicted by (5.31) is $O(\sqrt{a}/\sigma_k)$, similar to the relative error based on (5.10); see Example 3 below.

Following through the three-step analysis of Section 5.3.1, with (5.10) replaced with (5.31), we obtain a qualitatively different bound on the misclassification error of Algorithm 4. The key is that center separation of P treated as a k-means matrix is different from that of $\bar{Z}_1 U_\psi Q$. Note that P is indeed a valid k-means matrix according to Definition (5.15); in

fact, $P \in \mathbb{M}^{k_1}_{n_1,n_2}$. Similarly, $P^T \in \mathbb{M}^{k_2}_{n_2,n_1}$. Let us define

$$\Psi^2_{1,\wedge} := \min_{(s,t):\, s\neq t} \sum_{\ell=1}^{k_2} \pi_{2\ell}(\Psi_{s\ell} - \Psi_{t\ell})^2, \tag{5.32}$$

$$\widetilde{\Psi}^2_{1,\wedge} := \min_{(s,t):\, s\neq t} \Big[\pi_{1t} \sum_{\ell=1}^{k_2} \pi_{2\ell}(\Psi_{s\ell} - \Psi_{t\ell})^2\Big]. \tag{5.33}$$

Theorem 5. *Consider the spectral algorithm* SC-RR *given in Algorithm 4. Assume that for a sufficiently small $C_1 > 0$,*

$$ka\, \widetilde{\Psi}^{-2}_{1,\wedge} \le C_1(1+\kappa)^{-2}. \tag{5.34}$$

Then, under the SBM model of Section 5.2, w.h.p.,

$$\overline{\mathrm{Mis}}(\mathcal{P}_\kappa(A^{(k)}_{re}), P) \;\le\; C_1^{-1}\,(1+\kappa)^2\Big(\frac{ka}{\Psi^2_{1,\wedge}}\Big).$$

where a is defined in (5.9).

Proof. We only need to calculate $\delta(P) = \delta_\wedge(P)$ the minimum center separation of P viewed as an element of $\mathbb{M}^{k_1}_{n_1,n_2}$. Recall that

$$P = \bar{Z}_1 \bar{B} \bar{Z}_2^T = Z_1 N_1^{-1/2} (\bar{N}_1^{1/2} \Psi \bar{N}_1^{1/2}) \bar{Z}_2^T = n_1^{-1/2} Z_1 \Psi \bar{N}_1^{1/2} \bar{Z}_2^T.$$

Let e_s be the sth standard basis vector of $\mathbb{R}^{k_1}$. Unique rows of P are $q_s^T := n_1^{-1/2} e_s^T (\Psi \bar{N}_2^{1/2}) \bar{Z}_2^T$ for $s \in [k_1]$. We have

$$\begin{aligned}\|q_s - q_t\|_2^2 &= n_1^{-1} \|\bar{Z}_2 \bar{N}_2^{1/2} \Psi^T (e_s - e_t)\|_2^2 \\ &= n_1^{-1} \|\bar{N}_2^{1/2} \Psi^T (e_s - e_t)\|_2^2 = n_1^{-1} \sum_{\ell=1}^{k_2} \pi_{2\ell}(\Psi_{s\ell} - \Psi_{t\ell})^2.\end{aligned}$$

It follows that $\delta^2(P) = \min_{t\neq s} \|q_s - q_t\|_2^2 = n_1^{-1}\Psi^2_{1,\wedge}$. We apply Corollary 2, with $X^* = P$ and $\hat{X} = A^{(k)}_{re}$, taking $\varepsilon^2 = 8C^2 ka$ according to Lemma 3. Condition (b) of the corollary

holds if

$$32C^2(1+\kappa)^2 ka = 4(1+\kappa)^2\varepsilon^2 < n_{1t}\,\delta_t^2(P) = \pi_{1t} \min_{s:\, s\neq t} \sum_{\ell=1}^{k_2} \pi_{2\ell}(\Psi_{s\ell} - \Psi_{t\ell})^2$$

for all $t \in [k_1]$, which is satisfied under assumption (5.34). Corollary 2, and specifically (5.27) gives the desired bound on misclassification rate $\leq 4(1+\kappa)^2\varepsilon^2/(n_1\delta^2(P))$. □

As is clear from the proof, one can take $C_1 = 1/(32C^2)$ where C is the constant in concentration bound (5.9). Condition (5.34) can be replaced with the stronger assumption

$$ka\,(\pi_{1,\wedge}\Psi_{1,\wedge}^2)^{-1} \leq C_1(1+\kappa)^{-2} \tag{5.35}$$

where $\pi_{1,\wedge} := \min_{t\in[k_1]} \pi_{1t}$, since $\widetilde{\Psi}_{1,\wedge}^2 \geq \pi_{1,\wedge}\Psi_{1,\wedge}^2$.

Although the bounds of Theorems 4 and 5 are different, surprisingly, in the case of the planted partition model, they give the same result as the next example shows.

Example 3 (Planted partition model, symmetric case)**.** Let us consider the simplest symmetric SBM, the symmetric balanced planted partition (SBPP) model, and consider the consequences of Theorems 4 and 5 in this case. Recall that in the symmetric case we drop index r from k_r, n_r, n_{rj}, $\bar{N}_r$, β_r, $\Psi_{r,\wedge}$ and so on. SBPP is characterized by the following assumptions:

$$\Psi = bE_k + (a-b)I_k, \quad a \geq b, \quad \pi_j = n_j/n = \frac{1}{k},\ \forall j \in [k].$$

Here, $E_k \in \mathbb{R}^{k\times k}$ is the all ones matrix and *balanced* refers to all the communities being of equal size, leading to cluster proportions $\pi_j = 1/k$. In particular, $\beta = 1$, as defined in (A1). We have $\bar{B} = \bar{N}^{1/2}\Psi\bar{N}^{1/2} = \Psi/k$, recalling $\bar{N} = \mathrm{diag}(\pi_j)$. Hence, the smallest singular value of $\bar{B}$ is $\sigma_k = (a-b)/k$. Theorem 4 gives the following result:

Corollary 3. *Under the SBPP model, as long as $k^3a/(a-b)^2$ is sufficiently small,* SC-1 *has average misclassification error of $O(k^2a/(a-b)^2)$ with high probability.*

Now consider SC-RR. Using definitions (5.32), we have $k\widetilde{\Psi}_\wedge^2 = \Psi_\wedge^2 = 2(a-b)^2/k$. Then, Theorem 5 gives the exact same result for SC-RR:

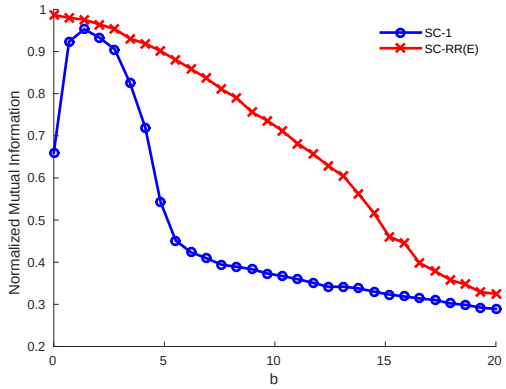


Figure 5.1: An example of the performance boost of SC-RR (or SC-RRE) relative to SC-1. The data is generated from the bipartite version of Example 4 with $n_2 = 2n_1 = 1000$, $k_1 = k_2 = 4$, $\pi_{r\ell} = n_r/k_r$ for all $\ell \in [k_r]$, $r = 1, 2$, and $\Psi = 2bE_4 + \mathrm{diag}(16, 16, 16, 2)$ similar to (5.36). The key is the significant difference in the two smallest diagonal elements of Ψ. The plot shows the normalized mutual information (a measure of cluster quality) between the output of the two spectral clustering algorithms and the true clusters, as b varies. Only row clusters are considered. The plot shows a significant improvement for SC-RR(E) relative to SC-1 over a range of b. As b increases, the relative difference between Ψ_{33} and Ψ_{44} reduces and the model approaches that of Example 3, leading to similar performances for both algorithms as expected. It is interesting to note that the monotone nature of the performance of SC-RR(E) as a function of b and the non-monotone nature of that of SC-1 is reflected in the upper bounds (5.37) and (5.38).

Corollary 4. *Corollary one holds with* SC-1 *replaced with* SC-RR.

Results of Corollary 3 and 4 are consistency results as long as $k^2 a/(a-b)^2 = o(1)$. A typical example is when $k = O(1)$, $a = a_0 f_n$, $b = b_0 f_n$, $a_0 \asymp 1$ and $b_0 \asymp 1$ for some $f_n \to \infty$ as $n \to \infty$. Then, SC-1 and SC-RR are both consistent at a rate $O(f_n^{-1})$. □

Let us now give an example where SC-1 and SC-RR behave differently.

Example 4. Consider the symmetric balanced SBM, with

$$\Psi = bE_k + \mathrm{diag}(\alpha_1, \ldots, \alpha_k), \quad \pi_j = n_j/n = \frac{1}{k}, \ \forall j \in [k]. \tag{5.36}$$

As in Example 3, we have dropped the index r determining the side of network in the bipartite case. Let us assume that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_k \geq 0$. We have

$$k\widetilde{\Psi}^2_\wedge = \Psi^2_\wedge = k^{-1} \min_{s \neq t} \sum_\ell (\Psi_{s\ell} - \Psi_{t\ell})^2 = k^{-1}(\alpha^2_{k-1} + \alpha^2_k).$$

Thus, Theorem 5 gives the following: With ρ defined as follows:

$$\rho := k^2 \frac{\alpha_1 + b}{\alpha_{k-1}^2 + \alpha_k^2}, \tag{5.37}$$

as long as $k\rho$ is sufficiently small, SC-RR has average misclassification error $O(\rho)$ with high probability.

To determine the performance of SC-1, we need to estimate σ_k, the smallest singular value of $\bar{B} = \bar{N}^{1/2}\Psi\bar{N}^{1/2} = \Psi/k$. Since Ψ is obtained by a rank-one perturbation of a diagonal matrix, it is well-known that when $\{\alpha_t\}$ are distinct, the eigenvalues of Ψ are obtained by solving $\sum_{t=1}^{k} 1/(\alpha_t - \lambda) = -1/b$; the case where some of the $\{\alpha_t\}$ are repeated can be reasoned by the taking the limit of the general case. By plotting $\lambda \mapsto \sum_{t=1}^{k} 1/(\alpha_t - \lambda)$ and looking at the intersection with $\lambda \mapsto -1/b$, one can see that the smallest eigenvalue of Ψ, equivalently its smallest singular value, is in $[\alpha_k, \alpha_{k-1}]$, and can be made arbitrarily close to α_k by letting $b \to 0$. Letting $\alpha_k + \varepsilon_k(\alpha; b)$ denote this smallest singular value, we have $0 \le \varepsilon_k(\alpha; b) \to 0$ as $b \to 0$.

It follows that $\sigma_k = \sigma_k(\bar{B}) = k^{-1}(\alpha_k + \varepsilon_k(\alpha; b))$. Theorem 4 gives the following: With ρ defined as

$$\rho := k^2 \frac{\alpha_1 + b}{(\alpha_k + \varepsilon_k(\alpha; b))^2}, \tag{5.38}$$

as long as $k\rho$ is sufficiently small, SC-1 has average misclassification error $O(\rho)$ with high probability.

Comparing (5.38) with (5.37), the ratio of the two bounds is $(\alpha_{k-1}^2 + \alpha_k^2)/(\alpha_k + \varepsilon_k(\alpha; b))^2 \to 1 + (\alpha_{k-1}/\alpha_k)^2$ as $b \to 0$. This ratio could be arbitrarily large depending on the relative sizes of α_k and α_{k-1}. Thus, when the bounds give an accurate estimate of the misclassification rates of SC-1 and SC-RR, we observe that SC-RR has a clear advantage. This is empirically verified in Figure 5.1, for moderately dense cases. (In the very sparse case, the difference is not very much empirically.) In general, we expect SC-RR to perform better when there is a large gap between σ_k and σ_{k-1}, the two smallest nonzero singular values of $\bar{B}$.

Algorithm 5 SC-RRE

1: Apply degree-reduction regularization of Theorem 3 to A to obtain A_{re}.
2: Obtain $A_{re}^{(k)} = \widehat{Z}_1 \hat{\Sigma} \widehat{Z}_2^T$, the k-truncated SVD of A_{re}.
3: Output $\mathscr{K}(\widehat{Z}_1 \hat{\Sigma})$ where $\mathscr{K}$ is an isometry-invariant κ-approximate $k$means algorithm.

Efficient reduced-rank SC. The SC-RR algorithm discussed above has the disadvantage of running a k-means algorithm on vectors in $\mathbb{R}^n$ (the rows of $A_{re}^{(k)}$, or in the ideal case the rows of P). We now introduce a variant of this algorithm that has the same performance as SC-RR in terms of misclassification rate, while computationally is as efficient as SC-1. This approach which we call efficient reduced-rank spectral clustering, SC-RRE, is detailed in Algorithm 5. The efficiency comes from running the k-means step on vectors in $\mathbb{R}^k$ which is usually a much smaller space than $\mathbb{R}^n$ ($k \ll n$ in applications).

For the k-means step in SC-RRE, we need a k-means (type) algorithm $\mathscr{K}$ that only uses the pairwise distances between the data points. We call such k-means algorithms *isometry-invariant*:

Definition 3. A k-means (type) algorithm $\mathscr{K}$ is isometry-invariant if for any two matrices $X^{(r)} \in \mathbb{R}^{n \times d_r}, r = 1, 2$, with the same pairwise distances among points—i.e., $d(x_i^{(1)}, x_j^{(1)}) = d(x_i^{(2)}, x_j^{(2)})$ for all distinct $i, j \in [n]$, where $(x_i^{(r)})^T$ is the ith row of $X^{(r)}$—one has

$$\overline{\text{Mis}}(\mathscr{K}(X^{(1)}), \mathscr{K}(X^{(2)})) = 0.$$

Although the rows of $\mathscr{K}(X^{(1)})$ and $\mathscr{K}(X^{(2)})$ lie in spaces of possibly different dimensions, it still makes sense to talk about their relative misclassification rate, since this quantity only depends on the membership information of the k-means matrices and not their center information. We have implicitly assumed that $d(\cdot, \cdot)$ defines a family of distances over all Euclidean spaces $\mathbb{R}^d, d = 1, 2, \ldots$. This is obviously true for the common choice $d(x, y) = \|x - y\|_2$. If algorithm $\mathscr{K}$ is randomized, we assume that the same source of randomness is used (e.g., the same random initialization) when applying to either of the two cases $X^{(1)}$ and $X^{(2)}$.

The following result guarantees that SC-RRE behaves the same as SC-RR when one uses an isometry-invariant approximate k-means algorithm in the final step.

