Community Recovery on the Weighted Stochastic Block Model and Its Information-Theoretic Limits

Min Xu[†]
minx@wharton.upenn.edu

Varun Jog[‡]
vjog@wisc.edu

Po-Ling Loh[‡*]
loh@ece.wisc.edu

Department of Statistics[†]
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104

Departments of ECE[‡] & Statistics[*]
Grainger Institute of Engineering
University of Wisconsin - Madison
Madison, WI 53706

January 2017

Abstract

Identifying communities in a network is an important problem in many fields, including social science, neuroscience, military intelligence, and genetic analysis. In the past decade, the Stochastic Block Model (SBM) has emerged as one of the most well-studied and well-understood statistical models for this problem. Yet, the SBM has an important limitation: it assumes that each network edge is drawn from a Bernoulli distribution. This is rather restrictive, since weighted edges are fairly ubiquitous in scientific applications, and disregarding edge weights naturally results in a loss of valuable information. In this paper, we study a weighted generalization of the SBM, where observations are collected in the form of a weighted adjacency matrix, and the weight of each edge is generated independently from a distribution determined by the community membership of its endpoints. We propose and analyze a novel algorithm for community estimation in the weighted SBM based on various subroutines involving transformation, discretization, spectral clustering, and appropriate refinements. We prove that our procedure is optimal in terms of its rate of convergence, and that the misclassification rate is characterized by the Renyi divergence between the distributions of within-community edges and between-community edges. In the regime where the edges are sparse, we also establish sharp thresholds for exact recovery of the communities. Our theoretical results substantially generalize previously established thresholds derived specifically for unweighted block models. Furthermore, our algorithm introduces a principled and computationally tractable method of incorporating edge weights to the analysis of network data.

1 Introduction

The recent explosion of interest in network data has created a need for new statistical methods for analyzing network datasets and interpreting results [8, 11, 17, 23]. One active area of research with diverse applications in many scientific fields pertains to community detection and estimation, where the information available consists of the presence or absence of edges between nodes in the graph, and the goal is to partition the nodes into disjoint groups based on their relative connectivity [9, 14, 20, 24, 25, 28].

A standard assumption in statistical modeling is that conditioned on the community labels of the nodes in the graph, edges are generated independently according to fixed distributions governing the connectivity of nodes within and between communities in the graph. This is the setting of the stochastic block model (SBM) [15, 29, 30]. In the homogeneous case, edges follow one distribution when both endpoints are in the same community, regardless of the community label; and edges follow a second distribution when the endpoints are in different communities. The majority of existing literature on stochastic block models has focused on the case where no other information is available beyond the unweighted adjacency matrix, and much work in the information theory and statistics has focused on deriving thresholds for *exact* or *weak* recovery of community labels in terms of the underlying probability parameters and the size of the graph (e.g., [1–3, 12, 13, 19, 21, 22, 32]).

However, the pairwise connections in many real-world networks possess a natural weighting structure [7, 16]. For example, in social networks, information may be available quantifying the strength of a tie, such as the frequency of interactions between the individuals [27]; in cellular networks, information may be available quantifying the frequency of communication between users [6]; in gene co-expression networks, edges weights range from -1 to 1 and indicate the correlation between the expression levels of a gene pair; and in neural networks, edge weights may symbolize the level of neural activity between regions in the brain [26]. Of course, the connectivity data could be condensed into an adjacency matrix consisting of only zeros and ones, but this would result in a loss of valuable information that could be used to recover node communities.

In this paper, we analyze the "weighted" setting of the stochastic block model [4], where, after an edge is generated from a Bernoulli distribution, it is given an edge weight generated from one of two arbitrary densities $p(x), q(x)$ depending on whether the edge is between-cluster or within-cluster. The weighted SBM presents a serious challenge in the design of algorithms because $p(x), q(x)$ are unknown and must be estimated. Nonparametric estimation of a density is a difficult problem in its own right and it is made much harder in the weighted SBM because one does not know whether an edge weight is drawn from $p(x)$ and $q(x)$ without the latent cluster structure. There are various approaches to the weighted SBM. For example, Newman [16] assumes that the edge weights have discrete units and then converts a weighted graph into a multigraph; Aicher et al [4] assumes $p(x), q(x)$ to be from a known exponential family and performs variational Bayesian inference. These approaches can be effective but they rely on strong assumptions to simplify the problems and nothing is known about their theoretical properties.

Our paper proposes a new discretization based approach that imposes weak assumptions and possesses strong guarantees. In the case of finitely-supported distributions, which correspond to a "labeled" or "colored" SBM, we demonstrate a method for choosing an initial label on which we apply a standard SBM estimation method to obtain an initial clustering. We then show how to use this initial rough clustering, together with the full set of edge labels, to obtain more accurate estimates of the true cluster assignments. In the case of continuous weight distributions, we propose a discretization strategy that will allow us to apply a recovery algorithm for the labeled case after appropriate preprocessing. Our method does not rely on prior knowledge of the densities $p(x)$ and $q(x)$ and does not rely on parametric assumptions.

Importantly, we show that the output of our algorithm is optimal, in the sense that under mild regularity assumptions on $p(x)$ and $q(x)$, the misclustering error of our algorithm converges to zero at an optimal rate. Our analysis generalizes the results of Zhang and Zhou [32] and Gao et al [**?**], which show that the optimal rate of convergence of unweighted SBM is driven by the Renyi divergence of order 1/2 between two Bernoulli distributions, corresponding to the probability of generation for within-community and between-community edges. In fact, a similar phenomenon holds for the weighted SBM setting in our paper: the optimal error rate is also driven by a Renyi divergence of order 1/2 between two mixed distributions that capture both the divergence between the edge probabilities and the divergence between the edge weight densities $p(x)$ and $q(x)$. Note that in order to achieve the optimal error rate, our discretization strategy must be chosen carefully when $p(x)$ and $q(x)$ are continuous distributions. Our proposed algorithm first transforms the distributions to be supported on $[0, 1]$, then bins the interval appropriately; in general, since $p(x)$ and $q(x)$ may vary with the size of the graph, the number of bins used will also need to grow slowly as the number of nodes increases. Our results has an interesting implication: although our algorithm is nonparametric, it is adaptive in the sense that it achieves the same optimal rate even if the edge weight densities $p(x), q(x)$ take on a parametric form such as Gaussian or Laplace. This is in contrast to most problems in statistics where nonparametric methods usually have slower rate of convergence than parametric methods in settings where a parametric form is known. This observation captures an important intuition behind our results, that on the weighted SBM, one do not need to estimate the densities well in order to cluster well.

We also explore the related problem of exact recovery for weighted SBMs. Exact recovery refers to the case where the communities are partitioned perfectly, and a corresponding estimator is called *strongly consistent*. We analyze the performance of our algorithm in the case when the average number of edges scales according to $\Theta(\log n)$, known as the *sparse* regime in SBM literature. Again, we show that the thresholds for exact recovery may be expressed in terms of the Renyi divergence between weighted distributions, in the sense that our algorithm exactly recovers the true community labels when the Renyi divergence exceeds a certain threshold, and every algorithm fails with nontrivial probability when the Renyi divergence lies below the threshold.

The remainder of the paper is organized as follows: Section 2 introduces the mathematical framework of the weighted stochastic block model and defines the problems which we are trying to solve. Section 3 describes our proposed algorithm for finding communities on the weighted SBM. In Sections 4 and 5, we provide the statements of our main results concerning the behavior of our algorithm in terms of misclassification error rates and exact recovery. Section 6 highlights the key technical components employed in the analysis of our algorithm. We close in Section 7 with further implications of our work and open questions related to our results. Note that some of the content in this paper overlaps with an earlier version of our manuscript [18], which focused on finitely-supported edge weight distributions.

2 Model and problem formulation

We begin with a formal definition of the weighted SBM and a description of our error metrics for clustering.

2.1 Model definition

Consider a network with n nodes and $K \geq 2$ communities. In this paper, we suppose that the communities are approximately balanced; that is, there exists a *cluster-imbalance constant* β such that the cluster size n_k for each cluster $k = 1, \ldots, K$ satisfies $\frac{\beta n}{K} \geq n_k \geq \frac{n}{\beta K}$. For each node u, we let $\sigma(u) \in \{1, 2, \ldots, K\}$ denote the community assignment of the nodes.

Definition 2.1. (Homogeneous Stochastic Block Model) An edge random variable A_{uv} has the following distribution:

$$A_{uv} \sim \begin{cases} Ber(p) & \text{if } \sigma(u) = \sigma(v) \\ Ber(q) & \text{if } \sigma(u) \neq \sigma(v) \end{cases}$$

In the more general case of *heterogenous* SBM, we have a $K \times K$ matrix P where each entry $P_{ij} \in [0, 1]$. The edge random variable is drawn from $A_{uv} \sim Ber(P_{\sigma(u),\sigma(v)})$. We focus on the homogeneous case in this paper but discuss how to extend our results to the heterogenous setting.

SBM gives a distribution over the set of all networks whose edges are binary. To adapt to networks with continuous edge weights, we generalize the homogenous SBM by adding a second step to the data generating process: an edge weight is sampled from a continuous distribution after it is generated.

Definition 2.2. (Weighted Homogeneous SBM) Let $0 < P_0, Q_0 < 1$ and let $p(x), q(x)$ be two densities. We first generate the edge presence indicator Z_{uv}:

$$Z_{uv} \sim \begin{cases} Ber(1-P_0) & \text{if } \sigma(u) = \sigma(v) \\ Ber(1-Q_0) & \text{if } \sigma(u) \neq \sigma(v) \end{cases}$$

The edge weight random variable is then:

$$A_{uv} \sim \begin{cases} 0 & \text{if } Z_{uv} = 0 \\ p(x) & \text{if } Z_{uv} = 1 \text{ and } \sigma(u) = \sigma(v) \\ q(x) & \text{if } Z_{uv} = 1 \text{ and } \sigma(u) \neq \sigma(v) \end{cases}$$

In this model, an edge is missing with probability either P_0 or Q_0 depending on whether the potential edge connects two nodes in the same cluster or in different clusters. If the edge is present, then it is given an edge weight drawn from either the density $p(x)$ or $q(x)$, depending again on the nature of the edge. If $p(x)$ and $q(x)$ are Dirac Delta mass at 1, then the weighted homogenous SBM reduces to homogeneous SBM with $p = 1 - P_0$ and $q = 1 - Q_0$.

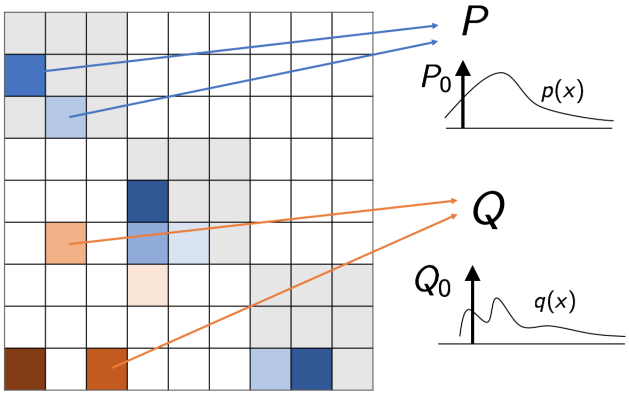

Figure 1: weighted stochastic block model

The model defined in 2.2 is the focus of our method. However, it is useful to note that we can further generalize model 2.2 by allowing both weights and labels.

Definition 2.3. (Weighted and Labeled Homogenous SBM) Let P, Q be two general mixed distributions. The edge random variable A_{uv} is drawn as

$$A_{uv} \sim \begin{cases} P & \text{if } \sigma(u) = \sigma(v) \\ Q & \text{if } \sigma(u) \neq \sigma(v) \end{cases}$$

In the case where P, Q are mixed distributions with continuous part $(1-P_0)p(x)$ and $(1-Q_0)q(x)$ respectively and a discrete point mass of P_0, Q_0 at zero respectively, then we get back the weighted SBM.

2.2 Community estimation

We study in this paper two problems on the weighted stochastic block model. Our first goal is to find a tractable community recovery algorithm whose misclustering error can be shown to converge to zero at an optimal rate. Our second goal is to find the threshold at which exact recovery of the communities shift from almost never possible to almost always possible.

2.2.1 Misclustering error rate

The goal of a community recovery algorithm is to take as input the adjacency matrix A and try to recover the community assignments. We evaluate a community recovery algorithm by looking at its mis-clustering error rate. To be precise, if σ_0 is the true clustering and $\widehat{\sigma}$ is the clustering generated by a community recovery algorithm, then the misclustering error rate is the following loss function:

$$l(\widehat{\sigma}, \sigma_0) \equiv \min_{\tau \in S_K} \frac{1}{n} \text{Hamming}(\widehat{\sigma},\, \tau \circ \sigma_0)$$

where Hamming$(\cdot,\, \cdot)$ denotes the Hamming distance. In the definition of mis-clustering error rate, we minimize over the set of permutations τ on K objects because clusterings are idenfiable only up to a permutation of their labels. It is important to note that $\widehat{\sigma}$ is a random quantity both because the community recovery algorithm may be stochastic and because the network A – the input to the algorithm – is random. Thus, we aim to bound $l(\widehat{\sigma}, \sigma_0)$ in probability.

Zhang and Zhou [32] and Gao et al [**?**] show that the minimax optimal rate of convergence for the unweighted stochastic block model is of the order $\exp\left(-(1+o(1))\frac{nI_{\text{Ber}}}{K}\right)$. $I_{\text{Ber}} = -2\log\sqrt{P_0 Q_0} + \sqrt{(1-P_0)(1-Q_0)}$ is the Renyi divergence of order 1/2 between Ber(P_0) and Ber(Q_0), where P_0, Q_0 are the probabilities of absence for within-community and between-communities edges. Yun and Proutiere have also characterized, though they present the results differently, the optimal rate of convergence for the labeled stochastic block model. Our work extends these results to the weighted SBM and show that the optimal rate is again governed by a Renyi divergence.

Although Renyi divergence is of central importance in homogenous stochastic block model where the cluster sizes are approximately balanced, it is important to note that, in the case of cluster imbalance or in the case of *heterogenous* stochastic block model, Abbe and Sandon [3] and Yun and Proutiere [31] have shown that an information divergence that generalizes the Renyi is what drives the intrinsic difficulty of community recovery – a generalization that is referred to as the CH-divergence.

2.2.2 Exact recovery

A closely related problem is that of finding the exact recovery threshold. We say that the weighted stochastic block model has an exact recovery threshold if there is some function of the parameters $\theta(P_0, Q_0, p(x), q(x), K, \beta, n)$ such that exact recovery is asymptotically almost always impossible if $\theta < 1$ and almost always possible if $\theta > 1$. For the homogeneous unweighted stochastic block model, Abbe et al [2] have shown that, when $\beta = 1, K = 2, 1 - P_0 = \frac{a \log n}{n}$, and $1 - Q_0 = \frac{b \log n}{n}$ (that is, the average degree is of order $\log n$) for some constant a, b, then the exact threshold is $\sqrt{a} - \sqrt{b}$, that is, no exact recovery algorithm can succeed if $\sqrt{a} - \sqrt{b} < 1$ and there exists a recovery algorithm that can succeed with probability tending to one if $\sqrt{a} - \sqrt{b} > 1$. This result was generalized by Zhang and Zhou [32] beyond the $\log n$ degree setting where $\frac{nI_{\text{Ber}}}{K \log n}$ was shown to be the threshold. Our paper again extends these results to the weighted stochastic block model where we show that the exact recovery threshold is a natural generalization of the unweighted analogues.

Apart from exact recovery (also known as strong consistency) and weak recovery, a notion of partial recovery (also known as weak consistency) has also been considered [5, 22, 32]. This notion lies between the other two notions of recovery, and only requires the fraction of misclassified nodes to converge in probability to 0 as n becomes large. A very general result for the $K = 2$ case, characterizing when exact and partial recovery are possible for the unweighted homogeneous stochastic block model, is provided in Mossel et al. [22].

3 Recovery algorithm

The weighted stochastic block model presents an extra layer of difficulty on top of the stochastic block model because the densities $p(x), q(x)$ are unknown. For example, one consequence of not knowing $p(x), q(x)$ is that the MLE does not exist. To see this, let us first see the MLE for stochastic block model: $\widehat{\sigma}_{MLE}^{SBM} = \arg\max_{\sigma} \sum_{\substack{(u,v)\in E \\ \sigma(u)=\sigma(v)}} \log \frac{p(1-q)}{q(1-p)}$. Since $\log \frac{p(1-q)}{q(1-p)} > 0$, the estimator $\widehat{\sigma}_{MLE}^{SBM}$ may be computed by searching for the clustering that maximizes the number of within cluster edges. In contrast, one can show, after straightforward algebraic manipulation, that likelihood maximization for wSBM takes on the form

$$\sup_{\sigma,\, p(x), q(x)\in\mathcal{P}} \sum_{\substack{(u,v)\in E \\ \sigma(u)=\sigma(v)}} \log \frac{p(A_{uv})(1-Q_0)}{q(A_{uv})(1-P_0)}$$

where $\mathcal{P}$ is the set of all densities. The maximum does not exist here because the maximizer of the likelihood does not exist for nonparametric density estimation. This remains true even if we restrict $\mathcal{P}$ to be the set of all smooth densities with say bounded second derivatives. Our approach therefore is to combine the idea of discretization from nonparametric density estimation with existing clustering techniques on the unweighted stochastic block model.

3.1 Algorithm overview

The key idea behind our method is to convert the edge weights into a finite set of labels by discretization. We then cluster on the labeled network. We first give a broad overview of our algorithm and then describe each steps in detail. Given a weighted network represented as an adjacency matrix A, our estimation method has four steps. We summarize the flow of the algorithm below and also in figure 3.

1. **Transformation.** We take as input a weighted matrix A and apply an invertible transformation function $\Phi : \mathbb{R} \to [0,1]$ to it. The resulting output $\Phi(A)$ is a matrix whose weights are between 0 and 1.

2. **Discretization.** We divide the $[0,1]$ interval into $l = 1, ..., L$ equally spaced subintervals, which we call bins. We replace the real-valued weight entries $\Phi(A)$ with a categorical label $l \in \{1, ..., L\}$: $[\Phi(A)]_{uv}$ is assigned label l if the value $[\Phi(A)]_{uv}$ falls into bin l. We output a network whose edges are colored with L possible colors.