Theorem 6. *Consider the spectral algorithm* SC-RRE *given in Algorithm 5. Assume that for a sufficiently small $C_1 > 0$, (5.34) holds. Then, under the SBM model of Section 5.2, w.h.p.,*

$$\overline{\mathrm{Mis}}(\mathscr{K}(\widehat{Z}_1\hat{\Sigma}), P) \;\le\; C_1^{-1}\,(1+\kappa)^2\Big(\frac{ka}{\Psi_{1,\wedge}^2}\Big).$$

Proof. Recall that $A_{re}^{(k)} = \widehat{Z}_1\hat{\Sigma}\widehat{Z}_2^T$ is the k-truncated SVD of A_{re}. Let $X^{(1)} = \widehat{Z}_1\hat{\Sigma}$ and $X^{(2)} = A_{re}^{(k)}$, and let $(x_i^{(1)})^T$ and $(x_i^{(2)})^T$ be their ith rows, respectively. Then,

$$\|x_2^{(i)} - x_2^{(j)}\| = \|\widehat{Z}_2(x_1^{(i)} - x_1^{(j)})\|_2 = \|x_1^{(i)} - x_1^{(j)}\|_2, \quad \forall i \neq j,$$

using $\widehat{Z}_2 \in \mathbb{O}^{n_2\times k}$ and (5.1). Isometry-invariance of $\mathscr{K}$ implies $\overline{\mathrm{Mis}}(\mathscr{K}(\widehat{Z}_1\hat{\Sigma}), \mathscr{K}(A_{re}^{(k)})) = 0$. Since $\overline{\mathrm{Mis}}$ is a pseudo-metric on k-means matrices, using the triangle inequality, we get

$$\begin{aligned}\overline{\mathrm{Mis}}\left(\mathscr{K}(\widehat{Z}_1\hat{\Sigma}), P\right) &\le \overline{\mathrm{Mis}}\left(\mathscr{K}(\widehat{Z}_1\hat{\Sigma}), \mathscr{K}(A_{re}^{(k)})\right) + \overline{\mathrm{Mis}}\left(\mathscr{K}(A_{re}^{(k)}), P\right)\\ &= \overline{\mathrm{Mis}}\left(\mathscr{K}(A_{re}^{(k)}), P\right).\end{aligned}$$

(In fact, using the triangle inequality in the other direction, we conclude that the two sides are equal.) The result now follows from Theorem 5. □

5.4.1 Results in terms of mean parameters

One useful aspect of SC-RR(E) is that one can state its corresponding consistency result in terms of the *mean parameters* of the block model. Such results are useful when comparing to the optimal rates achievable in recovering the clusters. The row mean parameters of the SBM in Section 5.2 are defined as $\Lambda_{s\ell} := B_{s\ell}\, n_{2\ell}$ for $(s,\ell) \in [k_1]\times[k_2]$ which we collect in a matrix $\Lambda = (\Lambda_{s\ell}) \in \mathbb{R}^{k_1\times k_2}$. To get an intuition for Λ note that

$$\mathbb{E}[AZ_2] = PZ_2 = Z_1BN_2 = Z_1\Lambda.$$

Each row of AZ_2 is obtained by summing the corresponding row of A over each of the column clusters to get a k_2 vector. In other words, the rows of AZ_2 are the sufficient statistics for estimating the row clusters, had we known the true column clusters. Note that

$\mathbb{E}[(AZ_2)_{i*}] = z_{1i}^T \Lambda$, where the notation $(\cdot)_{i*}$ denotes the ith row of a matrix. In other words, we have $\mathbb{E}[(AZ_2)_{i*}] = \Lambda_{s*}$ if node i belongs to row cluster s. Let us define the minimum separation among these row mean parameters:

$$\Lambda_\wedge^2 := \min_{t \neq s} \|\Lambda_{s*} - \Lambda_{t*}\|^2. \tag{5.39}$$

We have the following corollary of Theorem 5 which is proved in Appendix A.8.

Corollary 5. *Assume that* $\pi_{r,\wedge} := \min_t \pi_{rt} \geq (\beta_r k_r)^{-1}$ *for* $r = 1, 2$, *and let* $k = \min\{k_1, k_2\}$ *and* $\alpha = n_2/n_1$. *Consider the spectral algorithm* SC-RR *given in Algorithm 4. Assume that for a sufficiently small* $C_1 > 0$,

$$\beta_1 \beta_2 \, k \, k_1 k_2 \, \alpha \frac{\|\Lambda\|_\infty}{\Lambda_\wedge^2} \leq C_1 (1+\kappa)^{-2}. \tag{5.40}$$

Then, under the SBM model of Section 5.2, w.h.p.,

$$\overline{\mathrm{Mis}}(\mathcal{P}_\kappa(A_{re}^{(k)}), P) \;\leq\; C_1^{-1} \, (1+\kappa)^2 \beta_2 \, k \, k_2 \, \alpha \frac{\|\Lambda\|_\infty}{\Lambda_\wedge^2}.$$

We note that the exact same result as Corollary 5 holds for SC-RRE assuming the k-means step uses an isometry invariant algorithm as discussed in Section 5.4.

CHAPTER 6

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 A, iteratively. In this section, we introduce an algorithm, namely Algorithm 3, that partitions A into subblocks and operates iteratively on collections of these blocks to maintain the desired independence. For this version of the pseudo-likelihood algorithm, 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 notations. Let us write

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

to denote labels obtained by applying the spectral clustering, respectively, on rows and columns of the adjacency matrix A, the details of which are discussed in Chapter 5 below. We have $\text{colSC}(A) = \text{rowSC}(A^T)$. We also recall the LR classifier defined in (4.14). 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 subblock 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 Chapter 7.1—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 $\tilde{y}^{(q)} \leftarrow \text{LR}(A^{(q,q+2)}, \hat{\Lambda}^{(q+2)}, \tilde{z}^{(q+2)})$, 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

Algorithm 3 Provable 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_{top} 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: $\widehat{y}_{\text{top}} \leftarrow \text{LR}(A_{\text{top}}, \hat{\Lambda}, \tilde{z})$.

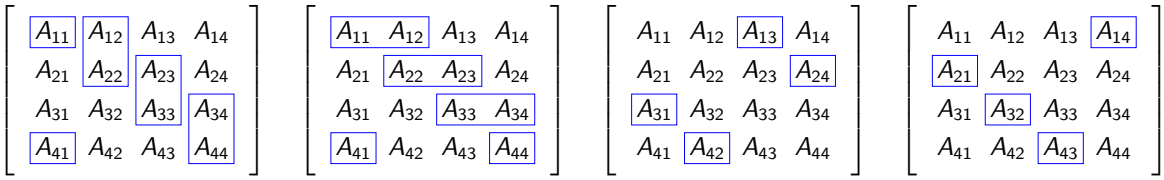
12: Swap A_{top} and A_{bottom}, then repeat steps 2–9 to obtain $\widehat{y}_{\text{bottom}}$.

13: $\widehat{y} \leftarrow [\widehat{y}_{\text{top}}\,;\, \widehat{y}_{\text{bottom}}]$.

14: Apply step 1 to 10 on A^T to obtain $\widehat{z}$.

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.

Figure 6.1 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 6.1(a). Note that $\{G_1^{\text{col}}, G_2, G_3\}$ form a complete partition of the matrix into disjoint blocks. Also, G_1^{col} and G_1^{row} involve the same elements of the matrix,



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

Figure 6.1: 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 the 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).

i.e. they *cover* the same portion of A. Thus, $\{G_1^{\text{row}}, G_2, G_3\}$ is also a complete cover of A with disjoint blocks. Let us write G_1 for the common portion of A 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 6.1(a) and obtain the label vectors (from the leftmost submatrix to the rightmost one):

$$[\tilde{y}'^{(1)}\,;\,\tilde{y}^{(4)}],\ [\tilde{y}^{(1)}\,;\,\tilde{y}'^{(2)}],\ [\tilde{y}^{(2)}\,;\,\tilde{y}'^{(3)}],[\tilde{y}^{(3)}\,;\,\tilde{y}'^{(4)}]. \tag{6.1}$$

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 subblock labels (6.1) 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 Chapter 6.1. The resulting updated global row and column labels only depend on G_1 portion of A. 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 6.1(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.

First LR classifier (steps 7–8): Using the estimates of the (local) row mean parameters, in Step 7, we apply the LR classifier, $\tilde{y}^{(q)} \leftarrow \mathrm{LR}(A^{(q,q+2)}, \hat{\Lambda}^{(q+2)}, \tilde{z}^{(q+2)})$ to each of the submatrices in G_2 (in Figure 6.1(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 the uniform consistency of LR classifier in parameters (Lemma 5). However, we note that $\tilde{z}^{(q+2)}$ is a function of G_1 blocks of A, 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 6.1(d)). This step is similar to step 6, but due to the much more accurate labels, the parameter estimates are much more accurate as well. 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. It is worth recalling that the true local mean parameters, do not depend on the block row index; see (4.6).

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 $\widehat{y}_{\mathrm{top}} \leftarrow \mathrm{LR}(A_{\mathrm{top}}, \hat{\Lambda}, \tilde{z})$ 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 $\approx \exp(-I)$), as argued in Chapter 8.

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. No matching is required in step 13 when concatenating the top and bottom pieces to form a global set of row labels $\widehat{y}$. This is because the LR classifiers produce the same cluster labels; see Chapter 8.

6.1 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

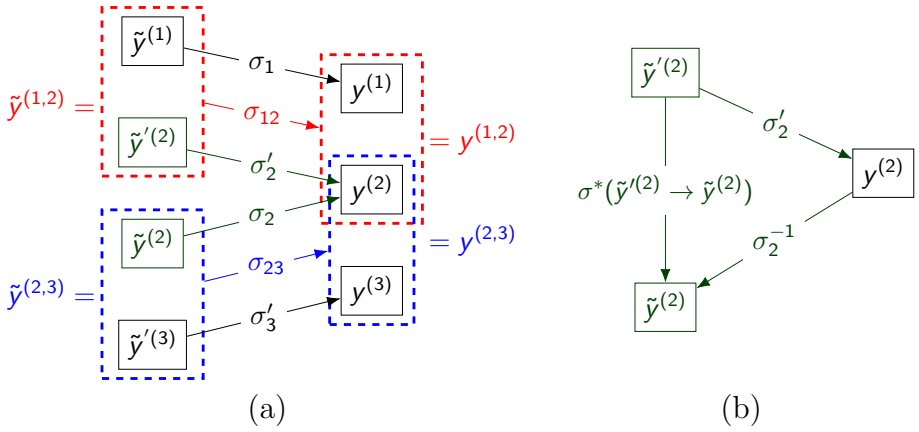


Figure 6.2: Pictorial depiction of the matching step. (a) Two-block and subblock optimal permutations to the truth. When $\tilde{y}^{(1,2)} \approx y^{(1,2)}$, we have $\sigma_1 = \sigma'_2 = \sigma_{1,2}$ and similarly $\tilde{y}^{(2,3)} \approx y^{(2,3)}$ implies $\sigma_2 = \sigma'_3 = \sigma_{2,3}$. (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 Chapter 6.1 for details.

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 also 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. Chapter 5.1). Let us define

$$\sigma_{q-1,q} := \sigma^*\big(\tilde{y}^{(q-1,q)} \to y^{(q-1,q)}\big), \quad \sigma_q := \sigma^*(\tilde{y}^{(q)} \to y^{(q)}), \quad \sigma'_q := \sigma^*(\tilde{y}'^{(q)} \to y^{(q)}). \tag{6.2}$$

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 6.2(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 subblocks 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 subblock has "almost" the same misclassification error as the bigger two-block. To see this, recall the *direct misclassification rate* introduced in Chapter 5.1, i.e., misclassification rate without applying any permutation (or equivalently with the identity permutation). We have

$$\text{dMis}\left(\sigma_{2,3}(\tilde{y}^{(2,3)}), y^{(2,3)}\right) = \text{Mis}\left(\tilde{y}^{(2,3)}, y^{(2,3)}\right) \leq \varepsilon. \tag{6.3}$$

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 (6.3) implies

$$\text{dMis}\left(\sigma_2(\tilde{y}^{(2)}), y^{(2)}\right) = \text{dMis}\left(\sigma_{2,3}(\tilde{y}^{(2)}), y^{(2)}\right) \leq 2\varepsilon = \varepsilon' \tag{6.4}$$

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,

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

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 subblocks as opposed to Mis. Similarly to (6.4), we obtain $\text{dMis}\left(\sigma_3'(\tilde{y}^{(3)}), y^{(3)}\right) \leq \varepsilon'$ considering the second component of $\tilde{y}^{(2,3)}$ and $y^{(2,3)}$. Applying the same argument to indices $(1,2)$, we conclude similarly that $\text{dMis}\left(\sigma_1(\tilde{y}^{(1)}), y^{(1)}\right) \leq \varepsilon'$ and $\text{dMis}\left(\sigma_2'(\tilde{y}'^{(2)}), y^{(2)}\right) \leq \varepsilon'$.

Now consider the following three block vector undergoing 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 $\leq$ 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 data, we can construct a consistent three-block label having Mis $\leq \varepsilon'$.

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

$$\sigma_2^{-1} \circ \sigma_2' = \sigma^*(\tilde{y}'^{(2)} \to \tilde{y}^{(2)}) \tag{6.5}$$

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 6.2(b). In order to make the above argument precise, we need to justify the first and third claims. We will discuss the details in Chapter 7.4. 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).

6.2 Results for Algorithm 3

6.2.1 General initialization

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

of (B3), with γ_1 satisfying

$$\gamma_1 \leq \Big[\frac{1}{384\beta^2\omega}\Big(\frac{I_{\min}}{8L\|\Lambda\|_\infty} \wedge \frac{I^{\mathrm{col}}_{\min}}{16K\|\Gamma\|_\infty}\Big)\Big] \wedge \frac{1}{4}. \tag{6.6}$$

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 7. *Assume that the model parameters satisfy $I_{min} \wedge I^{col}_{min} \to \infty$, $\Lambda_{\min} \to \infty$, (A3) and (A4$'$), and the initial labels satisfy (6.6). Then, for some $\zeta = o(1)$, $\widehat{y}$ output by Algorithm 3 satisfies*

$$\mathrm{Mis}_k\left(\widehat{y},\, y\right) = O\Big(\omega \sum_{r\neq k}\Big(1 + \frac{1}{\varepsilon_{kr}}\Big)\exp\left(-I_{kr} - \Big(\frac{1}{2} - \zeta\Big)\log\Lambda_{\min}\right)\Big) \tag{6.7}$$

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

We refer to Chapter 3 for the definition of the parameters involved in the rate given in (6.7).

CHAPTER 7

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:

$$\mathrm{Mis}(\tilde{y}, y) \leq \frac{\gamma}{\beta K}, \quad \mathrm{Mis}(\tilde{z}, z) \leq \frac{\gamma}{\beta L}. \tag{B3}$$

We call such labels γ-*good*. 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,

$$\mathscr{B}_\Lambda(\delta) := \{\tilde{\Lambda} : \|\tilde{\Lambda} - \Lambda\|_\infty \leq \delta \|\Lambda\|_\infty\}, \tag{7.1}$$

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.

7.1 Fixed label analysis

We first present the analysis assuming that all the operations are performed on the entire adjacency matrix A. In Chapter 7.2, these results are extended to be applicable to subblocks of A. Recall the definitions of the mean parameters an their estimates from Chapter 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 (4.2) are close to the truth,Λ.

Lemma 4 (Parameter consistency)**.** *Let* $C_\gamma = C_{\gamma,\beta} = \beta^2\gamma/(1-\gamma)$*, assume that* $6C_\gamma\omega \leq 1$*, 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*

$$p_1 = p_1(\tau;\, n, \Lambda_{\min}, \beta) := KL\exp\Big(-\frac{n\Lambda_{\min}\, h_1(\tau)}{4\beta K}\Big), \quad \forall \tau > 0, \tag{7.2}$$

and $\hat{\Lambda}$ *is as defined in* (4.2)*. 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 Chapter 10.1. 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 (4.14) 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 (4.7), and define the following:

$$Y_{ikr}(b_{i*},\tilde{\Lambda}) := \Psi(b_{i*};\tilde{\lambda}_{r*} \mid \tilde{\lambda}_{k*}) = \sum_{\ell=1}^{L} b_{i\ell}\log\frac{\tilde{\lambda}_{r\ell}}{\tilde{\lambda}_{k\ell}} + \tilde{\lambda}_{k\ell} - \tilde{\lambda}_{r\ell}, \tag{7.3}$$

$$Z_{ik}(b_{i*},\tilde{\Lambda}) := 1\{Y_{ikr}(\tilde{\Lambda}) \ge 0, \text{ for some } r \ne k\}. \tag{7.4}$$

$$S_k(\boldsymbol{b},\tilde{\Lambda}) := \frac{1}{n_k(y)}\sum_{i:y_i=k} Z_{ik}(b_{i*},\tilde{\Lambda}), \tag{7.5}$$

where Ψ is the Poisson log-likelihood ratio defined in (4.13). 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 $Y_{ikr}(\tilde{\Lambda}) < 0, \forall r \ne k$ 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

$$J_{kr} = L\|\Lambda\|_\infty / I_{kr} \tag{7.6}$$

and recalling definitions of ε_{kr}, ω and β from Chapter 3, set

$$\eta' := \eta'(\delta;\Lambda) = 8\omega\delta L\|\Lambda\|_\infty = 8\omega\delta J_{kr} I_{kr}, \tag{7.7}$$

$$\begin{aligned}\eta_{kr} &:= \eta_{kr}(\delta;\omega,\beta,m,\Lambda)\\ &= 21\delta\omega L\|\Lambda\|_\infty + \frac{5\beta L^2\|\Lambda\|_\infty^2}{m} + \log\Big[11\omega\Big(\frac{1}{\varepsilon_{kr}-2\omega(1+\varepsilon_{kr})\delta}+1\Big)\Big] - \frac{1}{2}\log\Lambda_{\min}.\end{aligned} \tag{7.8}$$

We have the following key lemma:

Lemma 5 (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 (7.7),

$$\mathbb{P}\big(\exists\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta),\ Y_{ikr}(b_{i*},\tilde{\Lambda}) \geq 0\big) \ \leq\ \exp(-I_{kr}+\eta'). \tag{7.9}$$

(b) If in addition $\varepsilon_{kr} - 2\omega\delta > 0$, then with η_{kr} defined as in (7.8),

$$\mathbb{P}\big(\exists\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta),\ Y_{ikr}(b_{i*},\tilde{\Lambda}) \geq 0\big) \ \leq\ \exp\big(-I_{kr}+\eta_{kr}\big). \tag{7.10}$$

The proof of Lemma 5(b) appears in Chapter 10.5, and that of part (a) in Appendix A.5.

Remark 9 (Typical setting)**.** In the error exponent in Lemma 5(b), i.e. $-I_{kr}+\eta_{kr}$, the first three terms in (7.8) 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 $I_{kr} \le 2L\|\Lambda\|_\infty$. Another important quantity is C_γ in Lemma 4, which in the typical setting behaves as $C_\gamma \asymp \gamma$ when $\gamma \to 0$.

Combining Lemma 5 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 6. *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 (7.9) holds. Then, for any $u \in \mathbb{R}$, we have*

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

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 (7.10) holds.

Remark 10. Edge splitting (ES) was proposed in [AS15] to generate nearly independent copies from a single network. One might ask whether combining the edge splitting idea with Lemma 6 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 6 in the best case scenario, with the most favorable version of η_{kr}—i.e., ignoring the first three positive terms in (7.8)—gives a misclassification rate

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

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 4, $\hat{\Lambda}_2 \in \mathscr{B}_{\Lambda_2}(\delta)$ w.h.p for some $\delta \ge \gamma_2$. By Lemma 6,

and the perturbation of information (Lemma 10) we have

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

for some $u \to \infty$ w.h.p.. To obtain result (3.7) 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,

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

However, this is not sufficient to show $(1-q)\gamma_2\|\Lambda\|_\infty = o(\log\Lambda_{\min})$. Therefore, applying edge splitting and Lemma 6 does not lead to the main result of this paper.

7.2 Analysis on subblocks

We now extend the analysis of Chapter 7.1 to be applicable to the subblocks obtained by random partitioning. Some care needs to be taken since the true (row and column) mean parameters of the subblocks are changed by partitioning, due to the change in the distributions of the labels within each subblock among the $K \times L$ classes. The deviations of the subblock 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 Chapter 8.1.2). 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 (4.10). 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

$$\pi_\ell(z) = \frac{n_\ell(z)}{m}, \quad \pi_\ell(z^{(q)}) = \frac{n_\ell(z^{(q)})}{m/Q}, \quad \text{hence,} \quad \frac{\pi_\ell(z^{(q)})}{\pi_\ell(z)} = Q\frac{n_\ell(z^{(q)})}{n_\ell(z)}, \tag{7.11}$$

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,

$$\Big|\frac{\pi_\ell(z^{(q)})}{\pi_\ell(z)} - 1\Big| \le \xi L\beta \le \frac{1}{2} \implies \frac{1}{2}\frac{1}{\beta L} \le \pi_\ell(z^{(q)}) \le \frac{3}{2}\frac{\beta}{L} \tag{7.12}$$

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

$$p_2 = 2Q(K+L)\exp\big(-\min(n,m)\xi^2/Q\big). \tag{7.13}$$

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

Lemma 8. *Under* (B4a) *and* (B4b)*, the true local mean parameters* $\Lambda^{(q)} = (\lambda_{k\ell}^{(q)})$ *satisfy:*

$$\Big|\lambda_{k\ell}^{(q)} - \frac{\lambda_{k\ell}}{Q}\Big| \le (\xi L\beta)\frac{\lambda_{k\ell}}{Q} \le \frac{1}{2}\frac{\lambda_{k\ell}}{Q}, \quad \forall q, k, \ell. \tag{7.14}$$

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

Our main lemma for the subblocks 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 4. We recall the operator $\mathscr{L}$ from (4.1):

Lemma 9 (Local parameter consistency)**.** *Let* $C_\gamma = \beta^2\gamma/(1-\gamma)$ *and* $h_c(\tau) := \frac{3}{4c}\tau \log\left(1 + \frac{2c}{3}\tau\right)$ *as in Lemma 4 and assume that* $72\, C_\gamma \omega \leq 1$*. 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,*

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

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

$$p_3 = p_3(\tau;\, n, K, \Lambda_{\min}, Q) := KL \exp\left(-\frac{n\Lambda_{\min}\, h_1(\tau)}{32Q^2\beta K}\right). \tag{7.15}$$

We also have

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

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

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

Remark 11. 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 8 can be translated to the column version by making the following substitutions $\Lambda \to \Gamma$, $Q \to 2Q$ and $L \leftrightarrow K$. For Lemma 9, in addition we need to make $n \to 4m$. (The reason for this is that in (A.1), 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 (7.15), that is, p_3 has to be replaced with

$$p_3' := p_3(\tau; 4m, L, \Gamma_{\min}, 2Q) = p_3(\tau; m, L, \Gamma_{\min}, Q). \tag{7.16}$$

7.3 Perturbation of information

Recall the definition of Chernoff information from (3.1), 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 A.1, bounds the perturbations of $I_{kr}(\Lambda)$ in Λ:

Lemma 10. *Under (A1), for any $\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta)$, we have $|I_{kr}(\tilde{\Lambda}) - I_{kr}(\Lambda)| \le 2\omega\delta L\|\Lambda\|_\infty$.*

7.4 Analysis of the matching step

In this section, we fill in the details of the argument sketched in Chapter 6.1. Specifically, we need to give sufficient conditions so that the first and the third claims of Chapter 6.1 hold. We will use the following two lemmas. Recall the notation $\sigma^*(\tilde{y} \to y)$ introduced in Chapter 5.1 to denote the optimal permutation from the set of labels $\tilde{y}$ to another set y.

Lemma 11. *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 11 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 12. *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 follows from Lemma 11, under the further assumption:

$$\mathrm{Mis}(\tilde{y}^{(q-1,q)}, y^{(q-1,q)}) < \frac{1}{32\beta K}, \quad q \in [Q]. \tag{7.17}$$

Using the permutation notations (6.2) of Chapter 6.1, we have:

Corollary 6. *Under assumptions* (A2), (B4a), (B4b) *and* (7.17), $\sigma_{q-1,q} = \sigma_{q-1}$ *for all* $q \in [Q]$.

The third and final claim of Chapter 6.1 follows from Lemmas 11 and 12, by applying them to the subblock labels $y^{(2)}, \tilde{y}^{(2)}, \tilde{y}'^{(2)}$:

Corollary 7. *Under assumptions* (A2), (B4a), (B4b) *and* (7.17), $\sigma_q^{-1} \circ \sigma'_q = \sigma^*(\tilde{y}'^{(q)} \to \tilde{y}^{(q)})$ *for all* $q \in [Q]$.

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

CHAPTER 8

Analysis of Algorithm 3

8.1 Proof of Theorem 7

We start with the high-level analysis of Algorithm 3 in Chapter 8.1.1. 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 Chapter 8.1.2, we make specific choices about these parameters and finish the proof of Theorem 7.

8.1.1 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:

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

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 7, we have $\mathbb{P}(\mathfrak{P}) \geq 1 - p_2$ where p_2 is given in (7.13). 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 $\mathbb{P}(\mathfrak{D}^c \cap \mathfrak{P}) \leq q$. In this case, if $\mathfrak{P}$ holds with high probability, say $\geq 1 - p_2$, and q is small, then $\mathfrak{D}$ holds with high probability as well: $\mathbb{P}(\mathfrak{D}) \geq 1 - q - p_2$.

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

$$\delta_1 := 24\, C_{\gamma_1} + 6\tau_1 + \xi L\beta. \tag{8.1}$$

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 (8.4). 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 9. These parameters will be optimized at the end of the argument (see Chapter 8.1.2).

Lemma 13 (First local parameters)**.** *Assume* (B4b) *and* $72\, C_{\gamma_1}\omega \leq 1$*, and let* δ_1 *be as defined in* (8.1)*. 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* (7.15)*.*

Proof. Conditioning on blocks G_1 (cf. Chapter 6) 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 Chapter 7.2 apply. We will apply Lemma 9 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 9 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 $\leq 2p_3$. Let us write

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

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

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

for which we have the following bound on misclassification rate:

Lemma 14 (First LR classifier)**.** *Under the assumptions of Lemma 13, further assume that* $9\omega\delta_1 < 1$. *Let* $\eta^{step\ 7} := 2\eta'(\delta_1; \Lambda/Q)$ *where* $\eta'(\cdot)$ *is defined in* (7.8). *Then, on event* $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$,

$$\text{Mis}_k\left(\tilde{y}^{(q)}_{step\ 7},\ y^{(q)}\right) \ \leq\ \sum_{r \neq k} \exp\left(-\frac{I_{kr}}{Q} + \eta^{step\ 7} + u\right) =: \gamma^{row}_{2k}, \quad \forall q \in \mathbb{Z}_Q, \tag{8.2}$$

fails with probability at most $Q(e^{-u} + 2Q\, p_3)$ *where* $p_3 = p_3(\tau_1)$ *as given in* (7.15).

Proof. Fix $q \in \mathbb{Z}_Q$ and consider $\tilde{y}^{(q-2)}$. As in the proof Lemma 13, we condition on blocks in G_1 so that $\tilde{z}^{(q)}_{\text{step 5}}$ can be assumed deterministic. We will apply Lemma 5(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 5(b), and we have the needed condition $\Lambda^{(q-2,q)}(y, \tilde{z}) \in \mathscr{B}_{\Lambda^{(q)}}(\delta_1)$ from Lemma 9. 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 5(a) gives

$$\mathbb{P}\Big(\Big\{\exists\, \tilde{\Lambda} \in \mathscr{B}_{\Lambda^{(q)}}(\delta_1),\ Y_{ikr}\big(b^{(q-2,q)}_{i*}, \tilde{\Lambda}\big) \geq 0\Big\} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\ \Big|\ A^{(G_1)}_{\text{bottom}}, \Pi\Big) \ \leq\ \exp\left(-I^{(q)}_{kr} + \eta^{(q)}\right) \tag{8.3}$$

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

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

where $\eta'(\delta_1;\Lambda^{(q)}) = 8\omega\delta_1 L\|\Lambda^{(q)}\|_\infty$ as defined in (7.8). The first inequality uses Lemma 10 and the second is obtained using the definition of $\eta'(\cdot)$ combined with $\Lambda^{(q)} \in \mathscr{B}_{\Lambda/Q}(\delta_1)$ (Lemma 8) which implies $\|\Lambda^{(q)}\|_\infty \leq (1+\delta_1)\|\Lambda/Q\|_\infty$. By taking expectation in (8.3), the same bound holds unconditionally.

By Lemma 13, 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 6, we conclude that

$$\text{Mis}_k\left(\tilde{y}^{(q-2)}_{\text{step 7}},\, y^{(q-2)}\right) \ \leq\ \sum_{r\neq k} \exp\left(-I_{kr}^{(q)} + \eta^{(q)} + u\right)$$

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

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

Since $9\omega\delta_1 < 1$ implies $\delta_1 < 9^{-1}$ (recall $\omega \geq 1$), 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 (8.2) 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 $\tilde{z}^{(q)}_{\text{step 8}}$ relative to $z^{(q)}$ which we call γ_2^{col}. In deriving γ_2^{col}, we have to make the substitutions in Remark 11, and particular, $\Lambda \to \Gamma$ where Γ is the column mean parameters defined in Chapter 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 $(I^{\text{col}}_{\ell r})$ from (3.2). Letting

$$\delta_1^{\text{col}} := 24\,C_{\gamma_1} + 6\tau_1^{\text{col}} + \xi K\beta, \tag{8.4}$$

we obtain the following counterpart of Lemma 14:

Corollary 8 (First LR classifier, column version)**.** *Under the assumptions of Lemma 13, 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* (7.8). *Then, on event* $\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}$,

$$\mathrm{Mis}_\ell\left(\tilde{z}^{(q)}_{step\,8},\, z^{(q)}\right) \;\le\; \sum_{r\neq\ell} \exp\left(-\frac{I^{col}_{\ell r}}{2Q} + \eta^{step\,8} + u^{col}\right) =: \gamma^{col}_{2\ell}, \quad \forall q \in \mathbb{Z}_Q, \tag{8.5}$$

fails with probability at most $Q(e^{-u^{col}} + 2Q\,p_3')$, *where* $p_3' = p_3'(\tau_1^{col})$ *as given in* (7.16).