3. **Initialization Part 1.** For each color l, we create a sub-network by including in it only edges whose color is l. For each sub-network, we perform spectral clustering. We output as l^* the color that induces the maximally separated spectral clustering.

4. **Initialization Part 2.** For each $u \in \{1, \ldots, n\}$, we perform spectral clustering on A^*_{-u}. We output n clusters $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n$.

5. **Refinement.** For each node $u \in \{1, \ldots, n\}$, we update the cluster assignment of u by considering $\widetilde{\sigma}_u$ and maximizing the likelihood looking at only the neighborhood around u.

6. **Consensus.** We align the cluster assignments made in the previous step.

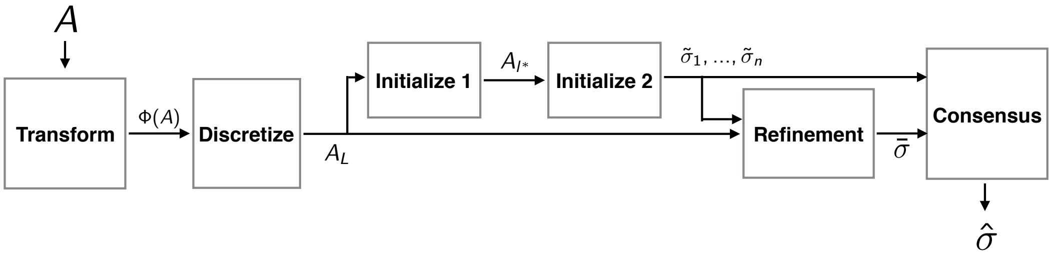

Figure 2: Pipeline for the our proposed algorithm

3.2 Transformation and discretization

These two steps are straightforward. In the transformation step, we apply an invertible CDF function $\Phi : \mathbb{R} \to [0,1]$ as the transformation function onto all the edge weights so that the transformed edge weights $\Phi(A)$ is in the interval $[0,1]$. In the discretization step, we divide the interval $[0,1]$ into L equally spaced bins labeled $l = 1, \ldots, L$. Each bin l is of the form $[a_l, b_l]$ where $a_1 = 0, b_L = 1$ and $b_l - a_l = 1/L$. We give an edge the label l if the weight of that edge falls into bin l.

Algorithm 3.1 Transformation and Discretization

Input: A weighted network A, a positive integer L, and an invertible function $\Phi : \mathbb{R} \to [0,1]$.
Output: A labeled network A_L

```
Divide $[0,1]$ into $L$ bins, labeled $Bin_1, \ldots, Bin_L$.
for every edge $(u,v)$ do
    let $l$ be the bin in which $\Phi(A_{uv})$ falls.
    Give the edge $(u,v)$ the label $l$ in the labeled network $A_L$
end for
Output $A_L$
```

3.3 Initialization

The initialization procedure takes as input a network whose edges are labeled with a color $l \in \{1, ..., L\}$. The goal of the initialization procedure is create a rough clustering $\widetilde{\sigma}$ that is sub-optimal but still consistent. As outlined in Algorithm 3.2, the rough clustering is based on a single color ℓ^*, which is chosen based on the maximum value of the estimated Renyi divergence between within-community and between-community distributions for the unweighted SBMs based on individual colors.

For technical reasons, we will actually create n separate rough clusterings $\{\widetilde{\sigma}_u\}_{u=1,\ldots,n}$ where each $\widetilde{\sigma}_u : [n-1] \to [K]$ is a clustering of a network of $n-1$ nodes where node u has been removed.

Spectral clustering: Note that Algorithm 3.2 involves several applications of spectral clustering. We describe the spectral clustering algorithm used as a subroutine in Algorithm 3.3 below:

Importantly, note that we may always choose the parameter μ sufficiently large such that Algorithm 3.3 generates a set S with $|S| = K$.

Algorithm 3.2 Initialization

Input: A labeled network A_L
Output: A set of clusterings $\{\widetilde{\sigma}_u\}_{u=1,\ldots,n}$

```
Separate $A_L$ into $L$ networks $\{A_l\}_{l=1,\ldots,l^*}$ where $A_l$ contains only edges with label $l$.   ▷ Stage 1
for each label $l$ do
    Compute $\overline{d} = \frac{1}{n}\sum_{u=1}^{n} d_u$ as the average degree.
    Perform spectral clustering with $\tau = \widetilde{C}\overline{d}$ and $\mu \geq C\beta$ to get $\widetilde{\sigma}_l$, where $\widetilde{C}, C$ are some large constants.
    estimate $\widehat{P}_l = \frac{\sum_{u\neq v\,:\,\widetilde{\sigma}_l(u)=\widetilde{\sigma}_l(v)}(A_l)_{uv}}{|u\neq v : \widetilde{\sigma}_l(u)=\widetilde{\sigma}_l(v)|}$ and $\widehat{Q}_l = \frac{\sum_{u\neq v\,:\,\widetilde{\sigma}_l(u)\neq\widetilde{\sigma}_l(v)}(A_l)_{uv}}{|u\neq v : \widetilde{\sigma}_l(u)\neq\widetilde{\sigma}_l(v)|}$.
    $\widehat{I}_l \leftarrow \frac{(\widehat{P}_l - \widehat{Q}_l)^2}{\widehat{P}_l \vee \widehat{Q}_l}$
end for
Choose $l^* = \arg\max_l \widehat{I}_l$. Let $A_{l^*}$ be the network with only edges labeled $l^*$.
for each node $u$ do   ▷ Stage 2
    Create network $A_{l^*} - \{u\}$ by removing node $u$ from $A_{l^*}$.
    Perform spectral clustering on $A_{l^*} - \{u\}$ to get $\widetilde{\sigma}_u$.
end for
Output the set of clusterings $\{\widetilde{\sigma}_u\}_{u=1,\ldots,n}$.
```

Algorithm 3.3 Spectral clustering

Input: An unweighted network A, trim threshold τ, number of communities K, tuning parameter μ
Output: A clustering σ

```
For each node $u$ whose degree $d_u \geq \tau$, set $A_{uv} = 0$ to get $T_\tau(A)$.
Let $\widehat{A}$ be the best rank-$K$ approximation to $T_\tau(A)$ in spectral norm.
For each node $u$, define the neighbor set $N(u) = \{v : \|\widehat{A}_u - \widehat{A}_v\|_2^2 \leq \mu K^2 \frac{\overline{d}}{n}\}$
Initialize $S \leftarrow 0$. Select node $u$ with the most neighbors and add $u$ into $S$ as $S[1]$
for $i = 2, \ldots, K$ do
    Among all $u$ such that $|N(u)| \geq \frac{n}{\mu K}$, select $u^* = \arg\max_u \min_{v\in S} \|\widehat{A}_u - \widehat{A}_v\|_2$.
    Add $u^*$ into $S$ as $S[i]$.
end for
for $u = 1, ..., n$ do
    Take $\arg\min_i \|\widehat{A}_u - \widehat{A}_{S[i]}\|_2$ and assign $\sigma(u) = i$.
end for
```

3.4 Refinement and consensus

Our refinement and consensus step closely follow the method described by Gao et al [**?**]. In the refinement step, we use the set of initial clusterings $\{\widetilde{\sigma}_u\}_{u=1,\ldots,n}$ to generate a more accurate clustering for the labeled network A_L. We do this by locally maximizing an approximate log-likelihood expression for each of the nodes $u = 1, \ldots, n$. The consensus step is to resolve a technical cluster label consistency problem that arises after the refinement stage.

Algorithm 3.4 Refinement

```
Input: A labeled network $A_L$ and a set of rough clusterings $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$
Output: a clustering $\widehat{\sigma}$ over the whole network
for each node $u$ do
    Estimate $\{\widehat{P}_l, \widehat{Q}_l\}_{l=0,\dots,L}$ from $\widetilde{\sigma}_u$.
    Let $\widehat{\sigma}_u : [n] \to [K]$ where $\widehat{\sigma}_u(v) = \widetilde{\sigma}_u(v)$ for all $v \neq u$ and
        $$\widehat{\sigma}_u(u) = \arg\max_k \sum_{v\,:\,\widetilde{\sigma}_u(v)=k,\, v\neq u} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$$
end for
Let $\widehat{\sigma}(1) = \widehat{\sigma}_1(1)$.                                  ▷ Consensus Stage
for each node $u \neq 1$ do
        $$\widehat{\sigma}(u) = \arg\max_k |\{v \,:\, \widehat{\sigma}_1(v) = k\} \cap \{v \,:\, \widehat{\sigma}_u(v) = \widehat{\sigma}_u(u)\}|$$
end for
Output $\widehat{\sigma}$
```

4 Analysis of misclustering error

On the unweighted stochastic block model, the key information quantity that governs the threshold behavior is $I = -2\log(\sqrt{pq} + \sqrt{(1-p)(1-q)})$. This is the Renyi divergence of order $\frac{1}{2}$ between the $Ber(p)$ distribution and the $Ber(q)$ distribution.

The Renyi divergence of order $\frac{1}{2}$ is defined on pairs of general measures as

$$I = -2\log \int \left(\frac{dP}{dQ}\right)^{1/2} dQ$$

Interestingly, this generalized form of the Renyi divergence is also what governs both the rate of convergence of our proposed algorithm and the threshold behavior of the weighted stochastic block model. In the weighted stochastic block model setting where P, Q have continuous part $p(x), q(x)$ and a point mass of probability P_0, Q_0 at zero, the Renyi divergence takes on the form

$$I = -2\log\left(\sqrt{P_0Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right)$$

When $I \to 0$, which is the scenario that we analyze, then I is also asymptotically equal to the Hellinger distance:

$$\begin{aligned} I &= \left\{(\sqrt{P_0} - \sqrt{Q_0})^2 + \int (\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)})^2 dx\right\}(1+o(1)) \\ &= \left\{(\sqrt{P_0} - \sqrt{Q_0})^2 + (\sqrt{1-P_0} + \sqrt{1-Q_0})^2 + \sqrt{(1-P_0)(1-Q_0)}\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right\} \\ &\quad \cdot (1+o(1)) \end{aligned} \tag{4.1}$$

Equation 4.1 shows that the Renyi divergence is driven by both the divergence between the edge probabilities $1-P_0, 1-Q_0$ as well as the divergence between the densities $p(x), q(x)$. This is a novel feature of the weighted stochastic block model.

When $p(x) = q(x)$, the Renyi divergence I reverts to the unweighted SBM case where it is a divergence between two Bernoulli distributions. This is intuitive because if $p(x) = q(x)$, then the edge weights give no additional information about the cluster structure. When $P_0 = Q_0$, then the Renyi divergence is driven only by the difference between the edge weight densities $p(x), q(x)$. This is also intuitive because if $P_0 = Q_0$, then the presence or absence of an edge offers no information on the cluster structure.

For the remainder of this paper, we define $H := \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$. Note that $H \leq 2$. It is important to note that $I \to 0$ quickly either when $H = o(1)$ or, if $\sqrt{(1-P_0)(1-Q_0)}$ is small, when $H = \Theta(1)$. Both of these are important cases to consider: in the first case, the challenge is to distinguish two densities $p(x), q(x)$ which are becoming increasingly similar; in the second case, the challenge is to estimate the density well when the amount of edges may be very sparse. The algorithm we propose in section 3 can handle both of these settings but the theoretical analyses are different.

4.1 Rate of convergence

Our analysis is asymptotic. We characterize the performance of the algorithm as $n \to \infty$. In our analysis, we treat $p(x), q(x), P_0, Q_0$ all as varying with n; we should properly write $p_n(x), q_n(x), P_{0n}, Q_{0n}$ but for the purpose of presentation, we omit the subscript and leave the dependency on n as implicit. All of our results will use the following assumption.

Assumption A0: There exist absolute constants c_0, C_0 such that $c_0 \leq \frac{1-P_0}{1-Q_0} \leq C_0$.

Assumption A0 says that the density of edges between the communities is of the same order as the density of edges across communities. This assumption is standard in the existing literature on unweighted stochastic block model.

Recall that $\Phi : \mathbb{R} \to [0,1]$ and that it must be invertible, differentiable, and a cumulative distribution function. We let ϕ denote Φ' and ϕ is thus the density of some distribution. We let $\Phi\{\cdot\}$ denote the Φ-measure of a set. Intuitively, the additional regularity conditions stated below require $p(x)$ and $q(x)$ to be smooth and the likelihood ratio $\frac{p(x)}{q(x)}$ to be well-behaved. Furthermore, the distribution ϕ must be heavier-tailed than $p(x)$ and $q(x)$. Note that $H = \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$ may either be $o(1)$ or $\Theta(1)$. Since these two cases lead to significant differences in their respective analyses, we require different sets of assumptions for $p(x)$ and $q(x)$ for each case.

4.1.1 The case $H = o(1)$

We state the assumptions for the case of $H = o(1)$, and then present our main result for this subproblem.

Regularity conditions:

A1 There exists a constant $C > 0$ such that $0 < p(x), q(x) \leq C$, and $p(x)$ and $q(x)$ are absolutely continuous. Moreover, the transformation density ϕ satisfies

$$\lim_{|x|\to\infty} \sup_n \frac{p(x) \vee q(x)}{\phi(x)} < \infty.$$

A2 There exists R a subinterval of $\mathbb{R}$ such that: (a) $\Phi\{R^c\} = o(H)$, and (b) $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$ for all $x \in R$ and for a constant ρ. (Recall that we define $H \equiv \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$.)

A3 Let $\alpha^2 = \int_R q(x) \left(\frac{p(x)-q(x)}{q(x)}\right)^2 dx$ and $\gamma(x) = \frac{q(x)-p(x)}{\alpha}$. There exists constants $M, r \geq 4$ such that

$$\int_R q(x) \left|\frac{\gamma(x)}{q(x)}\right|^r dx \leq M.$$

A4 Let $h(x) \geq \sup_n \max\left\{\left|\frac{\gamma'(x)}{q(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{\phi'(x)}{\phi(x)}\right|\right\}$. There exist constants M' and $1 \geq t \geq 2/r$ such that

$$\int_R |h(x)|^{2t/(1-t)}\phi(x)dx \leq M'.$$

Additionally, we require that the level set $\{x : |h(x)| \geq \kappa\}$ is a union of at most K_h intervals for all large enough κ where K_h is a constant. We also assume $\int \phi(x)^{\frac{1-t}{1+t}}dx < \infty$.

A5 There exists a constant $c' > 0$ such that $(\log p)'(x), (\log q)'(x) \geq (\log \phi)'(x) \geq 0$ for all $x < -c'$ and $(\log p)'(x), (\log q)'(x) \leq (\log \phi)'(x) \leq 0$ for all $x > c'$.

The simplest setting for which the assumptions are satisfied is when $p(x)$ and $q(x)$ are compactly supported (so the transformation Φ is not even necessary), have bounded first derivatives, likelihood ratio $\frac{p(x)}{q(x)}$ bounded away from 0 and infinity, and uniform convergence of $p(x) - q(x) \to 0$. However, this simple setting excludes many interesting cases, for example when $p(x)$ and $q(x)$ are Gaussian. To include such cases (cf. Section 4.1.2 below), we require the more technical conditions.

We then have the following result:

Theorem 4.1. *Suppose $\widehat{\sigma}$ is the output of the algorithm in section 3 with transformation Φ and discretization level L chosen such that $L \to \infty$, $L = o(\frac{1}{H})$, and $L = o(nI)$. Suppose that P_0, Q_0 satisfy assumption A0 and that $p(x), q(x)$ satisfy assumptions A1-A5 with respect to Φ. Suppose $\int(\sqrt{p(x)} - \sqrt{q(x)})^2dx = o(1)$, that K is fixed, and $I = o(1)$. Then, we have that*

$$\lim_{n\to\infty} P\left\{l(\widehat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI}{\beta K}(1 + o(1))\right)\right\} \to 1.$$

The proof of Theorem 4.1 is outlined in Appendix B.1.

4.1.2 Examples

Since conditions A1-A5 are rather technical, we illustrate with several concrete examples. Although we do not in general require $p(x)$ and $q(x)$ to belong to a parametric family, we will discuss cases where $p(x) = \exp(f_{\theta_1}(x))$ and $q(x) = \exp(f_{\theta_0}(x))$ where $f_\theta(x)$ is a set of functions indexed by θ where $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$ and where Θ is some compact subset of the Euclidean space. Although there is not a universal function Φ that works in all situations, it is generally sufficient, when $p(x), q(x)$ have subexponential tails, to take Φ as the CDF of the log-normal distribution. That is, we take $\phi(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left\{-\frac{\ln^2(x)}{2}\right\}$.

Example 4.1. (Gaussian with varying mean and variance)
Suppose $p(x) = N(\mu_1, \sigma_1^2)$ and $q(x) = N(\mu_0, \sigma_0^2)$ are both Gaussian with different mean and variance. Then, $\theta = (\mu, \sigma^2)$. We can take Θ to be any compact set where σ^2 is bounded away from 0. For example, we can let $\Theta = [-1, 1] \times [0.1, 2]$. Then, we have

$$f_\theta(x) = -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).$$

Since the log-normal distribution has all moments and has heavier tails than Gaussians, one can verify (through proposition 4.1 that the conditions A1-A5 are always satisfied.

Example 4.2. (Laplace with varying location and scale)
Suppose $p(x) = \frac{1}{2b_1}\exp\left(-\frac{|x-\mu_1|}{b_1}\right)$ and $q(x) = \frac{1}{2b_0}\exp\left(-\frac{|x-\mu_0|}{b_0}\right)$. Then $\theta = (\mu, b)$ and we can take Θ to be any compact set where b is bounded away from 0.

$$f_\theta(x) = -\frac{|x-\mu|}{b} - \log 2b.$$

Again, one can (through proposition 4.1) show that conditions A1-A5 are always satisfied.

Example 4.3. (Location Family) For a given density $\exp f(x)$ with mean zero, we can define a location family parametrized by μ as $\exp(f(x-\mu))$. We let $p(x) = \exp(f(x-\mu_1))$ and $q(x) = \exp(f(x-\mu_0))$. In this case, $\theta = \mu$ and we can let $\Theta = [-c, c]$ for some constant c.

In this case,

$$f_\theta = f(x-\mu).$$

One can show (through proposition 4.1) that conditions A1-A5 are always satisfied when $\sup_{\mu \in [-c,c]} |f'(x-\mu)|$ and $|f''(x-\mu)|$ are bounded by a polynomial of x.