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

$$\gamma_2 := \max\{\beta K \gamma_2^{\mathrm{row}}, \beta L \gamma_2^{\mathrm{col}}\}. \tag{8.6}$$

By (2.7), we have that (8.2) and (8.5) imply

$$\mathrm{Mis}\left(\tilde{y}^{(q)}_{\mathrm{step}\,7},\, y^{(q)}\right) \le \gamma_2^{\mathrm{col}} \le \frac{\gamma_2}{\beta K}, \quad \mathrm{Mis}\left(\tilde{z}^{(q)}_{\mathrm{step}\,8},\, z^{(q)}\right) \le \gamma_2^{\mathrm{row}} \le \frac{\gamma_2}{\beta L} \tag{8.7}$$

Thus, if we consider the following event:

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

after Step 8, we can work on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$ which holds with high probability: Combining Lemma 14 and Corollary 8, by union bound, $\mathbb{P}(\mathfrak{B}^c_{\gamma_2} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}) \le Q(2e^{-u} + 2Q(p_3 + p_3'))$, hence

$$\begin{aligned}
\mathbb{P}(\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}) \;&\ge\; \mathbb{P}\big(\mathfrak{B}_{\gamma_2} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\big) \\
&=\; \mathbb{P}\big(\mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\big) - \mathbb{P}\big(\mathfrak{B}^c_{\gamma_2} \cap \mathfrak{A}_{\gamma_1} \cap \mathfrak{P}\big) \\
&\ge\; 1 - \mathbb{P}(\mathfrak{A}^c_{\gamma_1}) - \mathbb{P}(\mathfrak{P}^c) - Q(2e^{-u} + Q(p_3 + p_3')).
\end{aligned} \tag{8.8}$$

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

$$\delta_2 := 24C_{\gamma_2} + 6\tau_2. \tag{8.9}$$

Lemma 15 (Second local parameters)**.** *Assume* (B4b) *and* $72\, C_{\gamma_2}\omega \le 1$*, and let* δ_2 *be as defined in* (8.9)*. 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* (7.15)*.*

Proof. Conditioning on blocks $G_1 \cup G_2$ (cf. Chapter 6) 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 Chapter 7.2 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 13. □

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 15, on $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$,

$$\|\hat{\Lambda}^{(q)}_{\text{step }9} - \Lambda/Q\|_\infty \le \delta_2 \|\Lambda/Q\|_\infty, \ \forall q \quad \text{hence,} \quad \|\hat{\Lambda}_{\text{step }10} - \Lambda\|_\infty \le \delta_2 \|\Lambda\|_\infty \tag{8.10}$$

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 12. 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 (8.10). 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 16 (Second LR classifier)**.** *Under the assumptions of Lemma 15, further assume that* δ_2 *defined in* (8.9) *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* $\mathfrak{B}_{\gamma_2} \cap \mathfrak{P}$*,*

$$\text{Mis}_k\left(\widehat{y}_{top},\ y_{top}\right) \ \le \ \sum_{r \ne k} \exp\left(-I_{kr} + \eta^{step\ 11}_{rk} + v\right) =: \gamma_3, \tag{8.11}$$

fails with probability at most $e^{-v} + 2Q\,p_3$ *where* $p_3 = p_3(\tau_2)$ *as given in* (7.15). *The same result holds for* $\eta_{kr}^{\text{step 11}} = \eta'(\delta_2; \Lambda)$.

Proof. As in the proof Lemma 15, 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 5(b) to A_{top}. Let $\text{Top} \subset [n]$ denote the row indices of A_{top}. Since all the columns are present in A_{top}, we can directly apply Lemma 5(b) (in contrast to the argument in Lemma 14), 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 4. Let $b_{i*}^{\text{step 11}} = b_{i*}(\tilde{z}_{\text{step 8}})$ be the block compression in step 11 of the algorithm. Then, Lemma 5(b) gives (after conditioning on $A_{\text{bottom}}^{(G_1 \cup G_2)}$ and then removing the conditioning as in (8.3))

$$\mathbb{P}\Big(\Big\{\exists\, \tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta_1),\ Y_{ikr}\big(b_{i*}^{\text{step 11}}, \tilde{\Lambda}\big) \geq 0\Big\} \cap \mathfrak{B}_{\gamma_2}\Big) \ \leq\ \exp\Big(-I_{kr} + \eta_{kr}^{\text{step 11}}\Big). \tag{8.12}$$

for any $i \in \text{Top}$ with $y_i = k$. By (8.10), 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 6, we conclude (8.11) as desired. The last statement of the theorem follows if we apply Lemma 5(a) in place of Lemma 5(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, $\text{Mis}_k(\widehat{y}, y) \leq \gamma_3$ 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

$$\begin{aligned}
\mathbb{P}\big(\,\text{Mis}_k(\widehat{y}, y) > \gamma_3\big) &\leq \mathbb{P}\big(\big\{\,\text{Mis}_k(\widehat{y}, y) > \gamma_3\big\} \cap \mathfrak{B}_{\gamma_2} \cap \mathfrak{P}\big) + \mathbb{P}\big((\mathfrak{B}_{\gamma_2} \cap \mathfrak{P})^c\big) \\
&\leq 2\big(e^{-v} + 2Q\,p_3(\tau_2)\big) + \mathbb{P}\big((\mathfrak{B}_{\gamma_2} \cap \mathfrak{P})^c\big) \\
&\leq 2\big(e^{-v} + 2Q\,p_3(\tau_2)\big) + \mathbb{P}(\mathfrak{A}_{\gamma_1}^c) + \mathbb{P}(\mathfrak{P}^c) + Q\big(2e^{-u \wedge u^{\text{col}}} + Q(p_3(\tau_1) + p_3'(\tau_1^{\text{col}})\big)
\end{aligned} \tag{8.13}$$

where γ_3 is the rate given in (8.11) and the second inequality uses (8.8).

8.1.2 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 (8.13) is $o(1)$.

Proof of Theorem 7. **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 14, we defined (recall (8.1))

$$\eta^{\text{step 7}} = 8\delta_1\omega L\|\Lambda/Q\|_\infty \;\le\; \left(384\beta^2\gamma_1 + 48\tau_1^{\text{row}} + 8\beta L\xi\right)\omega L\|\Lambda/Q\|_\infty. \tag{8.14}$$

By (6.6), $384\beta^2\gamma_1\omega L\|\Lambda/Q\|_\infty \le I_{\min}/(8Q)$. Take

$$\tau_1^{\text{row}} = \frac{I_{\min}}{384\omega L\|\Lambda\|_\infty}, \quad \xi = \frac{I_{\min} \wedge I_{\min}^{\text{col}}}{64\beta\omega(K \vee L)^2(\|\Lambda\|_\infty \vee \|\Gamma\|_\infty)}, \quad u = \frac{I_{\min}}{8Q}, \tag{8.15}$$

where u is the parameter in (8.2). Then from (8.14) 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 14 implies that on event $\mathfrak{P}$,

$$\text{Mis}_k\left(\tilde{y}^{(q)}_{\text{step 7}}, y^{(q)}\right) \;\le\; \gamma_{2k}^{\text{row}} := \sum_{r\neq k}\exp\left(-\frac{I_{kr}}{Q} + \frac{I_{\min}}{2Q}\right) \le K\exp\left(-\frac{I_{\min}}{2Q}\right), \quad \forall q \in \mathbb{Z}_Q \tag{8.16}$$

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),

$$\begin{aligned} Q^2 p_3(\tau_1^{\text{row}}) &= Q^2KL\exp\left(-\frac{n\Lambda_{\min}\,h_1(\tau_1^{\text{row}})}{32Q^2\beta K}\right) \\ &\le Q^2KL\exp\left(-\frac{nI_{\min}^2}{256(384^2)Q^2\beta KL^2\omega^3\|\Lambda\|_\infty}\right) = o(1) \end{aligned} \tag{8.17}$$

where we have used the definition (7.15) 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

$$\begin{aligned}\mathbb{P}(\mathfrak{P}^c) = p_2(\xi) &= 2Q(K+L)\exp\left(-(n\wedge m)\xi^2/Q\right)\\ &\le 2Q(K+L)\exp\Big(-\frac{n\wedge m}{64^2 Q\beta^2\omega^2(K\vee L)^2(\|\Lambda\|_\infty\vee\|\Gamma\|_\infty)^2}\Big) = o(1)\end{aligned} \tag{8.18}$$

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 (8.15). By (6.6), and a similar argument, we obtain $\eta^{\text{step 8}} + u^{\text{col}} \le I_{\min}^{\text{col}}/(4Q)$. By Corollary 8, on event $\mathfrak{P}$,

$$\text{Mis}_\ell\left(\tilde{z}^{(q)}_{\text{step 8}}, z^{(q)}\right) \le \gamma_{2\ell}^{\text{col}} \le \sum_{r\neq\ell}\exp\Big(-\frac{I_{\ell r}^{\text{col}}}{2Q}+\frac{I_{\min}^{\text{col}}}{4Q}\Big) \le L\exp\Big(-\frac{I_{\min}^{\text{col}}}{4Q}\Big), \quad \forall q\in\mathbb{Z}_Q, \tag{8.19}$$

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 (8.6) and combining with (8.16) and (8.19),

$$\gamma_2 \le \max\Big(\beta K^2\exp\Big(-\frac{I_{\min}}{2Q}\Big),\, \beta L^2\exp\Big(-\frac{I_{\min}^{\text{col}}}{4Q}\Big)\Big) = o\Big(\frac{\beta(K\vee L)^2}{(I_{\min}\wedge I_{\min}^{\text{col}})^b}\Big) \tag{8.20}$$

for any $b>0$, as $I_{\min}\wedge I_{\min}^{\text{col}}\to\infty$. By Lemma 16 and (7.8),

$$\begin{aligned}\eta_{kr}^{\text{step 11}} &:= \eta_{kr}\big(\delta_2;\ \omega,\beta,m,\Lambda\big)\\ &= 21\delta_2\,\omega L\|\Lambda\|_\infty + \frac{5\beta L^2\|\Lambda\|_\infty^2}{m} + \log\Big(11\omega\Big(\frac{1}{\varepsilon_{kr}-2\omega(1+\varepsilon_{kr})\,\delta_2}+1\Big)\Big) - \frac{1}{2}\log\Lambda_{\min}\\ &=: T_1+T_2+T_3+T_4,\end{aligned} \tag{8.21}$$

where we have called the four summands above $T_1, \ldots, T_4$ 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 (8.9) 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 (8.20)

$$T_{11} = O\big(\beta^2\omega L\|\Lambda\|_\infty\gamma_2\big) = o\Big(\frac{\beta^3\omega(K\vee L)^3\|\Lambda\|_\infty}{[(I_{\min}\wedge I^{\mathrm{col}}_{\min})/Q]^b}\Big). \tag{8.22}$$

Recall that we have $\beta\omega(K\vee L)\|\Lambda\|_\infty = o([(I_{\min}\wedge I^{\mathrm{col}}_{\min})/Q]^a)$ for some $a > 0$ by (A4′). Taking $b = 3a$ in (8.22), 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 16, we have by (A3)

$$\begin{aligned} Qp_3(\tau_2) &= QKL\exp\Big(-\frac{n\Lambda_{\min}\,h_1(\tau_2)}{32Q^2\beta K}\Big) \\ &\le QKL\exp\Big(-\frac{n}{256Q^2\beta KL^2\omega^3\|\Lambda\|_\infty}\Big) = o(1) \end{aligned} \tag{8.23}$$

where we have used $h_1(\tau) \ge \tau^2/8$ for $\tau \le 1$ and $\|\Lambda\|_\infty/\Lambda_{\min} \le \omega$. 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 (8.21). Recall that $J_{kr} = L\|\Lambda\|_\infty/I_{kr}$. By Lemma 23 in Chapter 10.3.2, $\varepsilon_{kr} \ge 2\big(J_{kr}^{-1}\wedge 1\big)$. 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 $J_{kr} \ge 1/2$ (see Remark 9), $(L\|\Lambda\|_\infty)^{-1} = o\big(J_{kr}^{-1}\wedge 2\big)$. Therefore, $2\omega\delta_2 = o(\varepsilon_{kr}\wedge 1)$. As a result, $2\omega(1+\varepsilon_{kr})\delta_2 = o(\varepsilon_{kr})$, hence

$$e^{T_3} := 11\omega\Big(1 + \frac{1}{\varepsilon_{kr} - 2\omega(1+\varepsilon_{kr})\delta_2}\Big) = O\Big(\omega\Big(1+\frac{1}{\varepsilon_{kr}}\Big)\Big). \tag{8.24}$$

Finally, we let $v = \sqrt{\log\Lambda_{\min}}$. Since $\Lambda_{\min}\to\infty$, $e^{-v} = o(1)$. Applying Lemma 16, combined with $T_1 + T_2 = O(1)$, and (8.24), 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. $\le 2(e^{-v} + 2Q\, p_3(\tau_2)) + \mathbb{P}(\mathfrak{P}^c) + Q(2e^{-u\wedge u^{\mathrm{col}}} + Q(p_3(\tau_1^{\mathrm{row}}) + p_3'(\tau_1^{\mathrm{col}}))) = o(1)$. When we swap A_{top} and A_{bottom} and repeat the algorithm, the same misclassification rate holds. The proof of Theorem 7 is complete. □

8.2 Proof of Theorem 1

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

Lemma 17. $\sum_{\ell\in[L]}(\lambda_{r\ell} - \lambda_{k\ell})^2 \ \ge\ 2\Lambda_{\min} I_{kr}$. *As a consequence,* $\Lambda_\wedge^2 \ge 2\Lambda_{\min} I_{min}$.

Combining Lemma 17 with Corollary 5, and noting that $\|\Lambda\|_\infty/\Lambda_\wedge^2 \le \omega/(2I_{\min})$ as a consequence of the lemma, we obtain the following guarantee for spectral clustering in terms of the information matrix (I_{kr}):

Corollary 9. *Consider the spectral algorithm given in Algorithm 5, assume that for a sufficiently small* $C_1 > 0$,

$$\frac{\beta^2\omega K L(K\wedge L)\,\alpha}{2I_{min}} \le C_1(1+\kappa)^{-2}. \tag{8.25}$$

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

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

We next modify Corollary 9 to be applicable on subblocks:

Lemma 18 (Spectral clustering on subblocks)**.** *Suppose* (A3) *holds, and we assume for a sufficiently small* $C_1 > 0$,

$$\frac{6Q\beta^2\omega^2 K L(K\wedge L)\,\alpha}{I_{min}} \le C_1(1+\kappa)^{-2}. \tag{8.26}$$

Using Algorithm 5 in Step 3 of Algorithm 3, w.h.p., the misclassification rate of $\tilde{y}^{(q)}$ *satisfies*

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

A similar result holds for misclassification rate of the spectral clustering for column labels,

with appropriate modifications.

Proof of Theorem 1. Assumption (A4) implies (8.26), 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 18 (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., $\text{Mis}(\tilde{y}^{(q)}, y^{(q)}) \le \gamma_1^{\text{row}}$ and $\text{Mis}(\tilde{z}^{(q)}, z^{(q)}) \le \gamma_1^{\text{col}}$ 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 (6.6) of Theorem 7, 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 7 hold with high probability, and so is its result. The proof is complete. □

8.3 Proof of Corollary 1

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

$$\text{Mis}_k\left(\widehat{y},\, y\right) = O\Big(\sum_{r \neq k} \omega\Big(1 + \frac{1}{\varepsilon_{kr}}\Big) \exp\Big(-I_{kr} - \frac{1}{2}\log\Lambda_{\min} + v\Big)\Big) \tag{8.27}$$

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

$$\chi_r := \omega\Big(1 + \frac{1}{\varepsilon_{kr}}\Big)\Lambda_{\min}^{-1/2} = o(1), \quad \text{uniformly in } r. \tag{8.28}$$

By Lemma 23 in Chapter 10.3.2, $\varepsilon_{kr} \geq 2\big(J_{kr}^{-1} \wedge 1\big)$. 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 9). Thus, to show (8.28), 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 (8.28), 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

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

completing the proof. □

8.4 Proof of Example 1

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 (8.27), 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. □

CHAPTER 9

Simulations

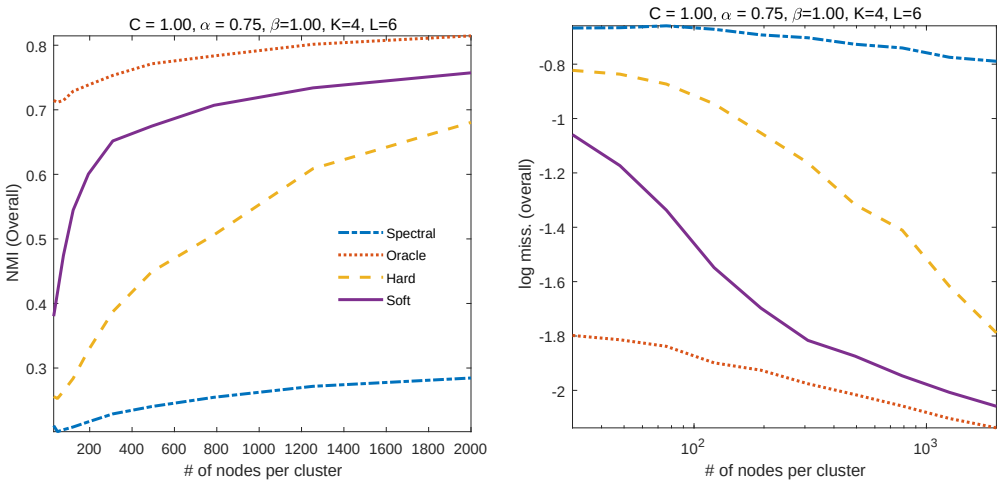
We provide some simulation results to corroborate the theory. We generate from the SBM model of Chapter 2.1 with the following connectivity matrix

$$P = C\frac{\big[\log(mn)\big]^{\alpha}}{\sqrt{mn}}B, \quad B = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 4 & 5 & 6 & 1 \\ 3 & 4 & 5 & 6 & 1 & 2 \\ 4 & 5 & 6 & 1 & 2 & 3 \end{bmatrix}. \tag{9.1}$$

Note that B does not have any clear assorative 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 1 (=perfect match). The NMI is quite sensitive to mismatch and tends to reveal discrepancies between methods more clearly. Figure 9.1(a) shows the overall NMI versus n_0. Figure 9.1(b) illustrates the corresponding log. misclassification rates.