4.1.3 The case $H = \Theta(1)$

We now state the assumptions we require in the case $H = \Theta(1)$.

Regularity conditions:

A1' There exists a constant $C > 0$ such that $0 \le p(x), q(x) \le C$, and $p(x)$ and $q(x)$ are absolutely continuous. Moreover, we assume that

$$\lim_{|x|\to\infty} \sup_n \frac{p(x) \vee q(x)}{\phi(x)} < \infty.$$

A2' For all large enough $\kappa > 0$, there exists a subinterval R of $\mathbb{R}$, and a constant $r > 2$ such that: (a) $\exp(-\kappa^{1/r}) \le \frac{p(x)}{q(x)} \le \exp(\kappa^{1/r})$, and (b) $\Phi\{R^c\} \le \frac{1}{2\kappa}$.

A3' Let $h(x) \ge \sup_n \max\left\{\left|\frac{p'(x)}{p(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{\phi'(x)}{\phi(x)}\right|\right\}$. There exist constants M' and $1 \ge t \ge 2/r$ such that $\int_R |h(x)|^{2t/(1-t)}\phi(x)dx \le M'$. Additionally, the level set $\{x : |h(x)| \ge \kappa\}$ is a union of at most K_h intervals for all large enough κ for a constant K_h. We also assume $\int \phi(x)^{\frac{1-t}{1+t}} dx < \infty$.

A4' There exists a constant $c' > 0$ such that $(\log p)'(x), (\log q)'(x) \ge (\log \phi)'(x) \ge 0$ for all $x < -c'$ and $(\log p)'(x), (\log q)'(x) \le (\log \phi)'(x) \le 0$ for all $x > c'$.

These assumptions are similar in nature to conditions A1-A5, and one can show that the examples in Section 4.1.2 also satisfy conditions A1'-A4'. Our main result here is as follows:

Theorem 4.2. *Suppose $\widehat{\sigma}$ is the output of the algorithm in section 3 with transformation Φ and discretization level L chosen such that $L \to \infty$ and $\frac{nI}{L\exp(L^{1/r})} \to \infty$. Suppose that P_0, Q_0 satisfy assumption A0 and that $p(x), q(x)$ satisfy assumptions A1'-A4' with respect to Φ. Suppose $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \Theta(1)$, that K is fixed, and that $I = o(1)$. Then, we have that*

$$\lim_{n\to\infty} P\left\{l(\widehat{\sigma}, \sigma_0) \le \exp\left(-\frac{nI}{\beta K}(1+o(1))\right)\right\} \to 1.$$

For the proof of Theorem 4.2, see Appendix B.2.

4.1.4 Additional discussion of assumptions

It is crucial to note that our algorithm does not require any knowledge about the form of $p(x), q(x)$. The same algorithm and the same guarantees apply whether $p(x)$ and $q(x)$ are Gaussian, Laplace, or any other (possibly non-parametric) distributions, as long as they satisfy conditions A1-A5 in conjunction with the transformation function Φ.

To aid the reader, we provide a brief non-technical interpretation of the regularity conditions.

Interpretation of assumptions A1-A5:

A1 Assumption A1 is simple; the second part states that Φ must have a tail just as heavy as that of $p(x)$ and $q(x)$.

A2 In Assumption A2, we require that the likelihood ratio $\frac{p(x)}{q(x)}$ be bounded away from 0 and ∞ except on a region $R^c \subset \mathbb{R}$. Since $H \to 0$, we have that $p(x), q(x)$ are becoming more similar and thus R^c is shrinking. We require that the measure of R^c, with respect to Φ, shrinks faster than H. This condition intuitively states that $|\frac{p(x)}{q(x)}|$ and its reciprocal tend to infinity slowly with respect to x. If Φ has a heavier tail, then A2 is a stronger condition on $\frac{p(x)}{q(x)}$. If Φ has a lighter tail, then A2 is a looser condition.

A3 In Assumption A3, note that because $H \to 0$, $\alpha \to 0$ as well. $\gamma(x) = \frac{p(x)-q(x)}{\alpha}$ is thus a function of constant order. The integrability condition on $\gamma(x)$ effectively says that $p(x) - q(x)$ must converge to 0 almost uniformly for all x in the region R. (Having an L_∞ bound on γ would imply uniform convergence.)

A4 Assumption A4 imposes smoothness on $q(x)$ as well as $\gamma(x)$. The second part of A4 is a weak condition that says $h(x)$ cannot oscillate with infinite frequency.

A5 Assumption A5 is another way of saying that ϕ must have a tail as heavy as that of $p(x)$ and $q(x)$.

Note that an analogous interpretation may be used to describe to the conditions A1'-A4'.

An alternative way to interpret these assumptions is that, for a given transformation Φ, there is a space $\mathcal{P}_\Phi(C, \rho, r, M, t, M', K_h, c')$ of densities that satisfy assumptions A1 to A5. We again emphasize that $\mathcal{P}_\Phi$ is actually a sequence of function spaces indexed by n; we make the dependence implicit in our notation. For a given Φ, assumption A1-A5 imposes a set of constraints on the densities $p(x), q(x)$. But suppose that $p(x), q(x)$ are given, it is difficult unfortunately to know how to choose an appropriate Φ from the statement of the assumptions.

Assumptions for parametric families: When $p(x)$ and $q(x)$ belong to a parametric family, as in the examples discussed in Section 4.1.2, it is helpful to consider a simpler set of assumptions. Suppose $p(x) = \exp(f_{\theta_1}(x))$ and $q(x) = \exp(f_{\theta_0}(x))$ where $f_\theta(x)$ is a set of functions indexed by θ where $\theta \in \Theta \subset \mathbb{R}^{d_\Theta}$ and where Θ is some compact subset of the Euclidean space. Consider the following conditions:

B1 For all $\theta \in \Theta$, $\liminf_{|x| \to \infty} ((\log \phi)(x) - f_\theta(x)) > -\infty$.

B2 Define the Fisher information matrix G_θ as

$$G_\theta = \int_{-\infty}^{\infty} (\nabla_\theta f_\theta(x))(\nabla_\theta f_\theta(x))^\mathsf{T} \exp(f_\theta(x)) dx.$$

We assume that this matrix is full-rank:

$$0 < c_{\min} < \inf_{\theta \in \Theta} \lambda_{min}(G_\theta) \le \sup_{\theta \in \Theta} \lambda_{max}(G_\theta) < c_{\max} < \infty$$

B3 There is some constant c such that $\sup_\theta \|\nabla_\theta f_\theta(x)\|$ is monotonically non-decreasing in $|x|$ for $|x| \ge c$.

B4 The following four integrability conditions hold:

$$\int_{-\infty}^{\infty} \sup_{\theta \in \Theta} \|\nabla_\theta f_\theta(x)\|^{r \vee \frac{2t}{1-t}} \phi(x) dx < \infty$$
$$\sup_{\theta \in \Theta} \int_{-\infty}^{\infty} \|\nabla_\theta f'_\theta(x)\|^{2t/(1-t)} \phi(x) dx < \infty$$
$$\sup_{\theta \in \Theta} \int_{-\infty}^{\infty} |f'_\theta(x)|^{2t/(1-t)} \phi(x) < \infty$$
$$\int_{-\infty}^{\infty} |(\log \phi)'(x)|^{2t/(1-t)} \phi(x) < \infty.$$

B5 There is some constant $c' > 0$ such that, for all θ, $f'_\theta(x) \geq (\log \phi)'(x) > 0$ for all $x \leq -c'$ and $f'_\theta(x) \leq (\log \phi)'(x) < 0$ for all $x \geq c'$.

Note that B4 translates to a moment condition on the transformation distribution Φ.

We then have the following result, proved in Section C:

Proposition 4.1. *Suppose assumptions B1-B5 hold. Then the following statements are true:*

(a) If $\|\theta_1 - \theta_0\| \to 0$, then $\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$ and that assumptions A1-A5 are satisfied.

(b) In the case where $\|\theta_1 - \theta_0\| = \Theta(1)$, then $\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \Theta(1)$ and that assumptions A1'-A4' are satisfied.

4.2 Lower bound

In this section, we give a lower bound on the performance of any clustering algorithms on the weighted stochastic block model. For technical reasons, we require the likelihood ratio $\frac{p(x)}{q(x)}$ to be bounded instead of approximately bounded as in Assumption A2 or A2'; we conjecture that the bounded likelihood ratio condition can be relaxed but leave its verification to future works. We also take the true clustering σ_0 to be random so that a clustering algorithm cannot use any prior information on σ_0; if σ_0 were fixed, then the algorithm that trivially outputs σ_0 would have error 0 for example. We also use a weak technical condition that $H = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$ cannot go to zero too quickly.

Theorem 4.3. *Suppose we have K clusters all of the same size and suppose the true clustering σ_0 is drawn uniformly at random.*

Suppose P_0, Q_0 satisfy assumption A0 and that $p(x), q(x)$ are two densities such that $\left|\log \frac{p(x)}{q(x)}\right| \leq C$ for some constant C. Suppose that $I \to 0$ and that $H = \omega\left(\sqrt{\frac{K}{n}}\right)$.

Then, we have that, for any community recovery algorithm $\widehat{\sigma}$:

$$\mathbb{E} l(\widehat{\sigma}, \sigma_0) \geq \exp\left(-(1 + o(1)) \frac{nI}{K}\right)$$

The proof of Theorem 4.3 is provided in Appendix D. The proof employs the same change of measure technique used by Yun and Proutiere to prove a similar lower bound on labeled stochastic block model [31]. We note that theorem 4.3 applies to any $p(x), q(x)$ that satisfy the assumptions; it does not take the supremum over a function space as with minimax lower bounds.

It is interesting to observe Theorem 4.3 in conjunction with Theorem 4.1 show that, in terms of rate of convergence, one does not have to pay a price for making nonparametric assumptions. That is, our nonparametric method achieves the same optimal rate even if the densities $p(x), q(x)$ take on a parametric form. This seemingly counter-intuitive phenonmenon arises because the cost of discretization is reflected in the $o(1)$ term in the exponent and is thus of lower order.

5 Exact recovery thresholds

We characterize the threshold for exact recovery of the communities on the weighted SBM, as Abbe et al [2] and Abbe and Sandon [3] have done for the unweighted SBM. We consider, in this subsection, the setting where $p(x)$, $q(x)$ are fixed with respect to n and that, for two positive constants a, b, that $P_0 = 1 - \frac{a \log n}{n}$ and $Q_0 = 1 - \frac{b \log n}{n}$. This is the *sparse* network setting where the average degree for any node is only of the order $\log n$. The following lemma characterizes the Renyi divergence in this setting.

Lemma 5.1. *In the setting where $p(x), q(x)$ are fixed and $P_0 = 1 - \frac{a \log n}{n}$ and $Q_0 = 1 - \frac{b \log n}{n}$, we have that*

$$I = (1 + o(1)) \frac{\log n}{n} \left((\sqrt{a} - \sqrt{b})^2 + \sqrt{ab} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \right)$$

where $o(1)$ is a term that goes to 0 as $n \to \infty$.

With this characterization of the Renyi divergence, we can describe the threshold for exact recovery on the weighted stochastic block model.

5.1 Error bounds for recovery algorithm

Theorem 5.1. *Suppose $K \geq 2$ is fixed and Φ is fixed, suppose that $p(x), q(x)$ satisfy assumptions A1'-A4' with respect to Φ, and suppose that*

$$(\sqrt{a} - \sqrt{b})^2 + \sqrt{ab} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 > K$$

Then, our algorithm, with a discretization level L that satisfies both $L \to \infty$ and $L = o(\log \log n)$, can exactly recover all the communities with probability converging to 1 as $n \to \infty$.

This is a corollary of theorem 4.2. If $(\sqrt{a} - \sqrt{b})^2 + \sqrt{ab} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx > K$, then we have that $\frac{nI}{K} > \log n$ and thus, the bound on the misclustering proportion is $\exp(-(1 + o(1))\frac{nI}{K}) < \frac{1}{n}$. Choosing $L = o(\log \log n)$ satisfies the conditions on L required in theorem 4.2.

5.2 Lower bounds

Theorem 5.1 shows that community recovery is possible when the Renyi divergence is above a certain threshold, the next theorem shows that community recovery is impossible when the Renyi divergence is below a certain threshold.

Theorem 5.2. *Suppose $K \geq 2$ is fixed, suppose that $\left| \log \frac{p(x)}{q(x)} \right|$ is bounded, and suppose that*

$$(\sqrt{a} - \sqrt{b})^2 + \sqrt{ab} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx < K$$

Then, every community recovery algorithm fails with probability at least $\frac{1}{3}$.

The proof of Theorem 5.2 is provided in Appendix E. It follows the technique of Abbe at al. [2]. Although we require a bounded log-likelihood-ratio condition that is stronger than the condition given in assumption A2', we not believe that this condition is necessary and it remains an open question how to relax this condition.

6 Proof sketch: Recovery algorithm

A large portion of the appendix is devoted to proving that our recovery algorithm succeeds and achieves the optimal error rates. Since this also constitutes the a significant part of the novel technical contribution of our paper, we provide an outline of the proof here. More details may be found in Appendix A.

We divide our argument into propositions that focus on successive stages of our algorithm. If we take a bird-eye view of our method, we find that it consists of two major components: first convert a weighted network into a labeled network, and then second, run community recovery algorithm on the

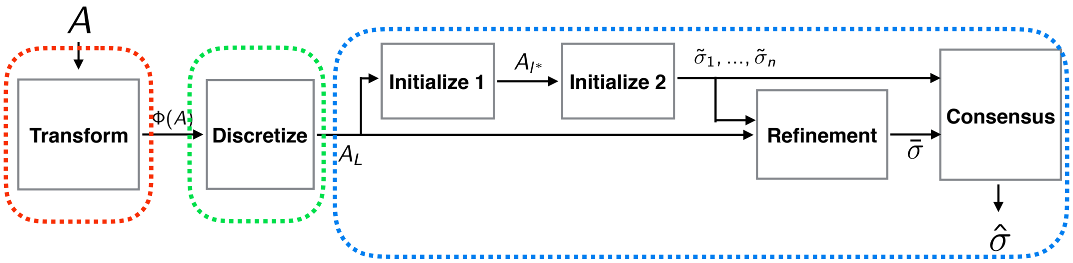

Figure 3: Analysis of the right-most blue region is in subsection 6.1, of the middle green region in subsection 6.2, and of the left-most red region in subsection 6.3

6.1 Analysis of community recovery on a labeled network

The workhorse behind our algorithm is a subroutine (right-most blue region in Figure 3) for recovering communities on a network where the edges have a discrete label $l = 1, \ldots L$. The following proposition characterizes the rate of convergence of the subroutine on the labeled stochastic block model where an edge within a community receives a label l with probability P_l and an edge between communities receives a label l with probability Q_l.

Proposition 6.1. *Suppose we have $l = 1, \ldots, L$ edge labels and suppose that the label probabilities satisfy $\frac{1}{\rho_L} \leq \frac{P_l}{Q_l} \leq \rho_L$ for a sequence $\rho_L = \Omega(1)$. Define $I_L = -2\log\sum_{l=0}^{L}\sqrt{P_l Q_l}$ and suppose $I_L \to 0$.*

Suppose $L = \Omega(1)$ and satisfies $\frac{nI_L}{L\rho_L^2} \to \infty$. Let $\widehat{\sigma}$ be the output of our algorithm. Then, we have that

$$\lim_{n\to\infty} P\left(l(\widehat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)\right) \to 1.$$

Yun and Proutiere [31] have proposed an algorithm for the labeled SBM that achieves the same rate of convergence. Proposition 6.1 is more general in that we allow the number of labels L and the bound on the ratio $\frac{P_l}{Q_l}$ to both go to infinity. This extension is critical for weighted SBM because to achieve consistency, we must let the discretization level L increase with n.

6.2 Discretization of the Renyi divergence

The rate of proposition 6.1 looks similar to that of theorems 4.1 and 4.2, except that instead of the actual Renyi divergence I, we have the discretized Renyi divergence I_L. A way to prove theorems 4.1

and 4.2 then is to show that the two quantities are close to each other; the following propositions do exactly that for distributions supported on $[0,1]$ and satisfying some additional assumptions.

It is easy to show that $I_L \leq I$ because discretization always loses information. If $p(x), q(x)$ are sufficiently regular in that they can be well approximated by discretization, then one might expect that I_L is not much smaller than I. Proposition 6.2 and 6.3 shows exactly that.

The following proposition is useful for proving theorem 4.1:

Proposition 6.2. *Let $p(z), q(z)$ be two densities supported on $[0,1]$. Suppose that $H \equiv \int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz = o(1)$. Let L be a sequence such that $L \to \infty$.*

Suppose the following assumptions are satisfied:

C1 Suppose $p(z), q(z) \leq C$ on $[0,1]$ and are absolutely continuous.

C2 There exists R a subinterval of $[0,1]$ such that $\frac{1}{\rho} \leq \left|\frac{p(z)}{q(z)}\right| \leq \rho$ and $\mu\{R^c\} = o(H)$ where μ is the Lebesgue measure.

C3 Define $\alpha^2 = \int_R \frac{(p(z)-q(z))^2}{q(z)} dz$ and $\gamma(z) = \frac{q(z)-p(z)}{\alpha}$. Suppose $\int_R q(z) \left|\frac{\gamma(z)}{q(z)}\right|^r dz \leq M$ for constants $M, r \geq 4$.

C4 Let $h(z) \geq \sup_n \max\left\{ \left|\frac{\gamma'(z)}{q(z)}\right|, \left|\frac{q'(z)}{q(z)}\right| \right\}$. Suppose $\int_R |h(z)|^t dz \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{z : |h(z)| \geq \kappa\}$ is a union of at most K intervals for all large enough κ.

C5 For all $z \leq \frac{1}{L}$, $p'(z), q'(z) \geq 0$ and for all $z \geq 1 - \frac{1}{L}$, we have that $p'(z), q'(z) \leq 0$.

Suppose $\frac{1}{c_0} \leq \frac{1-P_0}{1-Q_0} \leq c_0$. Let $bin_l = [a_l, b_l]$ for $l = 1, ..., L$ be a uniformly spaced binning of the interval $[0,1]$ and let $P_l = (1-P_0)\int_{a_l}^{b_l} p(z)dz$ and $Q_l = (1-Q_0)\int_{a_l}^{b_l} q(z)dz$. Suppose $L \to \infty$ but that $L \leq \frac{2}{H}$.

Define $I = -2\log\left(\sqrt{P_0Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(z)q(z)}dz\right)$ and $I_L = -2\log\left(\sqrt{P_0Q_0} + \sum_{l=1}^L \sqrt{P_lQ_l}\right)$.

Then, we have that

$$\left|\frac{I - I_L}{I}\right| = o(1)$$

and that $\frac{1}{4\rho c_0} \leq \frac{P_l}{Q_l} \leq 4\rho c_0$ for all l.

The following proposition is useful for proving theorem 4.2:

Proposition 6.3. *Let $p(z), q(z)$ be two densities supported on $[0,1]$. Suppose that $H \equiv \int (\sqrt{p(z)} - \sqrt{q(z)})^2 dz = \Theta(1)$.*

Let L be a sequence such that $L \to \infty$.

C1' Suppose $p(z), q(z) \leq C$ on $[0,1]$ and are absolutely continuous.

C2' There exists R a subinterval of $[0,1]$ such that $\exp(-L^{1/r}) \leq \frac{p(z)}{q(z)} \leq \exp(L^{1/r})$ and $\mu\{R^c\} \leq \frac{1}{2L}$.

C3' Let $h(z) \geq \sup_n \max\left\{ \left|\frac{p'(z)}{p(z)}\right|, \left|\frac{q'(z)}{q(z)}\right| \right\}$. Suppose $\int |h(z)|^t dz \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{z : |h(z)| \geq \kappa\}$ is a union of at most K intervals for all large enough κ.

C4' $p'(z), q'(z) \geq 0$ for all $z < \frac{1}{L}$ and $p'(z), q'(z) \leq 0$ for all $z > 1 - \frac{1}{L}$.

Suppose $\frac{1}{c_0} \leq \frac{1-P_0}{1-Q_0} \leq c_0$*. Let* $Bin_l = [a_l, b_l]$ *for* $l = 1, ..., L$ *be a uniformly spaced binning of the interval* $[0,1]$ *and let* $P_l = (1-P_0)\int_{a_l}^{b_l} p(z)dz$ *and* $Q_l = (1-Q_0)\int_{a_l}^{b_l} q(z)dz$*.*

Define $I = -2\log\left(\sqrt{P_0Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(z)q(z)}dz\right)$ *and* $I_L = -2\log\left(\sqrt{P_0Q_0} + \sum_{l=1}^{L}\sqrt{P_lQ_l}\right)$*.*

Then, we have that

$$\left|\frac{I - I_L}{I}\right| = o(1)$$

and that $\frac{1}{4\rho_L c_0} \leq \frac{P_l}{Q_l} \leq 4\rho_L c_0$ *for all* l*.*

6.3 Analysis on the transformation function

Proposition 6.2 and 6.3 considers densities supported on $[0,1]$. This is enough for us because once we transform the densities by an application of Φ, the new densities are compactly supported and, importantly, the Renyi divergence I and the Hellinger divergence H are invariant with respect to the transformation Φ.

To see this, let $p(x), q(x)$ denote densities over $\mathbb{R}$ and let $p_\Phi(z)$ and $q_\Phi(z)$ denote the transformed densities over $[0,1]$. It is easy to see that $p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$. Therefore, by a change of variables $z = \Phi^{-1}(x)$, we have that the following integrals are equal:

$$\int_{\mathbb{R}} \sqrt{p(x)q(x)}dx = \int_0^1 \sqrt{p_\Phi(z)q_\Phi(z)}dz$$

$$\int_{\mathbb{R}} (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \int_0^1 (\sqrt{p_\Phi(z)} - \sqrt{q_\Phi(z)})^2 dz$$

Therefore, the divergences I and H between $p(x), q(x)$ are the same as the divergences between $p_\Phi(z)$ and $q_\Phi(z)$.

To prove theorems 4.1 and 4.2, we then have to show that if the densities $p(x), q(x)$ satisfy assumptions A1-A5 (or A1'-A4'), then the transformed densities $p_\Phi(z), q_\Phi(z)$ satisfy assumptions C1-C5 (or C1'-C4') in proposition 6.2 (or proposition 6.3). This is done through proposition B.1 and B.2.

7 Conclusion

We have provided a rate-optimal community estimation algorithm for the homogeneous weighted stochastic block model. In the setting where the average degree is of order $\log n$ and the edge weight densities $p(x)$ and $q(x)$ are fixed, we have also characterized the exact recovery threshold. Our algorithm includes a preprocessing step consisting of transforming and discretizing the (possibly) continuous edge weights to obtain a simpler graph with edge weights supported on a finite discrete set. This approach may be useful for other network data analysis problems involving continuous distributions, where discrete versions of the problem are simpler to analyze.

Our paper is a first step toward understanding the weighted SBM under the same mathematical framework that has been so fruitful for the unweighted SBM. It is far from comprehensive, however, and many open questions remain. We describe a few here:

1. An important problem is to extend our analysis to the case of a *heterogenous* stochastic block model, where edge weight distributions depend on the exact community assignments of both endpoints. In such a setting, Abbe and Sandon [3] and Yun and Proutiere [31] have shown that a generalized information divergence—the CH divergence—governs the intrinsic difficulty of community recovery. We believe that a similar discretization-based approach should lead to

analogous results in the case of a heterogeneous weighted SBM. The key challenge would be to show that discretization does not lose much information with respect to the CH-divergence.

2. Real-world networks often have nodes with very high degrees, which may adversely affect the accuracy of recovery methods for the stochastic block model. To solve this problem, degree-corrected SBMs [10, 33] have been proposed as an effective alternative to regular SBMs. It is straightforward to extend the concept of degree-correction to the weighted SBM, but it is unclear whether our discretization-based approach would be effective in obtaining optimal error rates.

3. It is easy to extend our results to the weighted *and* labeled SBMs if the number of labels is finite or assumed to be slowly growing. However, this excludes some interesting cases, including the setting where edge labels represent counts from a Poisson distribution. We suspect that in such a situation, it may be possible to combine low-probability labels in a clever way to obtain a discretization that is again amenable to our approach.

References

[1] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[3] E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

[4] C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, page cnu026, 2014.

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

[6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.

[8] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge University Press, New York, NY, USA, 2010.

[9] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

[10] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.

[11] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, February 2010.

[12] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

[13] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.

[14] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.

[15] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[16] M. E. J. @inproceedingsbalakrishnan2011noise, title=Noise thresholds for spectral clustering, author=Balakrishnan, Sivaraman and Xu, Min and Krishnamurthy, Akshay and Singh, Aarti, booktitle=Advances in Neural Information Processing Systems, pages=954–962, year=2011 Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

[17] M. O. Jackson. *Social and Economic Networks.* Princeton University Press, 2010.

[18] V. Jog and P. Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

[19] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703. ACM, 2014.

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

[21] E. Mossel, J. Neeman, and A. Sly. Stochastic Block Models and Reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

[22] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.

[23] M. Newman, A.-L. Barabasi, and D. J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA, 2006.

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

[25] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[26] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. Computational Models of the Brain.

[27] D.S. Sade. Sociometrics of Macaca mulatta: I. Linkages and cliques in grooming matrices. *Folia Primatologica*, 18(3–4):196–223, 1972.

[28] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.

[29] S. Wasserman and C. Anderson. Stochastic a posteriori blockmodels: Construction and assessment. *Social Networks*, 9(1):1–36, 1987.

[30] H. C. White, S. A. Boorman, and R. L. Breiger. Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology*, 81(4):730–780, 1976.

[31] S. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

[32] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block model. *arXiv preprint arXiv:1507.05313*, 2015.

[33] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

A Proofs of propositions in Section 6

A.1 Proof of Proposition 6.1

The proof of Proposition 6.1 is quite involved. On a high level, we first show that the initialization stages output a set of rough clusterings $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n$ such that the error $l(\widetilde{\sigma}_u, \sigma_0)$ for any u is bounded by a sequence γ where $\gamma\rho_L^2 \to 0$, that is, the errors of the rough clusterings are small enough. Note that $\gamma\rho_L^2 \to 0$ also implies that $\gamma \to 0$. This step uses Propositions B.4 and B.5.

Then, we look at a single $\widetilde{\sigma}_u$ and show that because $l(\widetilde{\sigma}_u, \sigma_0)$ is small, the estimate $\widehat{P_l}, \widehat{Q_l}$ that we construct from it are close to the true P_l, Q_l. This is done through Proposition A.1. Since $\widehat{P_l}, \widehat{Q_l}$ are good estimators, we prove, in Proposition A.2, that we can use $\log \frac{\widehat{P_l}}{\widehat{Q_l}}$ derived from $\widetilde{\sigma}_u$ to correctly identify the true cluster membership of u with high probability. Finally, we analyze the consensus stage to show that we can at the end construct a single coherent clustering from all the $\widetilde{\sigma}_u$'s.

Formal details are provided below, with proofs of supporting propositions appearing in succeeding subsections.

Proof. (Proof of Proposition 6.1)

First we note that discarding color l where $P_l \vee Q_l \leq \frac{c}{n}$ does not affect the Renyi divergence I. To formalize this, define $P' = \sum_{l\,:\,P_l \wedge Q_l \geq c/n} P_l$ and $P'_l = P_l/P'$ and likewise for Q' and Q'_l. Define also $I' = -2\log \sum_{l\,:\,P_l, Q_l \geq c/n} \sqrt{P'_l Q'_l}$.

We claim that $I' = (1 + o(1))I_L$. We have

$$\begin{aligned}
I_L &= -2\log \sum_l \sqrt{P_l Q_l} \\
&\leq -2\log \sum_{l\,:, P_l \wedge Q_l \geq c/n} \sqrt{P_l Q_l} \\
&\leq -2\log \sum_{l\,:\,P_l \wedge Q_l \geq c/n} \sqrt{P'_l Q'_l} - 2\log \sqrt{P'Q'} \\
&\leq -2\log \sum_{l\,:\,P_l \wedge Q_l \geq c/n} \sqrt{P'_l Q'_l} - 2\log\left(1 - \frac{cL}{n}\right),
\end{aligned}$$

where the last inequality follows because $P' \geq 1 - \sum_{l\,:\,P_l \leq c/n} P_l \geq 1 - \frac{cL}{n}$, and likewise for Q'. Since $\frac{nI_L}{L} \to \infty$, we have that $\frac{cL}{n} = o(I_L) = o(1)$. The claim follows.

Now we prove the main result. Since $\frac{nI}{L\rho_L^2} \to \infty$, Proposition B.4 shows that the first initialization stage of our algorithm selects a color l^* that satisfies $\frac{n(P_{l^*} - Q_{l^*})^2}{(P_{l^*} \vee Q_{l^*})\rho_L^2} \to \infty$ with probability at least $1 - Ln^{3+\delta_p}$. Let us denote this event by E_1.

We then apply our analysis of spectral clustering (Proposition B.5) on each of the clusterings $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n$ and a union bound to show that $\max_u l(\widetilde{\sigma}_u, \sigma_0) \leq \gamma$ with probability at least $1 - n^3$, where

$$\gamma = CK^2 \frac{P_{l^*} \vee Q_{l^*}}{n(P_{l^*} - Q_{l^*})^2}$$

Let us denote this event by E_2.

Conditioned on E_1 and E_2 described above, we have that $\gamma\rho_L^2 \to 0$. Thus, we can apply Proposition A.1 on each of the rough clustering $\widetilde{\sigma}_u$'s and show that the conclusion of Proposition A.1 holds simultaneously for all $\widetilde{\sigma}_u$ with probability at least $1 - Ln^{2+\delta_p}$. Furthermore, the η that appears in

Proposition A.1 ($\eta = 4\left(\sqrt{2(5+\delta_p)K\gamma\log\frac{K}{\gamma}} + \frac{5+\delta_p}{c}K\gamma\log\frac{K}{\gamma} + \gamma K\right)$) satisfies $|\eta\rho_L| \to 0$. Let us denote this event by E_3.

Now we condition on E_3 and look at the refinement and consensus stage of the algorithm. Let $\widehat{\sigma}_u$ be the intermediate clustering on n nodes created in the refinement stage. By construction of $\widehat{\sigma}_u$, the error rate of $\widehat{\sigma}_u$ is at most $\gamma + \frac{1}{n}$ and thus, $l(\widehat{\sigma}_u, \sigma_0) < \frac{1}{8\beta K}$ for small enough γ. Let $\pi_u \in S_K$ denote the permutation such that $d(\pi_u(\widehat{\sigma}_u), \sigma_0) < \frac{n}{8\beta K}$.

It is also clear that for any u, the minimum cluster size of the clustering $\widehat{\sigma}_u$ is at least $\frac{n}{\beta K} - (n\gamma + 1) \geq \frac{n(1-\gamma k - 1/n)}{\beta K} \geq \frac{n}{2\beta K}$ for small enough γ. Therefore, we know by Lemma A.5 that π_u is the permutation that minimizes the $l(\widehat{\sigma}_u, \sigma_0)$.

Conditioned on E_3, we can apply proposition A.2 to conclude that

$$P(\widehat{\sigma}_u(u) \neq \pi_u^{-1}(\sigma_0(u)) \,|\, E_3) \leq (K-1)\exp\left(-(1+o(1))\frac{nI_L}{\beta K}\right)$$

By a triangle inequality argument, we have that $l(\widetilde{\sigma}_u, \widetilde{\sigma}_1) \leq 2\gamma + 2/n < \frac{1}{4\beta K}$. Thus, we can apply lemma A.4 on the pair $(\widetilde{\sigma}_1, \widetilde{\sigma}_u)$ to show that the consensus function ξ_u is the permutation that minimizes $d(\xi_u(\widetilde{\sigma}_1), \widetilde{\sigma}_u)$.

We claim then that $\pi_1 = \pi_u \cdot \xi_u$. We know that $d(\pi_1(\widetilde{\sigma}_1), \sigma_0) < \frac{n}{8\beta K}$ and that $d(\pi_u(\widetilde{\sigma}_u), \sigma_0) < \frac{n}{8\beta K}$. Therefore, $d(\widetilde{\sigma}_1, \pi_u(\pi_1^{-1}(\widetilde{\sigma}_u))) < \frac{n}{4\beta K}$. Since the minimum cluster size of both $\widetilde{\sigma}_1$ and $\widetilde{\sigma}_u$ is $\frac{n}{2\beta K}$, we have that $\pi_u^{-1} \cdot \pi_1 = \xi_u$ by Lemma A.5.

Let $\widehat{\sigma}$ be the output of the consensus stage. We have that

$$\begin{aligned} P(\widehat{\sigma}(u) \neq \pi_1^{-1}(\sigma_0(u)) \,|\, E_3) &= P(\widehat{\sigma}(u) \neq \xi_u^{-1} \circ \pi_u^{-1}(\sigma_0(u)) \,|\, E_3) \\ &= P(\widehat{\sigma}_u(u) \neq \pi_u^{-1}(\sigma_0(u)) \,|\, E_3) \end{aligned}$$

Now we are almost done with the proof.

$$\begin{aligned} P(\widehat{\sigma}(u) \neq \pi_1^{-1}(\sigma_0(u))) &\leq (K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + P(E_3^c \,|\, E_2, E_1) + P(E_2^c) + P(E_3^c) \\ &\leq (K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + Ln^{-(2+\delta_p)} + n^{-3} + Ln^{-3} \\ &\leq (K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + n^{-(1+\delta_p)} + n^{-3} + n^{-2} \\ &\leq (K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + n^{-(1+\delta_p)} \end{aligned}$$

where η' is some $o(1)$ sequence and where the third inequality follows because we take $L < n$. We take then $\eta'' = \eta' + \beta\sqrt{\frac{K}{nI_L}} = o(1)$.

First, suppose that $(K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) \geq n^{-(1+\delta_p/2)}$.

$$\begin{aligned} &P\left\{l(\widehat{\sigma}, \sigma_0) > (K-1)\exp\left(-(1-\eta'')\frac{nI_L}{\beta K}\right)\right\} \\ &\leq \frac{1}{(K-1)\exp\left(-(1-\eta'')\frac{nI_L}{\beta K}\right)}\frac{1}{n}\sum_{u=1}^{n} P(\widehat{\sigma}(u) \neq \pi_1^{-1}(\sigma_0(u))) \\ &\leq \exp\left\{-(\eta'' - \eta')\frac{nI_L}{\beta K}\right\} + \frac{Cn^{-(1+\delta_p)}}{(K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right)} \\ &\leq \exp\left\{-\sqrt{\frac{nI_L}{K}}\right\} + n^{-\delta_p/2} = o(1) \end{aligned}$$

Now, suppose that $(K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) \leq n^{-(1+\delta_p/2)}$.

$$\begin{aligned}
&P\left\{l(\widehat{\sigma},\sigma_0) > (K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right)\right\} \\
&= P(l(\widehat{\sigma},\sigma_0) > 0) \\
&\leq \sum_{u=1}^{n} P(\widehat{\sigma}(u) \neq \pi_1^{-1}(\sigma_0(u))) \\
&\leq n(K-1)\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + n^{-\delta_p} \leq n^{-\delta_p/2} = o(1)
\end{aligned}$$

□

A.1.1 Analysis of estimation error of $\widehat{P_l}$ and $\widehat{Q_l}$

Proposition A.1. *Let A_L be a labeled network with true clustering σ_0.*

Suppose σ is a random initial clustering and let us condition on the supposition that σ has error rate at most γ, that is, $l(\sigma,\sigma_0) \leq \gamma$.

Let $\Delta_l = |P_l - Q_l|$. Let $\widehat{P_l} = \frac{\sum_{u \neq v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(A_{uv}=l)}{\sum_{u\neq v\,:\,\sigma(u)=\sigma(v)} 1}$ and $\widehat{Q_l} = \frac{\sum_{u \neq v\,:\,\sigma(u)\neq\sigma(v)} \mathbf{1}(A_{uv}=l)}{\sum_{u\neq v\,:\,\sigma(u)\neq\sigma(v)} 1}$ be the MLE of P_l and Q_l based on σ. Let C_{thresh} be an absolute constant and let δ_p be a positive, fixed, and arbitrarily small real number. Let c be an absolute positive constant.