We have considered four algorithms: (1) `Spectral`: the spectral clustering of Algorithm 5. (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 Chapter 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 9.1 shows the results for $\alpha = .75$ (regime where no exact recovery is possible) and $C = 1$. Both the soft and hard versions of Algorithm 1 are initialized with the spectral



(a) (b)

Figure 9.1: Plots of (a) the (overall) NMI and (b) the corresponding log. misclassification rate, for the SBM model with connectivity matrix (9.1). The four algorithms considered are the `Spectral` clustering of Algorithm 5, `Soft` and `Hard` versions of Algorithm 1 and the `Oracle` algorithm of Chapter 2.2.

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 also interesting to note that the slope for the log. misclassification rate of Algorithm 1 approaches that of the oracle (esp. clear for the soft version in Figure 9.1(b)) as predicted by the theory. Simulation results for various other settings can be found in Appendix B, showing qualitatively similar behavior.

CHAPTER 10

Proofs of the main lemmas

In this section, we give the proof of the three main lemmas of Chapter 7.1. We we first give the proofs of Lemma 4 and 6 in Chapter 10.1 and 10.2. The proof of Lemma 5(b) is more technical and occupies the remainder of this section, including auxiliary results on the error exponents and Poisson-binomial approximations, in Sections 10.3 and 10.4.

Throughout, we will use the following concentration inequality [GN15, p. 118]:

Proposition 3 (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*

$$\mathbb{P}\big(S \geq vt\big) \leq \exp[-vh_c(t)], \quad t \geq 0, \quad \textit{where } h_c(t) := \frac{3}{4c} t \log\big(1 + \frac{2c}{3}t\big). \tag{10.1}$$

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 $t \to \infty$.

10.1 Proof of Lemma 4

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

$$R_{k\ell}(\tilde{z}, z) = \frac{1}{m} \sum_{j=1}^{m} 1\{\tilde{z}_j = k, z_j = \ell\} = \frac{|j : \tilde{z}_j = k, z_j = \ell|}{m}. \tag{10.2}$$

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 (4.3) of the (global) row mean parameters,

$$\lambda_{k\ell'}(y, \tilde{z}) = \sum_{j=1}^{m} \sum_{\ell=1}^{L} P_{k\ell} 1\{z_j = \ell, \tilde{z}_j = \ell'\} = m P_{k*} R_{*\ell'}(z, \tilde{z}). \tag{10.3}$$

To see (10.3), note that since we are using true labels y in the first argument of $\lambda_{k\ell'}(y, \tilde{z})$, the averaging $\frac{1}{n_k(y)} \sum_i 1\{y_i = k\}(\cdots)$ 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 (10.3).

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 (2.2). 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

$$\lambda_{k\ell'}(y, \tilde{z}) = \lambda_{k*}(y, z)\, U_{*\ell'}(z, \tilde{z}), \tag{10.4}$$

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 (4.3), 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 (10.3). Using the definition of the confusion

matrix in (10.2), adapted to row labels, and the definition of U, we have

$$\lambda_{k'\ell'}(\tilde{y}, \tilde{z}) = \frac{1}{\pi_{k'}(\tilde{y})} R_{k'*}(\tilde{y}, y)\, \lambda_{*\ell'}(y, \tilde{z}) = U_{k'*}(\tilde{y}, y)\, \lambda_{*\ell'}(y, \tilde{z}), \tag{10.5}$$

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 19. *(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

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

Recall that we can write (see Chapter 5.1)

$$\mathrm{Mis}(\tilde{y}, y) = \frac{1}{n}|i : \tilde{y}_i \neq y_i| = \frac{1}{n}\sum_k |i : \tilde{y}_i = k,\, y_i \neq k| = \frac{1}{n}\sum_k |i : \tilde{y}_i \neq k,\, y_i = k|. \tag{10.6}$$

Then, (B3) and the second equality in (10.6) implies $|i : \tilde{y}_i = k, y_i \neq k|/n \le \gamma/(\beta K)$, while the third equality in (10.6) gives $|i : \tilde{y}_i = y_i = k|/n \ge \pi_k(y) - \gamma/(\beta K)$. Letting $f(x) = x/(x+1)$,

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

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.,

$\text{Mis}_\ell(\tilde{z}, z)$, 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 $\ell \in [L]$. □

10.1.1 Proof of Lemma 4(a)

For the lower bound, by (10.4) 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

$$\lambda_{k\ell'}(y, \tilde{z}) - \lambda_{k\ell'}(y, z) \leq \sum_{\ell \neq \ell'} \lambda_{k\ell}(y, z) U_{\ell\ell'}(z, \tilde{z}). \tag{10.7}$$

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

$$\pi_{\ell'}(z) \;\geq\; \frac{1}{m}|j : z_j = \tilde{z}_j = \ell'| \;=\; \pi_{\ell'}(\tilde{z}) V_{\ell'\ell'}(z, \tilde{z}) \;\geq\; (1 - \gamma)\pi_{\ell'}(\tilde{z}). \tag{10.8}$$

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 (10.7)

$$\lambda_{k\ell'}(y,\tilde{z}) - \lambda_{k\ell'}(y,z) \;\le\; \frac{\beta^2}{1-\gamma}\sum_{\ell\neq\ell'}\lambda_{k\ell}(y,z)V_{\ell\ell'}(z,\tilde{z}) \;\le\; \frac{\beta^2\gamma}{1-\gamma}\|\lambda_{k*}(y,z)\|_\infty \tag{10.9}$$

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

10.1.2 Proof of Lemma 4(b)

Recalling (10.5), 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 $\gamma \le C_\gamma$.)

10.1.3 Proof of Lemma 4(c)

Recalling definitions of $\hat{\lambda}_{k\ell}$ and $\lambda_{k\ell}(\tilde{y},\tilde{z})$ from (4.2) and (4.3), 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 $\sum_{ij}\mathbb{E}X_{ij}^2 = \sum_{ij}\mathrm{var}(A_{ij}) \le \sum_{ij}\mathbb{E}A_{ij} = n_k(\tilde{y})\lambda_{k\ell}(\tilde{y},\tilde{z})$. 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 3, 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 (10.8). Similarly, since $\|\Lambda(\tilde{y},\tilde{z}) - \Lambda\|_\infty \le 3C_\gamma\|\Lambda\|_\infty$, we have

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

10.2 Proof of Lemma 6

Let $b_{i*} = b_{i*}(\tilde{z})$. Recall (7.3), (7.4) and (7.5), 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 $\mathbb{E}[X;\mathcal{A}] := \mathbb{E}[X1_{\mathcal{A}}]$. Consider the following event: $\mathcal{A} := \{\hat{\Lambda} \in \mathscr{B}_\Lambda(\delta)\}$. Pick some $i \in [N]$ with $y_i = k$. Then,

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

where the last inequality follows from Lemma 5 with η_{kr} defined there. Using Markov inequality

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

for any $t > 0$. The version of Markov inequality used follows from (pointwise) inequality: $1_{\{X\geq u\}}1_{\mathcal{A}} \leq (X1_{\mathcal{A}})/u$. Taking $t = e^u$ complete the proof.

10.3 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 [AS15], and is proved by the same technique. The result (and the technique inspired by [AS15]) is interesting since it provides a bound different than the classical Chernoff bound on the error exponent [Che52]; see also [Ver86] and [CT06, Theorem 11.9.1]. 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 Chapter 10.3.2 specialize to the case of interest in this paper, the Poisson family.

10.3.1 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 $\mathbb{R}$:

$$\pi(t;\gamma) = h(t)\exp\big(\gamma t - A(\gamma)\big). \tag{10.10}$$

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

$$p(x;\theta) = \prod_{\ell=1}^{L} \pi(x_\ell,\theta_\ell), \quad x=(x_\ell)\in\mathbb{R}^L, \quad \theta=(\theta_\ell)\in\mathbb{R}^L \tag{10.11}$$

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

Proposition 4. *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 (10.10) and (10.11). 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*

$$\theta_{s\ell} = (1-s)\theta_{0\ell} + s\theta_{1\ell}, \quad \text{and} \quad I_{s\ell} = \big[(1-s)A(\theta_{0\ell}) + sA(\theta_{1\ell})\big] - A(\theta_{s\ell}), \tag{10.12}$$

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

$$C(\alpha) := \int e^{-\alpha|t|}d\nu(t) = \begin{cases} \frac{2}{\alpha} & \nu \text{ is Lebesgue,} \\ \frac{1+e^{-\alpha}}{1-e^{-\alpha}} \le \frac{2}{1-e^{-\alpha}} & \nu \text{ is counting.} \end{cases} \tag{10.13}$$

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

$$P_{e,0}+P_{e,1} \;\le\; \inf_{\ell\in T}\inf_{s\in(0,1)} \Big[e^{-I_s}\|\pi(\,\cdot\,;\theta_{s\ell})\|_\infty C\Big(\min(s,1-s)|\theta_{0\ell}-\theta_{1\ell}|\Big)\Big]. \tag{10.14}$$

Remark 13. 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 (10.10) 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 14. Let s^* be the maximizer of $s \mapsto I_s$. Then, noting that $\alpha \mapsto C(\alpha)$ is decreasing, Proposition 4 implies

$$P_{e,0} + P_{e,1} \leq \exp\Big(-I_{s^*} + \log \|\pi(\,\cdot\,;\theta_{s^*\ell})\|_\infty + \log C(\alpha^*)\Big), \quad \text{where,} \tag{10.15}$$

$$\alpha^* = \min(s^*, 1-s^*) \max_{\ell\in[L]} |\theta_{0\ell} - \theta_{1\ell}|. \tag{10.16}$$

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.

10.3.2 Poisson case

The Poisson case corresponds to (10.10) 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 (10.12)

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

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

$$\lambda_{0\ell}/\lambda_{1\ell} \in [1/\omega, \omega],\ \forall \ell \in [L], \quad \text{for some } \omega > 1. \tag{10.17}$$

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

Lemma 20. *Assuming (10.17), we have* $s^* \in [\frac{1}{2\omega}, 1 - \frac{1}{2\omega}]$.

Proof of this lemma and subsequent results appear in Chapter A.4. From (10.16), we have $\alpha^* = \min(s^*, 1-s^*)\max_\ell |\log(\lambda_{0\ell}/\lambda_{1\ell})|$ in the Poisson case. Defining

$$\varepsilon_{01} := \varepsilon_{01}(\Lambda) := \max_{\ell\in[L]} \Big(\frac{\lambda_{0\ell}}{\lambda_{1\ell}} \vee \frac{\lambda_{1\ell}}{\lambda_{0\ell}}\Big) - 1, \quad \alpha_{01} := \frac{1}{2\omega}\log(1+\varepsilon_{01}) \tag{10.18}$$

we note that $\alpha^* = \min(s^*, 1-s^*)\log(1+\varepsilon_{01})$, hence Lemma 20 implies $\alpha^* \geq \alpha_{01}$, that is, $C(\alpha^*) \leq C(\alpha_{01})$ in (10.15), where $C(\cdot)$ has the form given in (10.13) for the counting measure, i.e.,

$$C(\alpha_{01}) = \frac{1+e^{-\alpha_{01}}}{1-e^{-\alpha_{01}}} \leq \frac{2}{1-e^{-\alpha_{01}}} \overset{(a)}{\leq} \Big(\frac{4}{\varepsilon_{01}} + 3\Big)\omega \tag{10.19}$$

where the last inequality is by the following lemma:

Lemma 21. *Inequality (a) in (10.19) holds.*

Next we bound the maximum of the density:

Lemma 22 ([HLC60])**.** *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 20, 21 and 22, we have the following corollary which gives the following overall bound on the error exponent:

Corollary 10. *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 (10.17). Let $\Lambda_{\min} = \min_{r\ell}\lambda_{r\ell}$. Then, the sum of the error probabilities for the likelihood ratio test is bounded as*

$$P_{e,0} + P_{e,1} \;\leq\; \omega\Big(\frac{4}{\varepsilon_{01}} + 3\Big)\exp\Big(-I^* - \frac{1}{2}\log\Lambda_{\min}\Big). \tag{10.20}$$

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

Lemma 23. *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 23 holds without any further assumption, it is not always tight. The difference in our two sets of results, namely (3.6) and (3.7) is due to using the sharper bound (10.20) versus replacing ε_{01} with its universal lower bound.