Then, with probability at least $1 - Ln^{-(3+\delta_p)}$, the following event happens:
For all l such that $P_l \vee Q_l \geq \frac{c}{n}$,

Case 1 if $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq C_{thresh}$, then

$$|\widehat{P_l} - P_l| \leq \eta\Delta_l$$
$$|\widehat{Q_l} - Q_l| \leq \eta\Delta_l$$

Case 2 if $\frac{n\Delta_l^2}{P_l \vee Q_l} \leq C_{thresh}$, then

$$|\widehat{P_l} - P_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}$$
$$|\widehat{Q_l} - Q_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}$$

Where $\eta = 4\left(\sqrt{2(5+\delta_p)K\gamma\log\frac{K}{\gamma}} + \frac{5+\delta_p}{c}K\gamma\log\frac{K}{\gamma} + \gamma K\right)$ is independent of the color l.

Proof. We first analyze the estimation error for a fixed σ and then take a union bound over all σ that satisfies $d_H(\sigma,\sigma_0) \leq \gamma n$.

Note that there are at most $\binom{n}{\gamma n}K^{\gamma n}$ possible assignments σ's that satisfy the error rate constraint.

$$
\begin{aligned}
\log \binom{n}{\gamma n} K^{\gamma n} &= \log\left(\frac{n(n-1)...(n-\gamma n+1)}{(\gamma n)!}\right) + \gamma n \log K \\
&\leq \log\left(\frac{n^{\gamma n} e^{\gamma n}}{(\gamma n)^{\gamma n}} \frac{1}{\sqrt{2\pi\gamma n}}\right) + \gamma n \log K \\
&\leq \log\left(\frac{e^{\gamma n}}{\gamma^{\gamma n}}\right) - \frac{1}{2}\log 2\pi\gamma n + \gamma n \log K \\
&\leq \gamma n \log \frac{e}{\gamma} + \gamma n \log K \\
&\leq 2\gamma n \log \frac{K}{\gamma}
\end{aligned}
$$

Next, we bound the bias of $\widehat{P}_l$.

Our estimator of P_l is

$$
\widehat{P}_l = \frac{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)} \mathbf{1}(A_{uv} = l)}{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)}}
$$

$\mathbb{E}\widehat{P}_l$ is a convex combination of P_l, Q_l.

$$
\begin{aligned}
\mathbb{E}\widehat{P}_l &= \frac{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)} \mathbf{1}(\sigma_0(u) = \sigma_0(v)) P_l + \mathbf{1}(\sigma_0(u) \neq \sigma_0(v)) Q_l}{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)} 1} \\
&= (1-\lambda)P_l + \lambda Q_l = P_l + \lambda(Q_l - P_l) \qquad \text{(A.1)}
\end{aligned}
$$

for $\lambda = \frac{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)} \mathbf{1}(\sigma_0(u) \neq \sigma_0(v))}{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)} 1}$.

Thus, we have that

$$
|\mathbb{E}\widehat{P}_l - P_l| \leq \lambda |Q_l - P_l|
$$

We need to upper bound λ. observe that

$$
\begin{aligned}
\lambda &= \frac{\sum_{u \neq v \,:\, \sigma(u)=\sigma(v)} \mathbf{1}(\sigma_0(u) \neq \sigma_0(v))}{\sum_{u \neq v} \mathbf{1}(\sigma(u) = \sigma(v))} \\
&= \frac{\sum_k \sum_{u \neq v \,:\, \sigma(u)=\sigma(v)=k} \mathbf{1}(\sigma_0(u) \neq \sigma_0(v))}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \\
&\leq \frac{\sum_k \sum_{u \neq v \,:\, \sigma(u)=\sigma(v)=k} \mathbf{1}(\neg(\sigma_0(u) = \sigma_0(v) = k))}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \\
&\leq \frac{\sum_k \sum_{u \neq v \,:\, \sigma(u)=\sigma(v)=k} \mathbf{1}(\sigma_0(v)) \neq k) + \mathbf{1}(\sigma_0(u) \neq k)}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)}
\end{aligned}
$$

Define $\gamma_k = \frac{1}{n}\sum_{u \,:\, \sigma(u)=k} \mathbf{1}(\sigma_0(u) \neq k)$ as the error rate within the estimated cluster k, and define $\widehat{n}_k = \sum_u \mathbf{1}(\sigma(u) = k)$. Then, we have that $\sum_k \gamma_k = \gamma$ and also $\sum_{u \,:\, \sigma(u)=k} \sum_{v \,:\, \sigma(v)=k} \mathbf{1}(\sigma_0(v) \neq k) = \gamma_k n \widehat{n}_k$. We continue the bound:

$$
\begin{aligned}
\lambda &\leq \frac{\sum_k 2\gamma_k n \widehat{n}_k}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \\
&= \frac{n}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \sum_k 2\gamma_k \widehat{n}_k \\
&\leq \frac{K}{n-K} n \sum_k 2\gamma_k \frac{\widehat{n}_k}{n} \\
&\leq 4\gamma K
\end{aligned}
$$

In the first inequality, we used the fact that $\sum_k \frac{\widehat{n}_k}{n}(\widehat{n}_k - 1) = n\sum_k \left(\frac{\widehat{n}_k}{n}\right)^2 - 1 \geq \frac{n}{K} - 1$ since $\sum_k \frac{\widehat{n}_k}{n} = 1$. In the last inequality, we used the assumption that $K < \frac{n}{2}$.

We then have an upper bound for $\lambda \leq 4\gamma K$. Therefore, we have that

$$|\mathbb{E}\widehat{P}_l - P_l| \leq 4\gamma K \Delta_l$$

where $\Delta_l = |Q_l - P_l|$. To simplify presentation, we define $\eta_1 = 4\gamma K$ so that

$$|\mathbb{E}\widehat{P}_l - P_l| \leq \eta_1 \Delta_l$$

From this it is clear that η_1 becomes arbitrarily small if γ is made arbitrarily small.

Having bounded the bias, we can now bound the variance.

Let $\widetilde{A}_{uv} = \mathbf{1}(A_{ij} = l)$. Then, by Bernstein's inequality,

$$P\left(\left|\sum_{u,v\,:\,\sigma(u)=\sigma(v)} (\widetilde{A}_{uv} - \mathbb{E}\widetilde{A}_{uv})\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{u,v\,\sigma(u)=\sigma(v)} \mathbb{E}\widetilde{A}_{uv} + \frac{2}{3}t}\right)$$

We first bound $\sum_{u,v\,\sigma(u)=\sigma(v)} \mathbb{E}\widetilde{A}_{uv}$:

$$\begin{aligned}\sum_{u,v\,\sigma(u)=\sigma(v)} \mathbb{E}\widetilde{A}_{uv} &= \sum_k \widehat{n}_k(\widehat{n}_k - 1)\mathbb{E}\widehat{P}_l \\ &\leq (P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1) \quad \text{(by Equation A.1)}\end{aligned}$$

Therefore,

$$P\left(\left|\sum_{u,v\,:\,\sigma(u)=\sigma(v)} (\widetilde{A}_{uv} - \mathbb{E}\widetilde{A}_{uv})\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1) + \frac{2}{3}t}\right)$$

Our goal is to choose an appropriate t such that the probability is upper bounded by $\exp(-2\gamma n \log \frac{K}{\gamma} - (3 + \delta_p)\log n)$.

We choose the following t

$$t^2 = 4\left\{\left(2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1)\right)\left(2\gamma n \log\frac{K}{\gamma} + (3+\delta_p)\log n\right)\right\} \vee 4\left\{\left(2\gamma n \log\frac{K}{\gamma} + (3+\delta_p)\log n\right)^2\right\}$$

We now verify that regardless of which term among $\{2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1),\, 2\gamma n \log\frac{K}{\gamma} + (3+\delta_p)\log n\}$ is larger, the probability term is at most $2\exp\left(-\left(2\gamma n \log\frac{K}{\gamma} + (3+\delta_p)\log n\right)\right)$. Let $A = 2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1)$ and $B = 2\gamma n \log\frac{K}{\gamma} + (3+\delta_p)\log n$.

- Suppose $A \geq B$, then $t^2 = 4AB$ and the probability term is at most $2\exp\left(-\frac{4AB}{A + \frac{4}{3}\sqrt{AB}}\right) \leq 2\exp\left(-\frac{4AB}{A+\frac{4}{3}A}\right) \leq 2\exp(-B)$
- Suppose $A \leq B$, then $t^2 = 4B^2$ and the probability term is at most $2\exp\left(-\frac{4B^2}{A+\frac{4}{3}B}\right) \leq 2\exp\left(-\frac{4B^2}{B+\frac{4}{3}B}\right) \leq 2\exp(-B)$.

Thus, with probability at most $\exp\left(-\left(2\gamma n\log\frac{K}{\gamma}+(3+\delta_p)\log n\right)\right)$,

$$|\widehat{P}_l-\mathbb{E}\widehat{P}_l|=\frac{\sum_{u,v\,\sigma(u)=\sigma(v)}(\widetilde{A}_{uv}-\mathbb{E}\widetilde{A}_{uv})}{\sum_{u,v}\mathbf{1}(\sigma(u)=\sigma(v))}>\frac{t}{\sum_{u,v}\mathbf{1}(\sigma(u)=\sigma(v))}$$

Now we derive a more manageable upper bound for t:
Note that

$$t^2\le 4\left\{\sqrt{\left(2(P_l\vee Q_l)\sum_k\widehat{n}_k(\widehat{n}_k-1)\right)\left(2\gamma n\log\frac{K}{\gamma}+(3+\delta_p)\log n\right)}+\left(2\gamma n\log\frac{K}{\gamma}+(3+\delta_p)\log n\right)\right\}^2$$

Note that we can without loss of generality assume $\gamma\ge\frac{1}{n}$. We then have that $\gamma n\log\frac{1}{\gamma}\ge\log n$. Thus, $2\gamma n\log\frac{K}{\gamma}+(3+\delta_p)\log n\le(5+\delta_p)\gamma n\log\frac{K}{\gamma}$.

$$\begin{aligned}\frac{t}{\sum_{u,v}\mathbf{1}(\sigma(u)=\sigma(v))}&\le 2\frac{\sqrt{(2(P_l\vee Q_l)\sum_k\widehat{n}_k(\widehat{n}_k-1))\left(2\gamma n\log\frac{K}{\gamma}+(3+\delta_p)\log n\right)}+\left(2\gamma n\log\frac{K}{\gamma}+(3+\delta_p)\log n\right)}{\sum_{u,v}\mathbf{1}(\sigma(u)=\sigma(v))}\\&\le 2\frac{\sqrt{2(P_l\vee Q_l)(5+\delta_p)\gamma n\log\frac{K}{\gamma}}}{\sqrt{\sum_k\widehat{n}_k(\widehat{n}_k-1)}}+2\frac{(5+\delta_p)\gamma n\log\frac{K}{\gamma}}{\sum_k\widehat{n}_k(\widehat{n}_k-1)}\\&\le 2\frac{\sqrt{2(P_l\vee Q_l)}\sqrt{(5+\delta_p)\gamma K\log\frac{K}{\gamma}}}{n-K}+2\frac{(5+\delta_p)\gamma K\log\frac{1}{\delta}}{n-K}\\&\le 4\sqrt{\frac{P_l\vee Q_l}{n}}\sqrt{(5+\delta_p)K\gamma\log\frac{K}{\gamma}}+4\frac{(5+\delta_p)K\gamma\log\frac{K}{\gamma}}{n}\end{aligned}$$

For the second to last inequality, we used the fact that $\sum_k(\widehat{n}_k-1)\frac{\widehat{n}_k}{n}=n\sum_k\left(\frac{\widehat{n}_k}{n}\right)^2-1\ge\frac{n}{K}-1$ because $\frac{\widehat{n}_k}{n}$ sums to 1. In the last inequality, we used the assumption that $n-K\ge\frac{n}{2}$.

To further simplify the expression, we have that $P_l\vee Q_l\ge\frac{c}{n}$ and thus, $\sqrt{\frac{P_l\vee Q_l}{n}}\ge\frac{c}{n}$. Therefore, we have that, with probability at least $1-\exp(-C_1\gamma n\log\frac{K}{\gamma}-(3+\delta)\log n)$,

$$|\widehat{P}_l-\mathbb{E}\widehat{P}_l|\le\sqrt{\frac{P_l\vee Q_l}{n}}\eta_2\qquad\text{(A.2)}$$

where $\eta_2=4\left(\sqrt{2(5+\delta_p)K\gamma\log\frac{K}{\gamma}}+\frac{(5+\delta_p)}{c}K\gamma\log\frac{K}{\gamma}\right)$. It is clear that η_2 can be made arbitrarily small by taking $\gamma K\log K$ to be arbitrarily small.

Taking the union bound across all clusterings with error γ and across all colors, we have that the probability of (A.2) holding simultaneously for all colors l is at least $1-Ln^{-(3+\delta_p)}$.

□

A.1.2 Analysis of probability of error for a single node

As a first step toward proving proposition 6.1, we first analyze the probability of error for a single node u.

Proposition A.2. *Let node u be arbitrarily fixed and suppose that $\frac{nI_L}{L}\to\infty$ and that for all l, $\frac{1}{\rho_L}\le\frac{P_l}{Q_l}\le\rho_L$. Suppose also that $\frac{1}{2}\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2\le\frac{1}{2}$. Conditioning on the event that the result* *of Proposition A.1 holds with a sequence η that satisfies $\eta\rho_L \to 0$, we have that, with probability at least $1-(K-1)\exp\left(-(1-o(1))\frac{n}{\beta K}I_L\right)$, the following event holds:*

$$\sigma_0(u) = \arg\max_k \sum_{v\,:\sigma_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l}\mathbf{1}(A_{uv}=l)$$

Proof. Throughout the proof, we let η' denote a sequence that converges to 0 and let C denote a $\Theta(1)$ sequence. Their value could change from line to line.

Let C_{thresh} be an absolute constant as defined in Proposition A.1.

First, define $L_1 = \{l \,:\, n\frac{\Delta_l^2}{P_l \vee Q_l} \geq C_{thresh}\}$. Then we claim that $C\sum_{l\in L_1}\frac{\Delta_l^2}{P_l\vee Q_l} = I_L(1-\eta')$. To see this, observe first that

$$\begin{aligned}
I_L &= -2\log\sum_l \sqrt{P_l Q_l} \\
&= C\sum_l (\sqrt{P_l}-\sqrt{Q_l})^2 \\
&= C\sum_l \frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2} \\
&= C\sum_l \frac{\Delta_l^2}{P_l \vee Q_l}
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
C\sum_{l\in L_1}\frac{\Delta_l^2}{P_l\vee Q_l} &= I_L - C\sum_{l\notin L_1}\frac{\Delta_l^2}{P_l\vee Q_l} \\
&\geq I_L - C_{\epsilon,c_1,c_2}\sum_{l\notin L_1}\frac{C_{thresh}}{n} \\
&\geq I_L - C_{\epsilon,c_1,c_2}\frac{LC_{thresh}}{n} \\
&\geq I_L - \eta' I_L
\end{aligned}$$

The first inequality follows from the definition of L_1. The third inequality follows because $\frac{nI_L}{L} \to \infty$ by assumption.

Now we proceed onto the main proof. Suppose without the loss of generality that $\sigma_0(u) = 1$. We want to then control the probability that for some cluster k,

$$\begin{aligned}
&\sum_{v\,:\,\sigma_u(v)=k}\sum_l \log\frac{\widehat{P}_l}{\widehat{Q}_l}\mathbf{1}(A_{uv}=l) \geq \sum_{v\,:\,\sigma_u(v)=1}\sum_l \log\frac{\widehat{P}_l}{\widehat{Q}_l}\mathbf{1}(A_{uv}=l) \quad \text{(iff)} \\
&\sum_{v\,:\,\sigma_u(v)=k}\overline{A}_{uv} - \sum_{v\,:\,\sigma_u(v)=1}\overline{A}_{uv} \geq 0 \qquad\qquad \text{(A.3)}
\end{aligned}$$

where $\overline{A}_{uv} \equiv \sum_l \log\frac{\widehat{P}_l}{\widehat{Q}_l}\mathbf{1}(A_{uv}=l)$.

Define $m_1 = |\{v \,:\, \sigma_u(v)=1\}|$ and $m_k = |\{v\,:\,\sigma_u(v)=k\}|$ as the size of clusters m_1, m_k under σ_u. Define $m_k' = \{v \,:\, \sigma_u(v)=k,\, \sigma_0(v)=k\}$, $m_1' = \{v\,:\,\sigma_u(v)=1,\,\sigma_0(v)=1\}$ as the points correctly clustered by σ_u.

With these definitions, the probability of the bad event Equation A.3 is upper bounded by the probability of the following:

$$\left(\sum_{i=1}^{m'_k}\widetilde{Y}_i+\sum_{i=1}^{m_k-m'_k}\widetilde{X}_i\right)-\left(\sum_{i=1}^{m'_1}\widetilde{X}_i+\sum_{i=1}^{m_1-m'_1}\widetilde{Y}_i\right)\geq 0 \quad \text{(iff)}$$
$$\exp(t\left(\sum_{i=1}^{m'_k}\widetilde{Y}_i+\sum_{i=1}^{m_k-m'_k}\widetilde{X}_i-\sum_{i=1}^{m'_1}\widetilde{X}_i-\sum_{i=1}^{m_1-m'_1}\widetilde{Y}_i\right))\geq 1$$

where $\widetilde{X}_i=\log\frac{\widehat{P}_l}{\widehat{Q}_l}$ with probability P_l and $\widetilde{Y}_i=\log\frac{\widehat{P}_l}{\widehat{Q}_l}$ with probability Q_l.

$$\begin{aligned}
&P\left(\exp(t\left(\sum_{i=1}^{m'_k}\widetilde{Y}_i+\sum_{i=1}^{m_k-m'_k}\widetilde{X}_i-\sum_{i=1}^{m'_1}\widetilde{X}_i-\sum_{i=1}^{m_1-m'_1}\widetilde{Y}_i\right))\geq 1\right)\\
&\leq \mathbb{E}\left(\exp(t\left(\sum_{i=1}^{m'_k}\widetilde{Y}_i+\sum_{i=1}^{m_k-m'_k}\widetilde{X}_i-\sum_{i=1}^{m'_1}\widetilde{X}_i-\sum_{i=1}^{m_1-m'_1}\widetilde{Y}_i\right))\right)\\
&\leq \mathbb{E}\left(\mathbb{E}[\exp(t\widetilde{Y}_i)\,|\,\widehat{P}_l,\widehat{Q}_l]\right)^{m'_k}\left(\mathbb{E}[\exp(t\widetilde{X}_i)\,|\,\widehat{P}_l,\widehat{Q}_l]\right)^{m_k-m'_k}\left(\mathbb{E}[\exp(-t\widetilde{X}_i)\,|\,\widehat{P}_l,\widehat{Q}_l]\right)^{m'_1}\left(\mathbb{E}[\exp(-t\widetilde{Y}_i)\,|\,\widehat{P}_l,\widehat{Q}_l]\right)^{m_1-m'_1}\\
&\leq \mathbb{E}\left(\sum_l e^{t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m'_k}\left(\sum_l e^{t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l\right)^{m_k-m'_k}\left(\sum_l e^{-t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l\right)^{m'_1}\left(\sum_l e^{-t\log\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m_1-m'_1}
\end{aligned}$$

We will set $t=\frac{1}{2}$, in which case, we have:

$$\left(\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m'_k}\left(\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l\right)^{m_k-m'_k}\left(\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}Q_l\right)^{m_1-m'_1}\left(\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l\right)^{m'_1}$$
$$=\left(\frac{\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}P_l}{\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l}\right)^{m_k-m'_k}\left(\frac{\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}Q_l}{\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l}\right)^{m_1-m'_1} \tag{A.4}$$
$$\left(\sum_l\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}Q_l\right)^{m_k}\left(\sum_l\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}}P_l\right)^{m_1} \tag{A.5}$$

We will bound term A.4 and A.5 separately. Loosely speaking, we will show that term A.4 is bounded in magnitude by $\exp(o(I_L)\frac{n}{K})$ and that term A.5 is bounded by $\exp(-\frac{n}{K}(1+o(1)I_L)$.