Lemma 24 (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 (10.18), and assume (10.17). Then $\varepsilon'_{01} \geq \varepsilon_{01} - 2\omega(1+\varepsilon_{01})\delta$.*

10.4 Approximation results for Lemma 5(b)

Let us collect some approximation lemmas that will be used in the proof of Lemma 5(b). The proofs can be found in Appendix A.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 25 (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,*

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

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

Corollary 11 (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*

$$\widetilde{\Phi}(x, (p^{(1)}, \dots, p^{(L)})) := \prod_{\ell=1}^L \widetilde{\varphi}(x_\ell; p^{(\ell)}), \quad \text{for each } x = (x_1, \dots, x_L) \in \mathbb{Z}_+^L \tag{10.21}$$

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

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

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

Lemma 26 (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 27 (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,*

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

Proof of Lemma 27. 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 3, we have

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

□

10.5 Proof of Lemma 5(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 $r \ne k \in [K]$, i such that $y_i = k$ and $\tilde{\Lambda} \in \mathscr{B}_\Lambda(\delta)$,

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

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

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

where we have defined (recalling the definition of Ψ from (4.13)):

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

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

$$E = \Big\{x \in \mathbb{Z}_+^L : \sum_{\ell=1}^L x_\ell \leq 5L\|\Lambda\|_\infty\Big\}. \tag{10.22}$$

Using Lemma 27, 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,

$$\|P\|_\infty \leq \frac{\|\Lambda\|_\infty}{\min_i n_i(z)} \leq \frac{\beta L\|\Lambda\|_\infty}{m} \tag{10.23}$$

where the first inequality follows from definition of Λ in (2.2), 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 (4.7), 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 (10.21). We also note that $\mathbb{E}[b_{i*}(\tilde{z})] = \lambda_{k*}(y, \tilde{z}) =: \lambda'_{k*}$. It follows from Corollary 11, noting that $\|P_{k,\mathrm{lab}(\tilde{z})}\|_\infty \le \|P\|_\infty$ combined with (10.23),

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

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 26,

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

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

$$\mathbb{P}(b_{i*} \in E \cap F) \;=\; \widetilde{\Phi}\big(E \cap F; P_{k*}\big) \;\le\; \zeta_1\, \Phi\big(E \cap F; \lambda'_{k*}\big) \;\le\; \zeta_1\zeta_2\, \Phi\big(E \cap F; \lambda_{k*} - \rho\big). \quad (10.24)$$

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

$$\begin{aligned}\sum_{x\in F} \Phi(x; \lambda_{k*} - \rho) &= \sum_{x \in \mathbb{Z}_+^L} p_0(x)\, 1\Big\{\log\frac{p_1(x)}{p_0(x)} \ge -2L\rho\Big\} \\ &= \sum_{x \in \mathbb{Z}_+^L} p_0(x)\, 1\Big\{\frac{e^{2L\rho} p_1(x)}{p_0(x)} \ge 1\Big\} \\ &\le \sum_{x \in \mathbb{Z}_+^L} \min\Big(e^{2L\rho} p_1(x),\; p_0(x)\Big) \le e^{2L\rho} \sum_{x \in \mathbb{Z}_+^L} \min\big(p_1(x),\; p_0(x)\big). \quad (10.25)\end{aligned}$$

Let us define

$$I_s(\lambda_0 \mid \lambda_1) = \sum_{\ell=1}^{L} \big[s\lambda_{0\ell} + (1-s)\lambda_{1\ell}\big] - \lambda_{0\ell}^s \lambda_{1\ell}^{1-s}, \quad \lambda_0, \lambda_1 \in \mathbb{R}_+^L. \tag{10.26}$$

We can now apply Corollary 10. 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 10 by 2ω. It follows that

$$\begin{aligned}
\sum_{x \in \mathbb{Z}_+^L} \min\big(p_1(x),\ p_0(x)\big) \ &\le\ \zeta_3 \exp\Big(-I_s(\lambda_{r*} + \rho \mid \lambda_{k*} - \rho) - \frac{1}{2}\log(\Lambda_{\min} - \rho)\Big) \\
&\le\ \zeta_3 \exp\Big(-I_s(\lambda_{k*} \mid \lambda_{r*}) + 2\omega L\rho - \frac{1}{2}\Big(\log\Lambda_{\min} + \log\frac{2}{3}\Big)\Big) \\
&\le\ 8\sqrt{\frac{3}{2}}\zeta_3 \exp\Big(-I_s(\lambda_{k*} \mid \lambda_{r*}) + 2\omega L\rho - \frac{1}{2}\log\Lambda_{\min}\Big) \qquad (10.27)
\end{aligned}$$

where $\zeta_3 = \omega/(\varepsilon_{kr} - 2\omega(1+\varepsilon_{kr})\delta) + \omega$ from Lemma 24, 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 (10.24), (10.25) and (10.27) 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

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

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 A

Remaining proofs

A.1 Proofs of Sections 7.2, 7.3 and 7.4

Proof of Lemma 7. We have $n_k(y^{(q)}) \sim \text{Hypergeometric}(n/4, n_k(y), n)$. For any fixed $k \in [K]$ and $q' \in [Q]$, the concentration of hypergeometric distribution [Chv79] gives $|\pi_k(y^{(q')}) - \pi_k(y)| \leq \xi$ with probability at least $1 - 2\exp(-n\xi^2/Q)$. The same probability bound holds for $|\pi_\ell(z^{(q)}) - \pi_\ell(z)| \leq \xi$, 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 8. Recall the definition of the true local mean parameters in (4.6), and the corresponding global parameters in Chapter 4.1. We have

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

Since $|\pi_\ell(z^{(q)}) - \pi_\ell(z)| \leq \xi$ and $\pi_\ell(z) \geq 1/(\beta L)$ by assumptions (B4a) and (A2), the first inequality in (7.14) follows, from which we have the second inequality by (B4a). □

Proof of Lemma 9. From Lemma 8, we have $\|\Lambda^{(q)} - \Lambda/Q\|_\infty \leq (\xi L\beta)\|\Lambda/Q\|_\infty$ and $\|\Lambda^{(q)}\|_\infty \leq \frac{3}{2}\|\Lambda/Q\|_\infty$. 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 4 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 8, we have

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

By (7.12), the condition (A2) holds with β replaced with 2β. We also need to replace $\Lambda_{\min}$ in Lemma 8 with $\Lambda_{\min}^{(q)} \geq \Lambda_{\min}/(2Q)$, and C_γ with $4C_\gamma$ (more precisely, we are replacing $C_{\gamma,\beta}$ with $C_{\gamma,2\beta}$). Thus, assuming $6(4C_\gamma)(3\omega) \leq 1$, we obtain

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

where $p_1(\cdot)$ is as in (7.2). Since $\|\Lambda^{(q)}\|_\infty \leq \frac{3}{2}\|\Lambda/Q\|_\infty$, $4(4C_\gamma + \tau)\|\Lambda^{(q)}\|_\infty \leq (24C_\gamma + 6\tau)\|\Lambda/Q\|_\infty$. Thus, on the event in (A.1), we have by triangle inequality

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

which is the desired result. □

Proof of Lemma 10. 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

$$I_s(\lambda_0 \mid \lambda_1) = \sum_{\ell=1}^{L}\big[(1-s)\lambda_{0\ell} + s\lambda_{1\ell}\big] - \lambda_{0\ell}^{1-s}\lambda_{1\ell}^{s}, \quad \lambda_0, \lambda_1 \in \mathbb{R}_+^L \tag{A.2}$$

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

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

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

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

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

Proof of Lemma 11. Assume $\mathrm{dMis}(\tilde{y}, y) \leq \alpha$ and let $n_k = |i : y_i = k|$ and $N_{kk'} = |i :\ y_i =$

$k,\ \tilde{y}_i = k'|$. Then,

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

It follows that $\sum_{k' \neq k} N_{kk'} \leq \varepsilon n_k$ for every k. We also obtain $N_{kk'} \leq \varepsilon n_k$ for all k and k' such that $k \neq k'$. Since $\sum_{k'} N_{kk'} = n_k$, we have $N_{kk} \geq (1-\varepsilon) n_k$. 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 12. 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 11 gives the desired result. □

Proof of Corollary 6. Take $q = 2$ for simplicity. Assume that (7.17) 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 (7.12) for row labels. Applying Lemma (11) we conclude that $\sigma_{12} = \sigma^*(\tilde{y}^{(1)} \to y^{(1)}) =: \sigma_1$. □

Proof Corollary 7. 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 6, $\sigma_{12} = \sigma_2$ and $\sigma_{23} = \sigma_2'$. Then, the argument leading to (6.4) 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 (7.12) for row labels, and the third inequality follows from the second inequality and Lemma 11(b). It thus follows from Lemma 12 that $\sigma_2^{-1} \circ \sigma_2' = \sigma^*(\tilde{y}'^{(2)} \to \tilde{y}^{(2)})$ which is the desired result. □

A.2 Proofs of Chapter 8.1.1

Proof of Lemma 17. 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 18. Recall our choice of ξ in (8.15)—which will also be assumed in this proof—giving $\mathbb{P}(\mathfrak{P}^c) = o(1)$ as shown (8.18). By Lemma 8, we have $\Lambda^{(q)} \in \mathscr{B}_{\Lambda/Q}(\xi L\beta)$ for all $q \in [Q]$, which combined with Lemma 10 (applied with $\delta = \xi L\beta$) gives

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

using (8.15). Thus

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

as $I_{\min} \to \infty$. We are now ready to apply Corollary 9 to the Algorithm 5 operating on subblocks in G_1^{col}. It remains to verify that assumption (8.26) translates to condition (8.25) for the subblocks. Indeed, we have to replace $I_{\min}$ with $I_{\min}(\Lambda^{(q)})$, ω with 3ω (by Lemma 8), β with 2β (by (7.12)), 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 (8.26),

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

verifying condition (8.25) on the subblocks. Applying Corollary 9, we have the misclassification rate of $\tilde{y}^{(q)}$ satisfies

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

which is the desired result. □

A.3 Proofs of Chapter 10.3.1

Proof of Proposition 4. **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,

$$P_{e,+} = \int p_0 1\{p_0 \le p_1\} d\mu + \int p_1 1\{p_1 < p_0\} d\mu \tag{A.4}$$

$$= \int \min(p_0, p_1) d\mu = \int p_0^{1-s} p_1^s \, \min(l^s, l^{s-1}) d\mu. \tag{A.5}$$

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

$$p_s := \frac{p_0^{1-s} p_1^s}{\int p_0^{1-s} p_1^s d\mu}, \quad \text{and} \quad p_{s\ell} := \frac{p_{0\ell}^{1-s} p_{1\ell}^s}{\int p_{0\ell}^{1-s} p_{1\ell}^s d\nu} \tag{A.6}$$

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 (10.11). That is, p_s defined in (A.6) 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 (A.4), we obtain

$$P_{e,+} = e^{-I_s} \int p_s \, \min(l^s, l^{s-1}) d\mu. \tag{A.7}$$

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

$$l(x) = \frac{p_0(x)}{p_1(x)} = \prod_\ell l_\ell(x_\ell), \quad \text{where} \quad l_\ell(x_\ell) = \frac{p_{0\ell}(x_\ell)}{p_{1\ell}(x_\ell)} = \exp\left[(\theta_{0\ell} - \theta_{1\ell})x - A(\theta_{0\ell}) + A(\theta_{1\ell})\right]. \tag{A.8}$$

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 $(x_2, \ldots, x_L)$, let

$x_1^* = x_1^*(x_2, \ldots, x_L)$ 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 (A.7):

$$e^{I_s} P_{e,+} = \int \prod_{\ell=2}^{L} p_{s\ell} \Big[\int p_{s1} \min(l^s, l^{s-1}) d\mu^1 \Big] d\mu^{2:L} \tag{A.9}$$

Let $J = J(x_2, \ldots, x_L)$ denote the inner integral in (A.9) (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 $(x_2, \ldots, x_L)$ 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 $J(x_2, \ldots, x_L) \leq C(\alpha)\|p_{s1}\|_\infty$ for all $(x_2, \ldots, x_L)$. It follows that the same bound holds for $P_{e,+}$ by (A.9), 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.

□

A.4 Proofs of Chapter 10.3.2

Proof of Lemma 20. 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

$$\lambda_{k1} - \lambda_{r1} + \lambda_{k1}^{1-s^*}\lambda_{r1}^{s^*}\log\left(\frac{\lambda_{r1}}{\lambda_{k1}}\right) = 0 \tag{A.10}$$

Let $x = \frac{\lambda_{r1}}{\lambda_{k1}}$. Now (A.10) 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 21. 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

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

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

Proof of Lemma 22. We have

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

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

Proof of Lemma 23. There exists $\ell \in L$ such that

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

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

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

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

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

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

Proof of Lemma 24. Without loss of generality, we assume $\lambda_{01}/\lambda_{11} = 1 + \varepsilon_{01}$. Letting $\rho = \delta\|\Lambda\|_\infty$, we have $\|\Lambda' - \Lambda\|_\infty \le \rho$ 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

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

Combined with

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

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

Proof of Lemma 25. Let $X_j \sim \mathrm{Poi}(p_j)$ independent over $j = 1,\dots,n$, so that $\sum_{j=1}^n X_j \sim \mathrm{Poi}(\lambda)$. 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 26. We have

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

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

Proof of Lemma 27. 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 3, we have

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

□

A.5 Proof of Lemma 5(a)

Proof of Lemma 5(a). Fix $\tilde z$ and let $b_{i*} = b_{i*}(\tilde z)$. For $r \ne k \in [K]$ 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 $\mathrm{Poi}(\mu)$ random variable. Recalling the assumption $\Lambda(y,\tilde{z}) \in \mathscr{B}_\Lambda$, we have

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

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 $\lambda_{k\ell} - \rho = \lambda_{k\ell} - \delta\|\Lambda\|_\infty \ge \lambda_{k\ell} - \frac{\|\Lambda\|_\infty}{3\omega} \ge \lambda_{k\ell} - \frac{1}{3}\Lambda_{\min} = \frac{2}{3}\lambda_{k\ell}$,

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

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

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

Taking infimum over $s > 0$, by Lemma 20, the maximizer s^* of I_{kr} is always bounded between 0 and 1, hence we have

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

□

A.6 Proofs of Chapter 3

Proof of Proposition 1. The upper bound has been provided by Corollary 10. Here we will show the lower bound, using the notation established in the proof of Proposition 4 and Chapter 10.3.2. 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 (A.7) that

$$P_{e,+} = \int \min(p_0, p_1) d\mu = e^{-I_s} \int p_s \, \min(l^s, l^{s-1}) d\mu, \tag{A.12}$$

where p_s is defined in (A.6) and l in (A.8). Since μ is the counting measure, we have

$$P_{e,+} \geq \max_{x \in \mathbb{Z}_+^L} \min(p_0(x), p_1(x)) = \max_{x \in \mathbb{Z}_+^L} e^{-I_s} p_s(x) \, \min(l^s(x), l^{s-1}(x)). \tag{A.13}$$

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. Chapter 10.3.2). 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 [Jam15]:

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

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 $\Gamma(x+1) = x\Gamma(x)$,

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

that is,

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

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

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

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

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

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,

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

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,

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

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

A.7 Alternative algorithm for the k-means step

In this appendix, we present a simple general algorithm that can be used in the k-means step, replacing the κ-approximate k-means solver used throughout the text. The algorithm is based on the ideas in [**Gao2015a**] and [**yun2014community**], and the version that we present here acheives the misclassification bound $\varepsilon^2/(n\delta^2)$ needed in Step 3 of the analysis (Section 5.3.1) without necessarily optimizing the k-means objective function. We present the results using the terminology of the k-means matrices (with rows in $\mathbb{R}^d$) introduced in Section 5.3.4, although the algorithm and the resulting bound work for data points in any metric space.

Let $X \in \mathbb{M}^k_{n,\,d}$ be a k-means matrix and let us denote its centers, i.e. distinct rows, as $\{q_r(X), r \in [k]\}$. As in Definition 2, we write $\delta_r(X)$ and $n_r(X)$ for the rth cluster center separation and size, respectively, and $\delta_\wedge(X) = \min_r \delta_r(X)$ and $n_\wedge = \min_r n_r(X)$. Assume that we have an estimate $\hat{X} \in \mathbb{R}^{n\times d}$ of X, and let us write $d(i,j) := d(\hat{x}_i, \hat{x}_j)$, $i,j \in [n]$ for the pairwise distances between the rows of $\hat{X}$.

Algorithm 4 which is a variant of the one presented in [**Gao2015a**], takes these pairwise distances and outputs cluster estimates $\widehat{\mathcal{C}}_1, \ldots, \widehat{\mathcal{C}}_k \subset [n]$, after k recursive passes through the data. A somewhat more sophisticated version of this algorithm appears in [**yun2014community**], where one also repeats the process for $i = 1, \ldots, \log n$ and radii $R_i = iR_1$ in an outer loop, producing clusters $\widehat{\mathcal{C}}^{(i)}_r, r \in [k]$; one then picks, among these $\log n$ possible clusterings, the one

Algorithm 4 k-means replacement

Require: Pairwise distance $d(i,j)$, $i,j \in [n]$ and radius ρ.