Bound for Term A.4.

Now, we can bound term A.4:

$$\begin{aligned}
\left|1 - \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right| &= \left|\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}(P_l - Q_l)}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right| \\
&\leq \frac{8}{\sum_l \sqrt{P_l Q_l}} \left|\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}}(P_l - Q_l)\right| \\
&\leq 16 \left|\sum_l \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right)(P_l - Q_l)\right| \\
&\leq 16 \left|\sum_{l \in L_1} \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right)(P_l - Q_l)\right| + 16 \sum_{l \notin L_1} \left|\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right| |P_l - Q_l| \\
&\leq 16 \sum_{l \in L_1} \frac{\Delta_l^2}{Q_l}(1 + \eta') + \sum_{i \notin L_1} 2C_{thresh} \frac{\Delta_l}{\sqrt{n(P_l \vee Q_l)}} \\
&\leq C I_L (1 + \eta') + 2C_{thresh}^2 \frac{L}{n} \\
&\leq C I_L (1 + \eta')
\end{aligned}$$

Where the second inequality follows from lemma A.3, second to last inequality follows under the assumption that $\sum_l \sqrt{P_l Q_l} \geq \frac{1}{2}$, and the last inequality follows from Lemma A.2.

Identical analysis shows that

$$\left|1 - \frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l}\right| = O(I_L)$$

Now, we note that $\exp(|1 - x|) \geq |x|$. Therefore, term A.4 can be bounded as

$$\begin{aligned}
&\left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right)^{m_k - m_k'} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l}\right)^{m_1 - m_1'} \\
&\leq \exp(O(I_L)(m_k - m_k' + m_1 - m_1')) \\
&\leq \exp(O(I_L)\gamma n) \\
&\leq \exp\left(\frac{n}{K} o(I_L)\right) \quad \text{(since } \gamma K \to 0)
\end{aligned}$$

Bound for Term A.5.

Define $\widehat{I} = -\log\left(\sum_l \frac{\widehat{P}_l}{\widehat{Q}_l} Q_l\right)\left(\sum_l \frac{\widehat{Q}_l}{\widehat{P}_l} P_l\right)$. With this definition,

$$\begin{aligned}
&\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{m_k} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{m_1} \\
&= \exp(-\widehat{I})^{\frac{m_k + m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{\frac{m_k - m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{\frac{m_1 - m_k}{2}}
\end{aligned}$$

We claim that the following three statements are true.

Claim 1 $m_k \geq n_k - \gamma n$ and likewise for m_1.

Claim 2 $\widehat{I} - I_L \geq -o(1) I_L$

Claim 3 $\left(\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} Q_l\right)^{\frac{m_k - m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} P_l\right)^{\frac{m_1 - m_k}{2}} = \exp(\frac{n}{k} o(I_L)$

Let us first suppose that these statements are true and see that term A.5 can be bounded.

$$
\begin{aligned}
&\exp(-\widehat{I})^{\frac{m_1+m_k}{2}} \left(\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} Q_l\right)^{\frac{m_k - m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} P_l\right)^{\frac{m_1 - m_k}{2}} \\
&\leq \exp(-(I^* + (\widehat{I} - I^*))^{\frac{m_1+m_k}{2}} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} P_l}{\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} Q_l}\right)^{\frac{m_1 - m_k}{2}} \\
&\leq \exp\left(-(1-o(1)) I_L (n_1 - \gamma n)\right) \left(\frac{\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} P_l}{\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} Q_l}\right)^{\frac{m_1 - m_k}{2}} \qquad \text{(by claim 1 and 2)} \\
&\leq \exp\left(-(1-o(1)) \frac{n}{\beta k} I_L\right) \quad \text{(by claim 3)}
\end{aligned}
$$

The last inequality holds because $\gamma = o\left(\frac{1}{K \log K}\right)$. We will prove each of the three claims in the remainder of the proof.

Claim 1: This is straightforward. σ_u has at most γn errors and therefore, $m'_1 \geq n_1 - \gamma n$ and $m_1 - m'_1 \leq \gamma n$.

Claim 2: We show that the estimation error of $\widehat{P_l}, \widehat{Q_l}$ does not make $\widehat{I}$ too small.

$$
\widehat{I} - I_L = -\log \frac{\left(\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} P_l\right)}{\left(\sum_l \sqrt{P_l Q_l}\right)^2} \tag{A.6}
$$

Let us consider the numerator.

$$
\begin{aligned}
&\left(\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} P_l\right) \\
&= \left(\sum_l \sqrt{P_l Q_l} \sqrt{\frac{\widehat{P_l}}{P_l} \frac{Q_l}{\widehat{Q_l}}}\right)\left(\sum_l \sqrt{P_l Q_l} \sqrt{\frac{P_l}{\widehat{P_l}} \frac{\widehat{Q_l}}{Q_l}}\right) \\
&= \sum_l P_l Q_l + 2 \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} + \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right) \\
&= \left(\sum_l \sqrt{P_l Q_l}\right)^2 + \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)
\end{aligned}
$$

where we define $T_{l,l'} = \frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}\frac{P_{l'}}{\widehat{P}_{l'}}\frac{\widehat{Q}_{l'}}{Q_{l'}}$. It will be later shown that $T_{l,l'} \to 1$ and thus, continuing equation A.6,

$$\begin{aligned}
\widehat{I} - I_L &= -\log\left(1 + \frac{\sum_{l<l'}\sqrt{P_lQ_lP_{l'}Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)}{\left(\sum_l \sqrt{P_lQ_l}\right)^2}\right) \\
&\geq -\log\left(1 + 4\sum_{l<l'}\sqrt{P_lQ_lP_{l'}Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)\right) \quad \text{(assuming that } \sum_l \sqrt{P_lQ_l} \geq 1/2\text{)} \\
&\geq -4\sum_{l<l'}\sqrt{P_lQ_lP_{l'}Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right) \qquad \text{(A.7)}
\end{aligned}$$

We proceed by first bounding $|T_{l,l'} - 1|$ and then taking the second order approximation of $\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)$ around 1.

$$\begin{aligned}
|T_{l,l'} - 1| &= \left|\frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}\frac{P_{l'}}{\widehat{P}_{l'}}\frac{\widehat{Q}_{l'}}{Q_{l'}} - 1\right| \\
&= \left|\left(1 - \frac{P_l - \widehat{P}_l}{P_l}\right)\left(1 - \frac{\widehat{Q}_l - Q_l}{\widehat{Q}_l}\right)\left(1 - \frac{\widehat{P}_{l'} - P_{l'}}{\widehat{P}_{l'}}\right)\left(1 - \frac{Q_{l'} - \widehat{Q}_{l'}}{Q_{l'}}\right) - 1\right| \\
&\leq \left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{\widehat{Q}_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{\widehat{P}_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}}\right) \\
&\leq 2\left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{P_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}}\right)
\end{aligned}$$

where the last inequality follows from lemma A.1.

Since we only work with pairs (l, l') such that $l' > l$ and we can choose whatever ordering we would like. Suppose that the l's are in decreasing order of $\frac{|\widehat{P}_l - P_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l}$ and therefore, we have that, for all pairs $l < l'$,

$$|T_{l,l'} - 1| \leq 4\left(\frac{|\widehat{P}_l - P_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l}\right)$$

By proposition A.1, we have that, for $l \in L_1$, $\frac{|P_l - \widehat{P}_l|}{P_l} \leq \eta\frac{\Delta_l}{P_l \vee Q_l}$ and for $l \notin L_1$, $\frac{|P_l - \widehat{P}_l|}{P_l} \leq \eta\frac{1}{\sqrt{n(P_l \vee Q_l)}}$ and likewise for the $\frac{|\widehat{Q}_l - Q_l|}{Q_l}$ term. We plug these bounds into the previous derivation and get that:

$$\begin{aligned}
|T_{l,l'} - 1| &\leq \eta'\frac{\Delta_l}{P_l \vee Q_l} \quad \text{for } l \in L_1 \\
|T_{l,l'} - 1| &\leq \eta'\frac{1}{\sqrt{n(P_l \vee Q_l)}} \quad \text{for } l \notin L_1
\end{aligned}$$

Where we have used the assumption that $\frac{1}{\rho_L} \leq \frac{P_l}{Q_l} \leq \rho_L$ and $\eta\rho_L \to 0$.

The Taylor approximation of $\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2$ around $T_{l,l'} = 1$ is:

$$\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 \leq \frac{1}{4}(T_{l,l'} - 1)^2 + O(T_{l,l'} - 1)^3$$

Continuing on from equation A.7, we have that

$$
\begin{aligned}
\widehat{I} - I_L &\geq -4 \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 \right) \\
&\geq -4 \sum_{l \in L_1} \sum_{l' > l} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 \right) - 4 \sum_{l \notin L_1} \sum_{l' > l} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 \right) \\
&\geq - \sum_{l \in L_1} \sum_{l' > l} \sqrt{P_l Q_l P_{l'} Q_{l'}} \eta' \left(\frac{\Delta_l}{P_l \vee Q_l} \right)^2 - \sum_{l \notin L_1} \sum_{l' > l} \sqrt{P_l Q_l P_{l'} Q_{l'}} \eta' \frac{1}{n(P_l \vee Q_l)} \\
&\geq -\eta' \left(\sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} \right) \left(\sum_{l'} \sqrt{P_{l'} Q_{l'}} \right) - \eta' \left(\sum_{l \notin L_1} \frac{1}{n} \right) \left(\sum_{l'} \sqrt{P_{l'} Q_{l'}} \right) \\
&\geq -o(I_L)
\end{aligned}
$$

The last inequality follows because $\sum_{l'} \sqrt{P_{l'} Q_{l'}} \leq 1$ and because $\sum_{l \notin L_1} \frac{1}{n} \leq \frac{L}{n} = o(I_L)$. This proves claim 2.

Claim 3.

$$
\begin{aligned}
&\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l \right)^{\frac{m_k - m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l \right)^{\frac{m_1 - m_k}{2}} \\
&= \left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l \right)^{\frac{m_k - m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l \right)^{\frac{m_1 - m_k}{2}} \left(\frac{\sum_l \sqrt{\widehat{P}_l \widehat{Q}_l}}{\sum_l \sqrt{\widehat{P}_l \widehat{Q}_l}} \right)^{\frac{m_1 - m_k}{2}} \\
&= \left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} \widehat{Q}_l} \right)^{\frac{m_k - m_1}{2}} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} \widehat{P}_l} \right)^{\frac{m_1 - m_k}{2}}
\end{aligned}
$$

Assume that $m_k \geq m_1$. The reverse case can be analyzed in the identical manner. Then,

$$
= \left(1 + \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} (Q_l - \widehat{Q}_l)}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} \widehat{Q}_l} \right)^{\frac{m_k - m_1}{2}} \left(1 + \frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} (\widehat{P}_l - P_l)}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l} \right)^{\frac{m_k - m_1}{2}}
$$

By lemma A.3, the denominators are of constant order. That is, $\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} \widehat{Q}_l = C$ and $\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l = C$.

To bound the numerator term, we apply lemma A.2.

$$\begin{aligned}
\left|\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}}(Q_l - \widehat{Q_l})\right| &= \left|\sum_l \left(\sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} - 1\right)(Q_l - \widehat{Q_l})\right| \\
&\leq \left|\sum_{l \in L_1} \left(\sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} - 1\right)(Q_l - \widehat{Q_l})\right| + \left|\sum_{l \notin L_1} \left(\sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} - 1\right)(Q_l - \widehat{Q_l})\right| \\
&\leq \sum_{l \in L_1} \eta' \frac{\Delta_l^2}{Q_l} + \sum_{l \notin L_1} \eta' \frac{1}{n} \\
&\leq \eta' I_L + \eta' \frac{L}{n} \\
&\leq \eta' I_L
\end{aligned}$$

The second inequality follows from lemma A.2 and the definition of L_1.

$$\begin{aligned}
&\left(1 + \frac{\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}}(Q_l - \widehat{Q_l})}{\sum_l \sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}}\widehat{Q_l}}\right)^{\frac{m_k - m_1}{2}} \left(1 + \frac{\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}}(\widehat{P_l} - P_l)}{\sum_l \sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}}P_l}\right)^{\frac{m_k - m_1}{2}} \\
&\leq \exp\left((m_k - m_1)\log(1 + o(I_L))\right) \\
&\leq \exp\left(\frac{n}{K} o(I_L)\right)
\end{aligned}$$

This proves claim 3.

Multiplying the bounds for term A.5 and A.4 shows that the probability of misclustering u into some cluster $k \neq 1$ is at most $\exp\left((-(1 + o(1))\frac{nI_L}{\beta K}\right)$. Taking a union bound over all clusters $k \neq$ completes the proof.

□

A.1.3 Lemmas

Here we collect various lemmas used in the proof.

We often use the bound that $\frac{1}{2}P \leq \widehat{P_l} \leq 2P_l$. The following lemma justifies this.

Lemma A.1. *Let l be any color and suppose that $\frac{1}{\rho_L} \leq \frac{P_l}{Q_l} \leq \rho_L$ where $\rho_L > 1$; suppose also that $P_l, Q_l \geq \frac{c}{n}$ for some absolute constant c.*

Let us condition on the event that the conclusion of proposition A.1 holds with a sequence η such that $\eta \rho_L \to 0$.

Then we have that

$$\max_l \frac{|\widehat{P_l} - P_l|}{P_l} \to 0 \quad \textit{and} \quad \max_l \frac{|\widehat{Q_l} - Q_l|}{Q_l} \to 0$$

In particular, for small enough η, we have that $\frac{1}{2}P \leq \widehat{P_l} \leq 2P_l$ for all l and likewise for Q_l.

Proof. We prove the statement first for P_l; the same argument goes for Q_l.

By proposition A.1, we have that either $|\widehat{P_l} - P_l| \leq \eta \Delta_l$ or $|\widehat{P_l} - P_l| \leq \eta \sqrt{\frac{P_l \vee Q_l}{n}}$.

Let us suppose $|\widehat{P_l} - P_l| \leq \eta \Delta_l$ first. Then,

$$\frac{|\widehat{P}_l - P_l|}{P_l} \leq \eta \frac{\Delta_l}{P_l} \leq \eta\rho_L \to 0$$

Now suppose that $|\widehat{P}_l - P_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}$. Then,

$$\frac{|\widehat{P}_l - P_l|}{P_l} \leq \eta\sqrt{\frac{\rho_L}{P_l n}} \leq \eta\rho_L\sqrt{\frac{1}{c}} \to 0$$

□

Lemma A.2. *Suppose that $\frac{1}{\rho_L} \leq \frac{P_l}{Q_l} \leq \rho_L$ for $\rho_L > 1$ and that $P_l, Q_l \geq \frac{c}{n}$ for some absolute constant c.*

Let us condition on the event that the conclusion of proposition A.1 holds with a sequence η such that $\eta\rho_L \to 0$.

Let C_{thresh} be the constant defined in proposition A.1. Then, the following are true:

1. *Suppose l satisfies $n\frac{\Delta_l^2}{P_l \vee Q_l} \geq C_{thresh}$. Then, we have that*

$$\left|\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right| \leq \left|\frac{P_l - Q_l}{Q_l}\right|(1 + \eta')$$

   where $\eta' \to 0$ and does not depend on the color l.

2. *Suppose l satisfies $n\frac{\Delta_l^2}{P_l \vee Q_l} \leq C_{thresh}$. Then, we have that*

$$\left|\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right| \leq 2C_{thresh}\frac{1}{\sqrt{n(P_l \vee Q_l)}}$$

And the same conclusion follows for $\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} - 1$.

Proof. First, suppose that l satisfies $n\frac{\Delta_l^2}{P_l \vee Q_l} \geq C_{thresh}$.

Note that

$$\frac{P_l}{Q_l} - 1 = \frac{P_l - Q_l}{Q_l}$$

We will show that $\frac{\widehat{P}}{\widehat{Q}}$ behaves similarly.

As a preliminary step, we note that, by lemma A.1, $\frac{\widehat{Q}_l - Q_l}{Q_l} = \eta'$ where $\max_l |\eta'| \to 0$.

In the following derivation, we use η' to denote a sequence such that $\max_l |\eta'| = o(1)$; the actual value of η' may change from instance to instance. We use η to denote a sequence where $\max_l |\eta\rho_L| = o(1)$.

$$\begin{aligned}
\frac{\widehat{P_l}}{\widehat{Q_l}} - 1 &= \frac{\widehat{P_l} - P_l + P_l}{\widehat{Q_l} - Q_l + Q_l} - 1 \\
&= \frac{\frac{\widehat{P_l} - P_l}{Q_l} + \frac{P_l}{Q_l}}{\frac{\widehat{Q_l} - Q_l}{Q_l} + 1} - 1 \\
&= \left(\frac{P_l}{Q_l} + \frac{\widehat{P_l} - P_l}{Q_l}\right)\left(1 - \frac{\widehat{Q_l} - Q_l}{Q_l}(1+\eta')\right) - 1 \\
&= \frac{P_l}{Q_l} + \frac{\widehat{P_l} - P_l}{Q_l} - \frac{P_l}{Q_l}\frac{\widehat{Q_l} - Q_l}{Q_l}(1+\eta') - \frac{\widehat{P_l} - P_l}{Q_l}\frac{\widehat{Q_l} - Q_l}{Q_l}(1+\eta') - 1 \\
&= \frac{P_l}{Q_l} + \eta\frac{\Delta_l}{Q_l} + \eta\rho_L\frac{\Delta_l}{Q_l} - 1 \\
&= \frac{P_l - Q_l}{Q_l}(1+\eta')
\end{aligned}$$

For the second to last equality, we used the fact that $|\widehat{P_l} - P_l| \leq \eta\Delta_l$ by the conclusion of proposition A.1.

If $P_l \geq Q_l$, then, for small enough value of η', we have that $\frac{\widehat{P_l}}{\widehat{Q_l}} \geq 1$. Hence,

$$\begin{aligned}
\sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} - 1 &\leq \frac{\widehat{P_l}}{\widehat{Q_l}} - 1 \\
&\leq \frac{P_l - Q_l}{Q_l}(1+\eta')
\end{aligned}$$

If $P_l \leq Q_l$, then, for small enough value of η', we have that $\frac{\widehat{P_l}}{\widehat{Q_l}} \leq 1$. Hence,

$$\begin{aligned}
\sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}} - 1 &\geq \frac{\widehat{P_l}}{\widehat{Q_l}} - 1 \\
&\geq \frac{P_l - Q_l}{Q_l}(1+\eta')
\end{aligned}$$

Symmetry yields that

$$\sqrt{\frac{\widehat{Q_l}}{\widehat{P_l}}} - 1 \begin{cases} \leq \frac{Q_l - P_l}{P_l}(1+\eta') & (\text{if } Q_l \geq P_l) \\ \geq \frac{Q_l - P_l}{P_l}(1+\eta') & (\text{if } Q_l \leq P_l) \end{cases}$$

This proves the first case.

The proof of the second case is almost identical. Let us assume that l is such that $n\frac{\Delta_l^2}{P_l \vee Q_l} \leq C_{thresh}$. In this case, we have that

$$\begin{aligned}
|\widehat{P_l} - P_l| &= \eta\sqrt{\frac{P_l \vee Q_l}{n}} \\
|\widehat{Q_l} - Q_l| &= \eta\sqrt{\frac{P_l \vee Q_l}{n}}
\end{aligned}$$

where $\max_l |\eta\rho_L| \to 0$.

By lemma A.1, $\frac{\widehat{Q}_l - Q}{Q_l} = \eta'$ where $\max_l |\eta'| = o(1)$.

In the following derivation, we use η' to denote a sequence such that $\max_l |\eta'| = o(1)$; the actual value of η' may change from instance to instance.

$$\begin{aligned}
\frac{\widehat{P}_l}{\widehat{Q}_l} - 1 &= \left(\frac{P_l}{Q_l} + \frac{\widehat{P}_l - P_l}{Q_l} \right) \left(1 - \frac{\widehat{Q}_l - Q_l}{Q_l}(1+\eta') \right) - 1 \\
&= \frac{P_l}{Q_l} + \frac{\widehat{P}_l - P_l}{Q_l} - \frac{P_l}{Q_l}\frac{\widehat{Q}_l - Q_l}{Q_l}(1+\eta') - \frac{\widehat{P}_l - P_l}{Q_l}\frac{\widehat{Q}_l - Q_l}{Q_l}(1+\eta') - 1 \\
&= \frac{P_l - Q_l}{Q_l} + \eta' \sqrt{\frac{1}{n(P_l \vee Q_l)}}
\end{aligned}$$

and, because $\frac{\widehat{P}_l}{\widehat{Q}_l} > 0$, it is clear that η' satisfies the condition that $\frac{P_l - Q_l}{Q_l} + \eta' \sqrt{\frac{1}{n(P_l \vee Q_l)}} + 1 > 0$.

$$\begin{aligned}
\left| \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right| &= \left| \sqrt{1 + \frac{P_l - Q_l}{Q_l} + \eta' \frac{1}{\sqrt{n(P_l \vee Q_l)}}} - 1 \right| \\
&\leq \left| \frac{P_l - Q_l}{Q_l} + \eta' \frac{1}{\sqrt{n(P_l \vee Q_l)}} \right| \\
&\leq 2C_{thresh} \frac{1}{\sqrt{n(P_l \vee Q_l)}}
\end{aligned}$$

The first inequality follows because $\sqrt{1+x} - 1 \leq x$ for $x \geq 0$ and $\sqrt{1+x} - 1 \geq x$ for $-1 < x < 0$. The second inequality follows because $\left| \frac{P_l - Q_l}{Q_l} \right| \leq C_{thresh} \frac{1}{\sqrt{n(P_l \vee Q_l)}}$.

□

The following lemma bounds $\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l$.

Lemma A.3. *Suppose that*

$$\frac{|\widehat{Q}_l - Q_l|}{Q_l} = \eta' \quad \frac{|\widehat{P}_l - P_l|}{P_l} = \eta'$$

where $\max_l |\eta'| = o(1)$.

Then, we have that for all small enough η,

$$\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l \geq \frac{1}{2} \sqrt{\widehat{P}_l \widehat{Q}_l} \geq \frac{1}{8} \sqrt{P_l Q_l}$$

Proof.

$$
\begin{aligned}
&\sqrt{\frac{\widehat{P_l}}{\widehat{Q_l}}Q_l} \\
&= \sqrt{\widehat{P_l}\widehat{Q_l}}\frac{Q_l}{\widehat{Q_l}} \\
&= \sqrt{\widehat{P_l}\widehat{Q_l}}\frac{1}{\frac{Q_l-\widehat{Q_l}}{Q_l}+1} \\
&= \sqrt{\widehat{P_l}\widehat{Q_l}}\left(1-\frac{\widehat{Q_l}-Q_l}{Q_l}(1+\eta')\right) \\
&= \sqrt{\widehat{P_l}\widehat{Q_l}}(1-\eta)
\end{aligned}
$$

where in the second to last equality, we used the fact that $\frac{\widehat{Q_l}-Q_l}{Q_l} = \eta' \to 0$.

Clearly then, for small enough η, we continue the bound as

$$\geq \frac{1}{2}\sqrt{\widehat{P_l}\widehat{Q_l}}$$

Also, for small enough η, we have that $\widehat{P_l} \geq \frac{1}{2}P_l$ and $\widehat{Q_l} \geq \frac{1}{2}Q_l$ and thus, we have the final bound

$$\frac{1}{8}\sqrt{P_lQ_l}$$

as desired.

□

We state Lemma 4 from [**?**] which analyzes the consensus step of the algorithm.

Lemma A.4. *Let σ, σ' be two clusters such that, for some constant $C \geq 1$, the minimum cluster size is at least $\frac{n}{Ck}$.*

Define a map $\xi : [k] \to [k]$ as $\xi(k) = \arg\max_{k'} |\{v : \sigma(v) = k\} \cap \{v : \sigma'(v) = k'\}|$.

Then, if $\min_{\pi \in S_k} l(\pi(\sigma), \sigma') < \frac{1}{Ck}$, we have that $\xi \in S_k$ and $l(\xi(\sigma), \sigma') = \min_{\pi \in S_k} l(\pi(\sigma), \sigma')$.

We include a simple additional lemma.

Lemma A.5. *Let $\sigma, \sigma' : [n] \to [k]$ be two clusterings where the minimum cluster size of σ is T. Let $\pi, \xi \in S_k$ be such that*

$$d(\pi(\sigma), \sigma') < T/2 \qquad d(\xi(\sigma), \sigma') < T/2$$

Then it must be that $\pi = \xi$.

Proof. Suppose not, then choose any k such that $\pi(k) \neq \xi(k)$.

$$|\{\sigma(u) = k\} \cap \{\sigma'(u) \neq \pi(k)\}| < d(\pi(\sigma), \sigma') < T/2$$

So, then, we have that $|\{\sigma(u) = k\} \cap \{\sigma'(u) = \pi(k)\}| > T/2$.

But then,

$$
\begin{aligned}
d(\xi(\sigma), \sigma') &\geq |\{\sigma(u) = k\} \cap \{\sigma'(u) \neq \xi(k)\}| \\
&\geq |\{\sigma(u) = k\} \cap \{\sigma'(u) = \pi(k)\}| \\
&\geq T/2
\end{aligned}
$$

□

A.2 Proof of Proposition 6.2

We first, for the convenience of the readers, restate the proposition:

Proposition A.3. *(Proposition 6.2)*
Let $p(x), q(x)$ be two densities supported on $[0,1]$. Suppose that $H \equiv \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = o(1)$ and suppose they satisfy the following assumptions:

- *C1 Suppose $p(x), q(x) \leq C$ on $[0,1]$ and are absolutely continuous.*
- *C2 There exists R a subinterval of $[0,1]$ such that $\frac{1}{\rho} \leq \left|\frac{p(x)}{q(x)}\right| \leq \rho$ and $\mu\{R^c\} = o(H)$ where μ is the Lebesgue measure.*
- *C3 Define $\alpha^2 = \int_R \frac{(p(x)-q(x))^2}{q(x)} dx$ and $\gamma(x) = \frac{q(x)-p(x)}{\alpha}$. Suppose $\int_R q(x) \left|\frac{\gamma(x)}{q(x)}\right|^r dx \leq M$ for constants $M, r \geq 4$.*
- *C4 Let $h(x) \geq \sup_n \max\left\{\left|\frac{\gamma'(x)}{q(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|\right\}$. Suppose $\int_R |h(x)|^t dx \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{x : |h(x)| \geq \kappa\}$ is a union of at most K_h intervals for all large enough κ.*
- *C5 For all $x \leq \frac{1}{L}$, $p'(x), q'(x) \geq 0$ and for all $x \geq 1 - \frac{1}{L}$, we have that $p'(x), q'(x) \leq 0$.*

Suppose $\frac{1}{c_0} \leq \frac{1-P_0}{1-Q_0} \leq c_0$. Let $Bin_l = [a_l, b_l]$ for $l = 1, ..., L$ be a uniformly spaced binning of the interval $[0,1]$ and let $P_l = (1-P_0)\int_{a_l}^{b_l} p(x)dx$ and $Q_l = (1-Q_0)\int_{a_l}^{b_l} q(x)dx$. Suppose $L \leq \frac{2}{H}$.
Define $I = -2\log\left(\sqrt{P_0Q_0} + \int\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right)$ and $I_L = -2\log\left(\sqrt{P_0Q_0} + \sum_{l=1}^L \sqrt{P_lQ_l}\right)$.

Then, we have that

$$\left|\frac{I - I_L}{I}\right| = o(1)$$

and that $\frac{1}{2\rho c_0} \leq \frac{P_l}{Q_l} \leq 2\rho c_0$ for all l.

Proof. We first prove the second claim. Define $\widetilde{P}_l = \int_{a_l}^{b_l} p(x)dx$ and $\widetilde{Q}_l = \int_{a_l}^{b_l} q(x)dx$.

It then follows from proposition A.8 that $\frac{1}{4\rho} \leq \frac{\widetilde{P}_l}{\widetilde{Q}_l} \leq 4\rho$. The claim follows from the bound on $\frac{1-P_0}{1-Q_0}$.

By lemma F.1 and A.6, we have that

$$\begin{aligned} I &= (1+o(1))\left\{(\sqrt{P_0} - \sqrt{Q_0})^2 + \int\left(\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)}\right)^2 dx\right\} \\ &= (1+o(1))\left\{(\sqrt{P_0} - \sqrt{Q_0})^2 + (\sqrt{1-P_0} - \sqrt{1-Q_0})^2 + \sqrt{(1-P_0)(1-Q_0)}\int\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx\right\} \end{aligned}$$

Likewise, we have that

$$\begin{aligned} I_L &= (1+o(1))\left\{(\sqrt{P_0} - \sqrt{Q_0})^2 + \sum_{l=1}^L\left(\sqrt{P_l} - \sqrt{Q_l}\right)^2 dx\right\} \\ &= (1+o(1))\left\{(\sqrt{P_0} - \sqrt{Q_0})^2 + (\sqrt{1-P_0} - \sqrt{1-Q_0})^2 + \sqrt{(1-P_0)(1-Q_0)}\sum_{l=1}^L(\sqrt{\widetilde{P}_l} - \sqrt{\widetilde{Q}_l})^2\right\} \end{aligned}$$

Proposition A.9 show that:

$$\left|\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx-\sum_{l=1}^{L}(\sqrt{\widetilde{P}_l}-\sqrt{\widetilde{Q}_l})^2\right|=o\left(\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx\right)$$

Thus,

$$I_L=(1+o(1))\left\{(\sqrt{P_0}-\sqrt{Q_0})^2+(\sqrt{1-P_0}-\sqrt{1-Q_0})^2+\sqrt{(1-P_0)(1-Q_0)}(1+o(1))\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx\right\}$$
$$=(1+o(1))I$$

□

A.2.1 Working with subinterval R

In this subsection, define an approximately uniform binning of an interval R to be the division of R into L bins $[a_l,b_l]$ such that the length of each bin is bounded $\frac{c_{bin}}{L}\le b_l-a_l\le\frac{C_{bin}}{L}$ for some constants c_{bin} and C_{bin}.

Proposition A.4. *Let $H^R=\int_R(\sqrt{p(x)}-\sqrt{q(x)})^2dx$. Define $\delta(x)=q(x)-p(x)$ and let α be a real number such that*

$$C'\int_R q(x)\left(\frac{\delta(x)}{q(x)}\right)^2dx\ge\alpha^2\ge c'\int_R q(x)\left(\frac{\delta(x)}{q(x)}\right)^2dx$$

for some constants C',c'. Define $\gamma(x)=\frac{\delta(x)}{\alpha}$. Suppose that $\limsup_n\int_R q(x)\left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^4dx\le M$. Then, we have that,

$$H^R=d\alpha^2(1+\eta)$$

where $d=\int_R p(x)\left(\frac{1}{2}\frac{\gamma(x)}{p(x)}\right)^2dx$ and $|\eta|\le(2\alpha+\alpha^2)\frac{C'M}{4}$. In particular, we have that if $H^R\to0$, then $\alpha\to0$ and $\eta\to0$.

Proof. First, let us note that

$$\frac{1}{c'}\ge\int_R q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2dx\ge\frac{1}{C'}$$

$$\begin{aligned}H^R&=\int_R(\sqrt{p(x)}-\sqrt{q(x)})^2dx\\&=\int_R(\sqrt{q(x)}-\sqrt{q(x)-\delta(x)})^2dx\\&=\int_R q(x)\left(1-\sqrt{1-\frac{\delta(x)}{q(x)}}\right)^2dx\end{aligned}$$

By convention, we let $\frac{\delta(x)}{q(x)}=0$ whenever $q(x),p(x)=0$.

Thus, we can define $\xi(x)=1-\frac{1}{2}\frac{\delta(x)}{q(x)}-\sqrt{1-\frac{\delta(x)}{q(x)}}$ for $x\in[0,1]$.

$$\begin{aligned}&=\int_R q(x)\left(1-(1-\frac{1}{2}\frac{\delta(x)}{q(x)}-\xi(x))\right)^2dx\\&=\int_R q(x)\left(\frac{1}{2}\frac{\delta(x)}{q(x)}-\xi(x)\right)^2dx\\&=\int_R q(x)\left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2(1-\xi_2(x))^2dx\end{aligned}$$

Where $\xi_2(x) = \frac{2\xi(x)}{\delta(x)/q(x)}$ if $\delta(x) \neq 0$ and $\xi_2(x) = 0$ if $\delta(x) = 0$.
Thus,

$$\int_R \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx = (1-\eta) \int_R q(x) \left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2 dx$$
$$= (1-\eta)\alpha^2 \int_R q(x) \left(\frac{1}{2}\frac{\gamma(x)}{q(x)}\right)^2 dx$$

where

$$\eta = \frac{\int_R q(x) \left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2 (\xi_2(x)^2 + 2\xi_2(x))dx}{\int_R q(x) \left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2}$$

By lemma F.3, $\xi_2(x) \leq 2\left|\frac{\delta(x)}{q(x)}\right|$. Thus, we have that

$$|\eta| \leq \left|\frac{\int_R q(x) \left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2 (\xi_2(x)^2 + 2\xi_2(x))dx}{\int_R q(x) \left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2}\right|$$
$$\leq 4\frac{\int_R q(x) \left(\frac{1}{2}\frac{\gamma(x)}{q(x)}\right)^2 \left(\left|\frac{\delta(x)}{q(x)}\right|^2 + \left|\frac{\delta(x)}{q(x)}\right|\right) dx}{\int_R q(x) \left(\frac{1}{2}\frac{\gamma(x)}{q(x)}\right)^2}$$
$$\leq C' \left\{4\alpha \int_R q(x) \left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^3 dx + \alpha^2 \int_R q(x) \left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^4 dx\right\}$$

Note that

$$\int_R q(x) \left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^3 dx \leq \left\{\int_R q(x) \left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^3 dx\right\}^{2/3} \left\{\int_R q(x) \left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^4 dx\right\}^{1/3}$$

Thus, we have that $\int_R q(x) \left(\frac{1}{2}\frac{\gamma(x)}{q(x)}\right)^3 dx \leq M$ as well. So,

$$|\eta| \leq (2\alpha + \alpha^2)C'M$$

□

Proposition A.5. *Let $L \geq 1$ be arbitrary. Let $P_l = \int_{a_l}^{b_l} p(x)dx$ and $Q_l = \int_{a_l}^{b_l} q(x)dx$. Define $H_L^R = \sum_{l=1}^L \left(\sqrt{P_l} - \sqrt{Q_l}\right)^2$.*

Define $\delta(x) = q(x) - p(x)$ and let α be a real number such that

$$C' \int_R q(x) \left(\frac{\delta(x)}{q(x)}\right)^2 dx \geq \alpha^2 \geq c' \int_R q(x) \left(\frac{\delta(x)}{q(x)}\right)^2 dx$$

for some constants C', c'. Define $\gamma(x) = \frac{\delta(x)}{\alpha}dx$. Suppose that $\limsup_n \int_R q(x) \left|\frac{1}{2}\frac{\gamma(x)}{q(x)}\right|^4 dx \leq M$. Then, we have that,

$$H_L^R = d_L \alpha^2 (1 + \eta_L)$$

where $d_L = \sum_{l=1}^L P_l \left(\frac{1}{2}\frac{\gamma_l}{P_l}\right)^2 dx$, $\gamma_l = \int_{a_l}^{b_l} \gamma(x)$ and $\sup_L |\eta_L| \leq (2\alpha + \alpha^2)2C'M$. In particular, we have that if $\alpha \to 0$, then $\eta_L \to 0$.