$S \leftarrow [n]$
for $r = 1, \ldots, k$ **do**
    For every $i \in S$, let $B_d(i;\rho) := \{j \in S : \ d(i,j) \leq \rho\}$.
    Pick $i_0 \in S$ that maximizes $i \mapsto |B_d(i;\rho)|$.
    Let $\widehat{\mathcal{C}}_r = B_d(i_0;\rho)$.
    $S \leftarrow S \setminus \widehat{\mathcal{C}}_r$.
end for

Ensure: Return clusters $\widehat{\mathcal{C}}_r : 1, \ldots, k$ and output remaining S as unlabeled.

that minimizes the k-means objective. The variant in [**yun2014community**] also leaves no unlabeled nodes by assigning the unlabeled to the cluster whose estimated center is closest. In the rest of this section, we will focus on the simple version presented in Algorithm 4 as this is enough to establish our desired bound. The following theorem provides the necessary guarantee:

Theorem 8. *Consider the cluster model above and let* $n_r = n_r(X)$, $n_\wedge = n_\wedge(X)$ *and* $\delta_\wedge = \delta_\wedge(X)$. *Assume that we have approximate data* $\hat{x}_1, \ldots, \hat{x}_n$ *such that* $\sum_{i=1}^n d(x_i, \hat{x}_i)^2 \leq \varepsilon^2$. *In addition, assume that for some* $\gamma \in (0,1)$ *and* $\beta \geq 1$*:*

(i) $n_r \leq \beta n_\wedge$ *for all* $r \in [k]$ *(Clusters are* β*-balanced.),*

(ii) $\dfrac{2\varepsilon}{\sqrt{\gamma n_\wedge}} < \dfrac{\delta_\wedge}{3}$ *(*ε^2 *small enough compared to* $n_\wedge \delta_\wedge^2$*.),*

(iii) $\xi\beta + \gamma < 1 - \gamma$ *where* $\xi := \gamma/(1-\gamma)$. *(Gamma small enough relative to* β*.)*

Let $M_n(\rho)$ *be the (average) misclassification rate of Algorithm 4 with input radius* ρ. *Then,*

$$M_n(\rho) \leq \frac{8\varepsilon^2}{n\rho^2}, \quad \forall \rho \in \Big[\frac{2\varepsilon}{\sqrt{\gamma n_\wedge}}, \frac{\delta_\wedge}{3}\Big).$$

Applying the algorithm with $\rho \geq \alpha\delta_\wedge$ for $\alpha < 1/3$ we obtain the misclassication bound $c_\alpha\, \varepsilon^2/(n\delta_\wedge^2)$ where $c_\alpha = 8/\alpha^2$. Thus, Algorithm 4 with a proper choice of the radius ρ satisfies the desired bound (5.11) of the k-means step.

Proof of Theorem 8. The proof follows the argument in [**Gao2015a**]. As in the proof of Proposition 2, let $\mathcal{C}_r$ denote the rth cluster of X, having center $q_r = q_r(X)$. We have

$|\mathcal{C}_r| = n_r$. Let x_i and $\hat{x}_i$ be the ith row of X and $\hat{X}$, respectively, and let

$$T_r := \{i \in \mathcal{C}_r : \ d(\hat{x}_i, q_r) < \rho/2\} = \{i \in \mathcal{C}_r : \ d(\hat{x}_i, x_i) < \rho/2\}$$

using $x_i = q_r$ for all $i \in \mathcal{C}_r$ which holds by definition. $\{T_r\}$ are disjoint and clearly $T_r \subset \mathcal{C}_r$. Let $T := \biguplus_r T_r$, a disjoint union, and $T^c = [n] \setminus T$. We have

$$|T^c|\rho^2/4 \ \leq \ \sum_{i \in T^c} d(\hat{x}_i, x_i)^2 \leq \varepsilon^2 \implies |T^c| \leq 4\varepsilon^2/\rho^2. \tag{A.14}$$

As a consequence of assumption (ii) and our choice of ρ, we have $4\varepsilon^2/(n_\wedge \rho^2) \leq \gamma$, hence

$$|T_r| = |\mathcal{C}_r| - |\mathcal{C}_r \setminus T_r| \ \geq \ |\mathcal{C}_r| - |T^c| \ \geq \ n_\wedge\Big(1 - \frac{4\varepsilon^2}{n_\wedge \rho^2}\Big) \geq n_\wedge(1-\gamma) \tag{A.15}$$

for all $r \in [k]$. On the other hand, $|T^c| \leq \gamma n_\wedge$. In particular, combining the two estimates

$$|T^c| \leq \xi\, |T_r|, \quad \forall r \in [k] \tag{A.16}$$

where $\xi = \gamma/(1-\gamma)$. These size estimates will be used frequently in the course of the proof.

Recall that $d(i,j) := d(\hat{x}_i, \hat{x}_j)$, $i, j \in [n]$, the collection of pairwise distances between the data points $\hat{x}_1, \ldots, \hat{x}_n$. Thus, with some abuse of notation, $(i,j) \mapsto d(i,j)$ defines a pseudo-metric on $[n]$ (and a proper metric if $\{\hat{x}_i\}$ are distinct). For any two subsets $A, B \subset [n]$ we write $d(A,B) = \inf\{d(i,j) : i \in A,\ j \in B\}$. For any $i \in [n]$, let $d(i,A) = d(\{i\}, A)$.

We say that node i_0 is near T if $d(i_0, T) \leq \rho$, i.e., i_0 belongs to the ρ-enlargement of T. Similarly, we say that i_0 is near T_r if $d(i_0, T_r) \leq \rho$ and far from T_r otherwise. Note that i_0 can be near at most one of $T_r, r \in [k]$. This is since $d(T_r, T_\ell) \geq \delta_\wedge - \rho$ for $r \neq \ell$, and we are assuming $\delta_\wedge > 3\rho$. In fact, i_0 *is near* T *iff* i_0 *is near exactly one of* $T_r, r \in [k]$.

To understand Algorithm 4, let us assume that we are at some iteration of the algorithm and we are picking the center i_0 and the corresponding cluster $\widehat{\mathcal{C}} := \{j : \ d(j, i_0) \leq \rho\}$. One of the following happens:

(a) We pick the new center $i_0 \in T_r$ for some r, in which case $\widehat{\mathcal{C}}$ will include the entire T_r, none of $T_\ell, \ell \neq r$, and perhaps some of T^c. That is, $\widehat{\mathcal{C}} \supset T_r$ and $\widehat{\mathcal{C}} \cap T_\ell = \emptyset$ for $\ell \neq r$.

(b) We pick i_0 *near* T_r for some r. In this case, $|\widehat{\mathcal{C}}| \geq |T_r|$, otherwise any member of T_r would have created a bigger cluster by part (a) above. Now, $\widehat{\mathcal{C}}$ cannot contain any of $T_\ell, \ell \neq r$, because i_0 is far from those if it is near T_r. Hence, $\widehat{\mathcal{C}} \subset T_r \cup T^c$. Since $|T^c| \leq \xi|T_r|$ by (A.16), and $|\widehat{\mathcal{C}}| \geq |T_r|$, we have $|\widehat{\mathcal{C}} \cap T_r| \geq (1-\xi)|T_r|$. That is, $\widehat{\mathcal{C}}$ contains a large fraction of T_r.

If either of the two cases above happen, we say that T_r is depleted, otherwise it is intact. If T_r is depleted, it will not be revisited in future iterations, as long as other intact $T_\ell, \ell \neq r$ exist. To see this, first note that $|T_r \cap \widehat{\mathcal{C}}^c| \leq \xi|T_r| \leq \xi\beta n_\wedge$, using assumption (i). Taking i_0 on or near T_r in a future iteration will give us a cluster of size at most $(\xi\beta + \gamma)n_\wedge < (1-\gamma)n_\wedge$ (by assumption (iii)) which is less that $|T_\ell|$ for an intact cluster.

To simplify notation, if either of (a) and (b) happen, i.e., we pick cluster center i_0 near T_r for some r, we name the corresponding cluster $\widehat{\mathcal{C}}_r$. This is to avoid carrying around a permutation of cluster labels different than the original one, and is valid since each T_r is visited at most once by the above argument. (In fact, each is visited exactly once, as we argue below.) That last possibility is

(c) We pick i_0 *far* from any T_r, that is $d(i_0, T) > \rho$. This gives $\widehat{\mathcal{C}} \subset T^c$, hence $|\widehat{\mathcal{C}}| \leq |T^c| \leq \gamma n_\wedge < (1-\gamma)n_\wedge \leq |T_\ell|$ for any intact T_ℓ. Thus as along as there are intact T_ℓ, this case does not happen.

The above argument gives the following picture of the evolution of the algorithm: At each step $t = 1, \ldots, k$, we pick i_0 near T_r for some previously unvisited r, making it depleted, creating estimated cluster $\widehat{\mathcal{C}}_r$ and proceeding to the next iteration. After the k-th iteration all $T_\ell, \ell \in [k]$ will be depleted. We have $|\widehat{\mathcal{C}}_r| \geq |T_r|$, and $\widehat{\mathcal{C}}_r \subset T_r \cap T^c$ for all $r \in [k]$.

By construction $\{\widehat{\mathcal{C}}_\ell\}$ are disjoint. Let $\widehat{\mathcal{C}} := \biguplus_{\ell \in [k]} \widehat{\mathcal{C}}_\ell$, and note that $|\widehat{\mathcal{C}}| \geq |T|$ hence $|\widehat{\mathcal{C}}^c| \leq |T^c|$. Since $\widehat{\mathcal{C}}_\ell \cap T_r = \emptyset$ for $\ell \neq r$, we have $T_r \subset \bigcap_{\ell \neq r} \widehat{\mathcal{C}}_\ell^c$, hence $T_r \cap \widehat{\mathcal{C}}_r^c \subset \widehat{\mathcal{C}}^c$. All the misclassified or unclassified nodes produced by the algorithm are contained in $[\bigcup_r (T_r \cap \widehat{\mathcal{C}}_r^c)] \cup T^c$ which itself is contained in $\widehat{\mathcal{C}}^c \cup T^c$. Hence, the misclassification rate is bounded above by

$$\frac{1}{n}|\widehat{\mathcal{C}}^c \cup T^c| \leq \frac{1}{n}(|\widehat{\mathcal{C}}^c| + |T^c|) \leq \frac{2}{n}|T^c| \leq \frac{8\varepsilon^2}{n\rho^2}.$$

where we have used (A.14). The proof is complete. □

Remark 15. The last part of the argument can be made more transparent as follows: Each $\widehat{\mathcal{C}}_r$ consists of two disjoint part, $\widehat{\mathcal{C}}_r \cap T_r$ and $\widehat{\mathcal{C}}_r \cap T^c$. We have $|\widehat{\mathcal{C}}_r \cap T^c| \geq |T_r \setminus \widehat{\mathcal{C}}_r|$ (equivalent to $|\widehat{\mathcal{C}}_r| \geq |T_r|$). Then,

$$\sum_k |T_r \setminus \widehat{\mathcal{C}}_r| \leq \sum_r |\widehat{\mathcal{C}}_r \cap T^c| = |\widehat{\mathcal{C}} \cap T^c| \leq |T^c|$$

and total misclassifications are bound by $\sum_k |T_r \setminus \widehat{\mathcal{C}}_r| + |T^c|$.

A.8 Proofs of Chapter 5.4.1

Proof of Corollary 5. We have

$$\|\Lambda_{s*} - \Lambda_{t*}\|^2 = \sum_{\ell=1}^{k_2} n_{2\ell}^2 (B_{s\ell} - B_{t\ell})^2 = \sum_{\ell=1}^{k_2} \frac{n_{2\ell}^2}{n_1 n_2} (\Psi_{s\ell} - \Psi_{t\ell})^2 = \alpha \sum_{\ell=1}^{k_2} \pi_{2\ell}^2 (\Psi_{s\ell} - \Psi_{t\ell})^2$$

where we have used $\alpha := n_2/n_1$. Recall from (5.39) that $\Lambda_\wedge^2 := \min_{t \neq s} \|\Lambda_{s*} - \Lambda_{t*}\|^2$. It follows that $\alpha^{-1}\|\Lambda\|_\infty^2 \leq \Psi_{1,\wedge}^2$, using $\pi_{2,\ell} \leq 1$. We also have $\widetilde{\Psi}_{1,\wedge}^2 \geq \pi_{1,\wedge}\Psi_{1,\wedge}^2 \geq \pi_{1,\wedge}\alpha^{-1}\Lambda_\wedge^2$. Recalling the definition of a from (5.9), and using $\Psi_{s\ell} = (\sqrt{n_1 n_2}/n_{2\ell})\Lambda_{s\ell}$, we have

$$a = \sqrt{\frac{n_2}{n_1}}\|\Psi\|_\infty \leq \sqrt{\frac{n_2}{n_1}}\frac{\sqrt{n_1 n_2}}{n_{2,\wedge}}\|\Lambda\|_\infty = \frac{1}{\pi_{2,\wedge}}\|\Lambda\|_\infty \leq \beta_2 k_2 \|\Lambda\|_\infty.$$

Hence, $ka\Psi_{1,\wedge}^{-2} \leq k\,\beta_2 k_2 \|\Lambda\|_\infty (\alpha^{-1}\Lambda_\wedge^2)^{-1} = \beta_2 k k_2 \alpha \|\Lambda\|_\infty \Lambda_\wedge^{-2}$ which is the desired bound. We also note that

$$ka\widetilde{\Psi}_{1,\wedge}^{-2} \leq ka\pi_{1,\wedge}^{-1}\,\Psi_{1,\wedge}^{-2} \leq (\beta_1 k_1)\,ka\Psi_{1,\wedge}^{-2}$$

which combined with the previous bound shows that the required condition (5.40) in the statement is enough to satisfy (5.34). □

A.9 Auxiliary lemmas

Lemma 28. *Let $Z, Y \in \mathbb{O}^{n\times k}$ and let $\Pi_Z = ZZ^T$ and $\Pi_Y = YY^T$ be the corresponding projection operators. We have*

$$\min_{Q\in\mathbb{O}^{k\times k}} \|Z - YQ\|_F \;\le\; \|\Pi_Z - \Pi_Y\|_F.$$

Proof. We first note that $\|\Pi_Z\|_F^2 = \mathrm{tr}(\Pi_Z^2) = \mathrm{tr}(\Pi_Z) = k$ (since projections are idempotent), and $\|Z\|_F^2 = \mathrm{tr}(Z^TZ) = \mathrm{tr}(\Pi_Z) = k$. Let $Z^TY = U\Sigma V^T$ be the SVD of Z^TY where $U, V \in \mathbb{O}^{k\times k}$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \succeq 0$. Then, using the change of variable $O = V^TQU$,

$$\begin{aligned}\frac{1}{2}\min_Q \|Z - YQ\|_F^2 &= k - \max_Q \mathrm{tr}(Z^TYQ)\\ &= k - \max_{O\in\mathbb{O}^{k\times k}} \mathrm{tr}(\Sigma O) = k - \|\Sigma\|_*,\end{aligned}$$

where $\|\Sigma\|_* = \sum_i \sigma_i$ is the nuclear norm of Σ. To see the last equality, we note that since O is orthogonal, we have $|O_{ii}| \le 1$ for all i, hence $\max_O \mathrm{tr}(\Sigma O) \le \max_{\forall i,\, |O_{ii}|\le 1} \sum_i \sigma_i O_{ii} = \sum_i \sigma_i$ by the duality of ℓ_1 and ℓ_∞ norms and $\sigma_i \ge 0$. The equality is achieved by $O = I_k$. On the other hand

$$\frac{1}{2}\|\Pi_Z - \Pi_Y\|_F^2 = k - \mathrm{tr}(\Pi_Z\Pi_Y) = k - \|Z^TY\|_F^2 = k - \|\Sigma\|_F^2.$$

Since $\|\Sigma\| = \|Z^TY\| = \|Z\|\|Y\| \le 1$, we have $\sigma_i \le 1$ for all i. It follows that $\|\Sigma\|_F^2 \le \|\Sigma\|_*$ completing the proof. □

APPENDIX B

Extra Simulation Results

Here we present extra simulation results under the setup of Chapter 9. The following figure shows the overall NMI and log. error rate for different values of C and α:

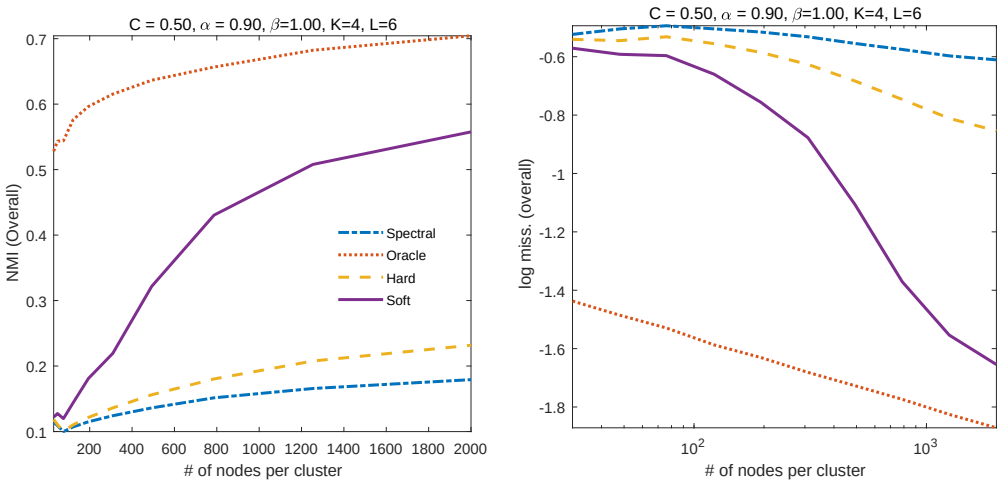


Figure B.1: Plots with different parameters.

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):

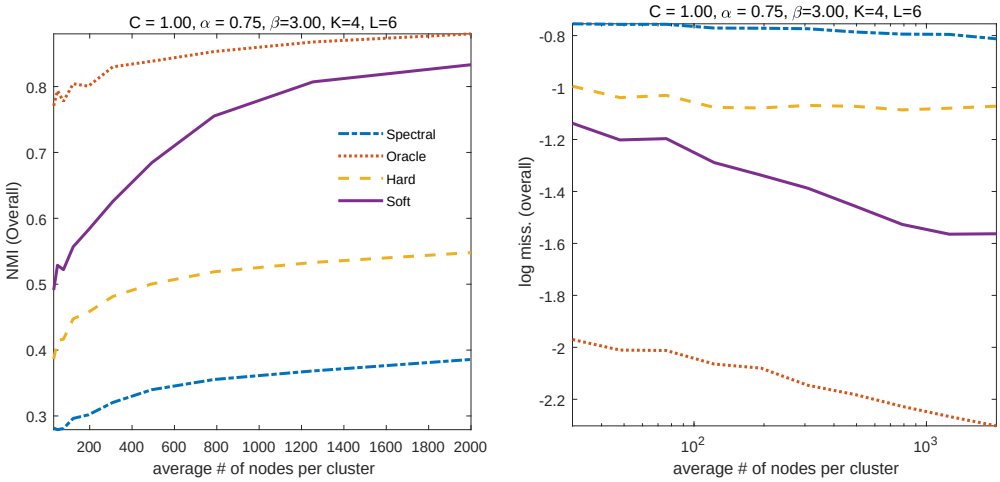


Figure B.2: Plots with unbalanced cluster sizes.

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:

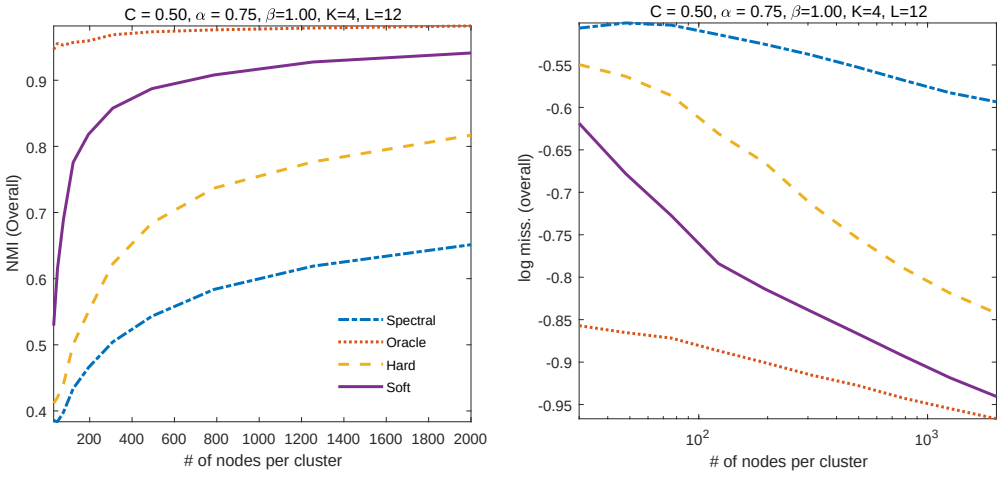


Figure B.3: Plots with different number of communities.

Bibliography

[Abb17] E. Abbe. "Community detection and stochastic block models: recent developments". In: *arXiv preprint arXiv:1703.10146* (2017).

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall. "Exact recovery in the stochastic block model". In: *IEEE Transactions on Information Theory* 62.1 (2016), pp. 471–487.

[ABKK17] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. "Multisection in the stochastic block model using semidefinite programming". In: *Compressed Sensing and its Applications*. Springer, 2017, pp. 125–162.

[ACBL+13] A. A. Amini, A. Chen, P. J. Bickel, E. Levina, et al. "Pseudo-likelihood methods for community detection in large sparse networks". In: *The Annals of Statistics* 41.4 (2013), pp. 2097–2122.

[AL14] A. A. Amini and E. Levina. "On semidefinite relaxations for the block model". In: *arXiv preprint arXiv:1406.5647* (2014).

[AL18] A. A. Amini and E. Levina. "On semidefinite relaxations for the block model". In: *The Annals of Statistics* 46.1 (2018), pp. 149–179.

[AS15] E. Abbe and C. 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*. IEEE. 2015, pp. 670–688.

[Ban15] A. S. Bandeira. "Random Laplacian matrices and convex relaxations". In: *Foundations of Computational Mathematics* (2015), pp. 1–35.

[BC09] P. J. Bickel and A. Chen. "A nonparametric view of network models and Newman–Girvan and other modularities". In: *Proceedings of the National Academy of Sciences* 106.50 (2009), pp. 21068–21073.

[BCCZ13] P. Bickel, D. Choi, X. Chang, and H. Zhang. "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *The Annals of Statistics* (2013), pp. 1922–1943.

[BLM15] C. Bordenave, M. Lelarge, and L. 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*. IEEE. 2015, pp. 1347–1357.

[BVH+16] A. S. Bandeira, R. Van Handel, et al. "Sharp nonasymptotic bounds on the norm of random matrices with independent entries". In: *The Annals of Probability* 44.4 (2016), pp. 2479–2506.

[CC00] Y. Cheng and G. M. Church. "Biclustering of expression data." In: *Ismb*. Vol. 8. 2000. 2000, pp. 93–103.

[CCT12] K. Chaudhuri, F. Chung, and A. Tsiatas. "Spectral clustering of graphs with general degrees in the extended planted partition model". In: *Conference on Learning Theory*. 2012, pp. 35–1.

[CDP12] A. Celisse, J.-J. Daudin, and L. Pierre. "Consistency of maximum-likelihood and variational estimators in the stochastic block model". In: *Electronic Journal of Statistics* 6 (2012), pp. 1847–1899.

[Che52] H. Chernoff. "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations". In: *The Annals of Mathematical Statistics* (1952), pp. 493–507.

[Chv79] V. Chvátal. "The tail of the hypergeometric distribution". In: *Discrete Mathematics* 25.3 (1979), pp. 285–287.

[CRV15] P. Chin, A. Rao, and V. Vu. "Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery". In: *Conference on Learning Theory*. 2015, pp. 391–423.

[CT06] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

[CX16] Y. Chen and J. Xu. "Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 882–938.

[Dhi01] I. S. Dhillon. “Co-clustering documents and words using Bipartite Co-clustering documents and words using Bipartite Spectral Graph Partitioning”. In: *Proc of 7th ACM SIGKDD Conf* (2001), pp. 269–274.

[Dhi03] I. S. Dhillon. “Information-Theoretic Co-clustering”. In: (2003), pp. 89–98.

[DKMZ11] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications”. In: *Physical Review E* 84.6 (2011), p. 066106.

[FH16] S. Fortunato and D. Hric. “Community detection in networks: A user guide”. In: *Physics Reports* 659 (2016), pp. 1–44.

[Fis+13] D. E. Fishkind et al. “Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown”. In: *SIAM Journal on Matrix Analysis and Applications* 34.1 (2013), pp. 23–39.

[GLM17] L. Gulikers, M. Lelarge, and L. Massoulié. “A spectral method for community detection in moderately sparse degree-corrected stochastic block models”. In: *Advances in Applied Probability* 49.3 (2017), pp. 686–721.

[GLMZ16] C. Gao, Y. Lu, Z. Ma, and H. H. Zhou. “Optimal estimation and completion of matrices with biclustering structures”. In: *Journal of Machine Learning Research* 17.161 (2016), pp. 1–29.

[GMZZ16] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. “Community Detection in Degree-Corrected Block Models”. In: *arXiv preprint arXiv:1607.06993* (2016).

[GMZZ17] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. “Achieving optimal misclassification proportion in stochastic block models”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 1980–2024.

[GN15] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Vol. 40. Cambridge University Press, 2015.

[GV16] O. Guédon and R. Vershynin. “Community detection in sparse networks via Grothendieck’s inequality”. In: *Probability Theory and Related Fields* 165.3-4 (2016), pp. 1025–1049.

[Har72] J. A. Hartigan. “Direct clustering of a data matrix”. In: *Journal of the american statistical association* 67.337 (1972), pp. 123–129.

[HLC60] J. L. Hodges and L. Le Cam. “The Poisson approximation to the Poisson binomial distribution”. In: *The Annals of Mathematical Statistics* 31.3 (1960), pp. 737–740.

[HWX16a] B. Hajek, Y. Wu, and J. Xu. “Achieving exact cluster recovery threshold via semidefinite programming”. In: *IEEE Transactions on Information Theory* 62.5 (2016), pp. 2788–2797.

[HWX16b] B. Hajek, Y. Wu, and J. Xu. “Achieving exact cluster recovery threshold via semidefinite programming: Extensions”. In: *IEEE Transactions on Information Theory* 62.10 (2016), pp. 5918–5937.

[Jam15] G. J. O. Jameson. “A simple proof of Stirling’s formula for the gamma function”. In: *The Mathematical Gazette* 99.544 (2015), p. 68.

[Jin15] J. Jin. “Fast community detection by SCORE”. In: *The Annals of Statistics* 43.1 (2015), pp. 57–89.

[JY13] A Joseph and B Yu. “Impact of regularization on Spectral Clustering”. In: *arXiv preprint arXiv:1312.1733* (2013). arXiv: arXiv:1312.1733v1.

[Krz+13] F. Krzakala et al. “Spectral redemption in clustering sparse networks”. In: *Proceedings of the National Academy of Sciences* 110.52 (2013), pp. 20935–20940.

[LCJ14] D. B. Larremore, A. Clauset, and A. Z. Jacobs. “Efficiently inferring community structure in bipartite networks”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 90.1 (2014), pp. 17–22. arXiv: 1403.2933.

[LLV17] C. M. Le, E. Levina, and R. Vershynin. “Concentration and regularization of random graphs”. In: *Random Structures & Algorithms* (2017).

[LR13] J. Lei and A. Rinaldo. “Consistency of spectral clustering in sparse stochastic block models. arXiv preprint”. In: *arXiv preprint arXiv:1312.2050* (2013).

[LR15] J. Lei and A. Rinaldo. “Consistency of spectral clustering in stochastic block models”. In: *The Annals of Statistics* 43.1 (2015), pp. 215–237.

[Mas14] L. Massoulié. "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM. 2014, pp. 694–703.

[MNS15] E. Mossel, J. Neeman, and A. Sly. "Consistency thresholds for the planted bisection model". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM. 2015, pp. 69–75.

[MS16] A. Montanari and S. 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*. ACM. 2016, pp. 814–827.

[MTSCO10] S. C. Madeira, M. C. Teixeira, I. Sa-Correia, and A. L. Oliveira. "Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm". In: *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 7.1 (2010), pp. 153–165.

[NS01] K. Nowicki and T. A. B. Snijders. "Estimation and prediction for stochastic blockstructures". In: *Journal of the American statistical association* 96.455 (2001), pp. 1077–1087.

[PW17] A. Perry and A. S. Wein. "A semidefinite program for unbalanced multisection in the stochastic block model". In: *Sampling Theory and Applications (SampTA), 2017 International Conference on*. IEEE. 2017, pp. 64–67.

[PZ17] M. Pensky and T. Zhang. "Spectral clustering in the dynamic stochastic block model". In: *arXiv preprint arXiv:1705.01204* (2017).

[RAL17] Z. S. Razaee, A. A. Amini, and J. J. Li. "Matched bipartite block model with covariates". In: *Preprint* (2017). arXiv: 1703.04943.

[RCY11] K. Rohe, S. Chatterjee, and B. Yu. "Spectral clustering and the high-dimensional stochastic blockmodel". In: *The Annals of Statistics* (2011), pp. 1878–1915.

[Roh15] K. Rohe. "Co-clustering for directed graphs : the Stochastic co-Blockmodel and spectral algorithm Di-Sim". In: (2015), pp. 1–39. arXiv: arXiv:1204.2296v2.

[RTJM16] F. Ricci-Tersenghi, A. Javanmard, and A. Montanari. “Performance of a community detection algorithm based on semidefinite programming”. In: *Journal of Physics: Conference Series*. Vol. 699. 1. IOP Publishing. 2016, p. 012015.

[TM10] D.-C. Tomozei and L. Massoulié. “Distributed user profiling via spectral methods”. In: *ACM SIGMETRICS Performance Evaluation Review*. 2010, pp. 383–384. arXiv: 1109.3318.

[Tro15] J. A. Tropp. “An introduction to matrix concentration inequalities”. In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230.

[TSS02] A. Tanay, R. Sharan, and R. Shamir. “Discovering statistically significant biclusters in gene expression data”. In: *Bioinformatics* 18.suppl_1 (2002), S136–S144.

[Ver86] S. Verdú. “Asymptotic error probability of binary hypothesis testing for poisson point-process observations (corresp.)” In: *IEEE Transactions on Information Theory* 32.1 (1986), pp. 113–115.

[Vu14] V. Vu. “A simple SVD algorithm for finding hidden partitions”. In: *arXiv preprint arXiv:1404.3918* (2014).

[WFL14] J. Wyse, N. Friel, and P. Latouche. “Inferring structure in bipartite networks using the latent block model and exact ICL”. In: 2013 (2014), p. 23. arXiv: 1404.2911.

[YP14] S.-Y. Yun and A. Proutiere. “Accurate community detection in the stochastic block model via spectral algorithms”. In: *arXiv preprint arXiv:1412.7335* (2014).

[ZLZ+12] Y. Zhao, E. Levina, J. Zhu, et al. “Consistency of community detection in networks under degree-corrected stochastic block models”. In: *The Annals of Statistics* 40.4 (2012), pp. 2266–2292.

[ZRMZ07] T. Zhou, J. Ren, M. Medo, and Y. C. Zhang. “Bipartite network projection and personal recommendation”. In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 76.4 (2007), pp. 1–7. arXiv: 0707.0540.

[ZZ+16] A. Y. Zhang, H. H. Zhou, et al. “Minimax rates of community detection in stochastic block models”. In: *The Annals of Statistics* 44.5 (2016), pp. 2252–2280.