Proof. Let us define $\delta_l = P_l - Q_l$.

$$\begin{aligned}
H_L^R &= \sum_{l=1}^{L} (\sqrt{P_l} - \sqrt{Q_l})^2 \\
&= \sum_{l=1}^{L} Q_l \left(1 - \sqrt{\frac{P_l}{Q_l}}\right)^2 \\
&= \sum_{l=1}^{L} Q_l \left(1 - \sqrt{1 - \frac{\delta_l}{Q_l}}\right)^2 \\
&= \sum_{l=1}^{L} Q_l \left(1 - \left(1 - \frac{1}{2}\frac{\delta_l}{Q_l} - \xi_l\right)\right)^2
\end{aligned}$$

where by convention, we define $\frac{\delta_l}{Q_l} = 0$ when $Q_l, P_l = 0$. Here, $\xi_l = 1 - \frac{1}{2}\frac{\delta_l}{Q_l} - \sqrt{1 - \frac{\delta_l}{Q_l}}$. Continuing,

$$\begin{aligned}
H_L^R &= \sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l} + \xi_l\right)^2 \\
&= \sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l}\right)^2 (1 + \xi_{2l})^2
\end{aligned}$$

where $\xi_{2l} = 0$ if $\frac{\delta_l}{Q_l} = 0$ and $\xi_{2l} = 2\xi_l \frac{Q_l}{\delta_l}$ otherwise.

$$H_L^R = (1 + \eta_L) \sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l}\right)^2$$

where $\eta_L = \frac{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l}\right)^2 (2\xi_{2l} - \xi_{2l}^2)}{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l}\right)^2}$.

By lemma F.3, $|\xi_{2l}| \le 2\left|\frac{\delta_l}{Q_l}\right|$. Therefore,

$$\begin{aligned}
|\eta_L| &= \left| \frac{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l}\right)^2 (2\xi_{2l} - \xi_{2l}^2)}{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\delta_l}{Q_l}\right)^2} \right| \\
&\le \frac{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 (2|\xi_{2l}| + \xi_{2l}^2)}{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2} \\
&\le 4 \frac{\alpha \sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^3 + \alpha^2 \sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^4}{\sum_{l=1}^{L} Q_l \left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2}
\end{aligned}$$

The denominator can be bounded by $1/(2C')$ for large enough L by proposition A.6.

To bound the numerator, we note that for a single l:

$$\int_{a_l}^{b_l} \frac{q(x)}{Q_l} \left|\frac{\gamma(x)}{q(x)}\right|^3 dx \geq \left|\int_{Bin_l} \frac{q(x)}{Q_l}\frac{\gamma(x)}{q(x)} dx\right|^3 = \left|\frac{\gamma_l}{Q_l}\right|^3$$

Therefore,

$$\sum_{l=1}^{L} Q_l \left|\frac{\gamma_l}{Q_l}\right|^3 \leq \int_R q(x)\left|\frac{\gamma(x)}{q(x)}\right|^3 \leq M$$
$$\sum_{l=1}^{L} Q_l \left|\frac{\gamma_l}{Q_l}\right|^4 \leq \int_R q(x)\left|\frac{\gamma(x)}{q(x)}\right|^4 \leq M$$

Thus,

$$|\eta_L| \leq (2\alpha + \alpha^2)2C'M$$

□

For the next proposition, we use the notation Bin_l to denote the bin $[a_l, b_l]$. Let $B_l = b_l - a_l$.

Proposition A.6. *Let* $d = \int_R q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx$ *and* $d_L = \sum_l Q_l \left(\frac{\gamma_l}{Q_l}\right)^2$.
Suppose assumptions A1-A4 hold. Then we have that

$$\lim_{L\to\infty} \sup_n \left|\frac{d}{d_L} - 1\right| = o(1)$$

The proof strategy is to relate both d and d_L to the Riemann sum $d_R = \sum_l B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2$ where x_l is some appropriately chosen element in Bin_l.

Proof. **Step 1:** The first step is to bound $|d_L - d_R|$.

Let $h(x)$ be as defined in Assumption C3; in particular, $|h(x)| \geq \left|\frac{\gamma'(x)}{q(x)}\right| \vee \left|\frac{q'(x)}{q(x)}\right|$.

Let $0 < \tau < 1$. We say that a bin l is good if

$$\sup_{x\in Bin_l} |h(x)| \leq L^\tau$$

the exponent τ will be chosen later to balance two error terms. We will now argue that the proportion of bad bins goes to 0 as $L \to \infty$.

For all large enough L, $\{x : |h(x)| \geq L^\tau\}$ is a union of at most K_h intervals, thus, we have that

$$\begin{aligned}\sum_{l\in\{l\,:\,|h(x)|\geq L^\tau\}} B_l &\leq \mu\left(\{x : |h(x)| \geq L^\tau\}\right) + 2K_h C_{bin} L^{-1}\\ &\leq M' L^{-\tau t} + 2K_h C_{bin} L^{-1}\\ &\leq C_{M',K} L^{-\tau t}\end{aligned}$$

The last inequality follows because $t \leq 1$ and thus $\tau t < 1$ and so the first term dominates. The second inequality follows from Lemma A.7.

We can now bound the number of bad bins:

$$\#\{l \,:\, |h(x)| \geq L^{\tau}\} \leq \frac{C_{M',K}L^{-\tau t}L}{c_{bin}} \leq C_{M',K}L^{1-\tau t}$$

The inequality follows because $1-\tau t > 0$.
For a bad bin l, we can bound $Q_l\left(\frac{\gamma_l}{Q_l}\right)^2$ as follows:

$$\begin{aligned}
Q_l\left(\frac{\gamma_l}{Q_l}\right)^2 &= Q_l\left(\frac{1}{Q_l}\int_{Bin_l}\gamma(x)dx\right)^2 \\
&= Q_l\left(\int_{Bin_l}\frac{\gamma(x)}{q(x)}\frac{q(x)}{Q_l}dx\right)^2 \\
&\leq Q_l\int_{Bin_l}\frac{q(x)}{Q_l}\left(\frac{\gamma(x)}{q(x)}\right)^2 dx \quad \text{by Jensen} \\
&\leq \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&\leq \left(\int_{Bin_l} q(x)\left|\frac{\gamma(x)}{q(x)}\right|^r dx\right)^{2/r}\left(\int_{Bin_l} q(x)dx\right)^{(r-2)/r} \\
&\leq M^{2/r}(CC_{bin})^{(r-2)/r}L^{-(r-2)/r} \\
&\leq C_{M,C}L^{-\frac{2}{r}}
\end{aligned}$$

Now, we have

$$\begin{aligned}
d_L &= \sum_{l=1}^{L} Q_l\left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 \\
&= \sum_{l\,good} Q_l\left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 + \sum_{l\,bad} Q_l\left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 \\
&= \sum_{l\,good} Q_l\left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 + C_{M,C}L^{-\frac{2}{r}}|\{l \,:\, l \text{ bad}\}| \\
&= \sum_{l\,good} Q_l\left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 + C_{M,M',C,K}L^{1-\tau t-\frac{(r-2)}{r}} \\
&= \sum_{l\,good} Q_l\left(\frac{1}{2}\frac{\gamma_l}{Q_l}\right)^2 + C_{M,M',C,K}L^{\frac{2}{r}-\tau t}
\end{aligned}$$

For each good bin l, define $x_l = \arg\max_{x\in Bin_l}|q(x)|$ The argmax is attainable since q is continuous and $q(x_l) < \infty$ since q is bounded.

Now, for a good bin, we have that

$$
\begin{aligned}
Q_l &= \int_{\mathrm{Bin}_l} q(x)dx \\
&= \int_{a_l}^{b_l} q(x)dx \\
&= \int_{a_l}^{b_l} q(x_l) + q'(c_x)(x - x_l)dx \quad \text{for some } c_x \in [a_l, b_l] \\
&= B_l q(x_l) + \int_{a_l}^{b_l} q'(c_x)(x - x_l)dx \\
&= B_l q(x_l) + B_l^2 \xi_l
\end{aligned}
$$

where we define $\xi_l = \frac{1}{B_l^2} \int_{a_l}^{b_l} q'(c_x)(x - x_l)dx$. We can bound $B_l \left| \frac{\xi_l}{q(x_l)} \right|$:

$$
B_l \left| \frac{\xi_l}{q(x_l)} \right| \le \frac{1}{B_l} \int_{a_l}^{b_l} \left| \frac{q'(c_x)}{q(x_l)} \right| |x - x_l| dx \le \frac{1}{B_l} \int_{a_l}^{b_l} \left| \frac{q'(c_x)}{q(c_x)} \right| |x - x_l| dx \le \frac{1}{B_l} \int_{a_l}^{b_l} L^\tau |x - x_l| dx \le \frac{1}{2} C_{bin} L^{\tau - 1}
$$

The second inequality follows because $q(c_x) \le q(x_l)$. The third inequality follows because l is a good bin and thus $\left| \frac{q'(c_x)}{q(c_x)} \right| \le L^\tau$. The last inequality follows because $B_l \le C_{bin} 1/L$.

We perform similar analysis on γ:

$$
\begin{aligned}
\gamma_L &= \int_{Bin_l} \gamma(x)dx \\
&= \int_{a_l}^{b_l} \gamma(x_l) + \gamma'(c_x)(x - x_l)dx \\
&= B_l \gamma(x_l) + B_l^2 \xi'_l
\end{aligned}
$$

where $\xi'_l = \frac{1}{B_l^2} \int_{a_l}^{b_l} \gamma'(c_x)(x - x_l)dx$. It is straightforward to verify that $B_l \left| \frac{\xi'_l}{q(x_l)} \right| \le \frac{1}{2} C_{bin} L^{\tau - 1}$.

For any bin l, we also have that

$$
Q_l = \int_{Bin_l} q(x)dx \le C B_l
$$

Now We look at a single $Q_l \left(\frac{\gamma_l}{Q_l} \right)^2$ term for a single good bin l.

$$
\begin{aligned}
Q_l \left(\frac{\gamma_l}{Q_l} \right)^2 &= \frac{\gamma_l^2}{Q_l} \\
&= \frac{(B_l \gamma(x_l) + B_l^2 \xi'_l)^2}{B_l q(x_l) + B_l^2 \xi_l} \\
&= B_l q(x_l) \left(\frac{\gamma(x_l)}{q(x_l)} + B_l \frac{\xi'_l}{q(x_l)} \right)^2 \left(\frac{1}{1 + B_l \frac{\xi_l}{q(x_l)}} \right) \\
&= B_l q(x_l) \left(\frac{\gamma(x_l)}{q(x_l)} + B_l \frac{\xi'_l}{q(x_l)} \right)^2 \left(1 - B_l \frac{\xi_l}{q(x_l)} + \eta_l (B_l \frac{\xi_l}{q(x_l)})^2 \right)
\end{aligned}
$$

To arrive at the last equality, we assume that $L^{\tau - 1} \le \frac{1}{2} C_{bin}$, which is satisfied so long as $L \ge (2/C_{bin})^{\frac{1}{1-\tau}}$. Under this assumption, $\left| B_l \frac{\xi'_l}{q(x_l)} \right| \le \frac{1}{2}$ and thus it is valid to take the Taylor approximation. Here, η_l is some scalar that satisfies $|\eta_l| \le 16$.

$$Q_l\left(\frac{\gamma_l}{Q_l}\right)^2 =$$

$$\left(B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 + B_l q(x_l)\frac{\gamma(x_l)}{q(x_l)}B_l\frac{\xi_l'}{q(x_l)} + B_l q(x_l)\left(B_l\frac{\xi_l'}{q(x_l)}\right)^2\right)\left(1 - B_l\frac{\xi_l}{q(x_l)} + \eta_l(B_l\frac{\xi_l}{q(x_l)})^2\right)$$

We again note that $\left|B_l\frac{\xi_l'}{q(x_l)}\right| \leq \frac{C_{bin}}{2}L^{\tau-1}$ and $\left|B_l\frac{\xi_l}{q(x_l)}\right| \leq \frac{C_{bin}}{2}L^{\tau-1}$. Suppose $L \geq \left(\frac{1}{2C_{bin}}\right)^{1-\tau}$ so that $\frac{C_{bin}}{2}L^{\tau-1} \leq \frac{1}{4}$, then

$$\left|B_l\frac{\xi_l}{q(x_l)}\right| + \left|\eta_l(B_l\frac{\xi_l}{q(x_l)})^2\right| \leq L^{\tau-1}$$
$$\left|1 - B_l\frac{\xi_l}{q(x_l)} + \eta_l(B_l\frac{\xi_l}{q(x_l)})^2\right| \leq 2$$

Now, we can bound

$$\left|Q_l\left(\frac{\gamma_l}{Q_l}\right)^2 - B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right|$$
$$\leq B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 L^{\tau-1} + 2B_l q(x_l)\frac{\gamma(x_l)}{q(x_l)}L^{\tau-1} + 2B_l q(x_l)L^{2(\tau-1)}$$

The third term is bounded by $2B_l CL^{2(\tau-1)} \leq 2C_{bin}CL^{2\tau-3}$. To bound the second term, we perform case analysis.

Case 1: $\left|\frac{\gamma(x_l)}{q(x_l)}\right| \geq 1$. In this case, $q(x)\left|\frac{\gamma(x_l)}{q(x_l)}\right| \leq q(x)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2$.

Case 2: $\left|\frac{\gamma(x_l)}{q(x_l)}\right| \leq 1$. Then, the second term is bounded by $2B_l CL^{\tau-1} \leq 2C_{bin}CL^{\tau-2}$.

In any case, we have that

$$\left|Q_l\left(\frac{\gamma_l}{Q_l}\right)^2 - Bq(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \leq 3B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 L^{\tau-1} + 4C_{bin}CL^{\tau-2}$$

Define $d_R = \sum_{l\,good} B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2$. Then,

$$\begin{aligned}
|d_L - d_R| &= \left|\sum_l Q_l\left(\frac{\gamma_l}{Q_l}\right)^2 - \sum_{l\,good} B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \\
&\leq \sum_{l\,good}\left|Q_l\left(\frac{\gamma_l}{Q_l}\right)^2 - B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| + C_{M,M',K,C}L^{\frac{2}{r}-\tau t} \\
&\leq 3d_R L^{\tau-1} + C_{C,C_{bin}}L^{\tau-2} + C_{M,M',K,C}L^{\frac{2}{r}-\tau t} \\
&\leq 3d_R L^{\tau-1} + C_{C,C_{bin}}L^{\tau-2} + C_{M,M',K,C}L^{\frac{2}{r}-\tau t} \\
&\leq 3d_R L^{\frac{2-rt}{r(1+t)}} + C_{M,M',K,C,C_{bin}}L^{\frac{2-rt}{r(1+t)}} \quad \text{setting } \tau = \frac{2+r}{r(1+t)}
\end{aligned}$$

The τ is chosen to balance $L^{\tau-1}$ and $L^{2/r-\tau t}$.

Since $2 > rt$, we have that $|d_L - d_R| = o(d_R) + o(1)$.

Step 2.
In like fashion, we bound $|d_R - d|$. We use the same definition of good and bad bins as before.

$$\begin{aligned}
d &= \int_R q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&= \sum_{l=1}^{L} \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&= \sum_{l\,good} \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx + \sum_{l\,bad} \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&\leq \sum_{l\,good} \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx + |\{l\,:\,l\,\text{bad}\}| C_{M,C} L^{-\frac{2}{r}} \\
&\leq \sum_{l\,good} \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx + C_{M,M',C,K} L^{\frac{2}{r}-\tau t}
\end{aligned}$$

The bound on the second term–the inequality–follows from the previous analysis. We now focus on the first term. Note that, for all $x \in Bin_l$

$$\begin{aligned}
q(x) &= q(x_l) + q'(c_l)(x - x_l) \\
\gamma(x) &= \gamma(x_l) + \gamma'(c_l')(x - x_l)
\end{aligned}$$

c_l, c_l' are in Bin_l and they are dependent on x; we leave that dependency implicit to make the notations simpler.

$$\begin{aligned}
\int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 &= \int_{Bin_l} \frac{(\gamma(x_l) + \gamma'(c_l')(x - x_l))^2}{q(x_l) + q'(c_l)(x - x_l)} \\
&= \int_{Bin_l} q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)} + \frac{\gamma'(c_l')}{q(x_l)}(x - x_l)\right)^2 \left(\frac{1}{1 + \frac{q'(c_l)}{q(x_l)}(x - x_l)}\right)
\end{aligned}$$

To make the algebraic manipulation more clear, let us use the following shorthand:

$$T_1 = \frac{q'(c_l)}{q(x_l)}(x - x_l) \quad T_2 = \frac{\gamma'(c_l')}{q(x_l)}(x - x_l)$$

We observe that $|x - x_l| \leq B_l$ and that

$$\left|\frac{\gamma'(c_l')}{q(x_l)}\right| \leq \left|\frac{\gamma'(c_l')}{q(c_l')}\right| \leq L^\tau$$

and likewise, $\left|\frac{q'(c_l)}{q(x_l)}\right| \leq L^\tau$. Thus, we have that $|T_1|, |T_2| \leq C_{bin} L^{\tau - 1}$.

Now, suppose $C_{bin} L^{\tau-1} \leq \frac{1}{2}$, which is satisfied if $L \geq (2C_{bin})^{\frac{1}{1-\tau}}$.

$$= \int_{Bin_l} q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)} + T_2\right)^2 \left(\frac{1}{1+T_1}\right) dx$$
$$= \int_{Bin_l} \left(q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 + q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right) T_2 + q(x_l)T_2^2\right) dx(1 - T_1 + \eta T_1^2)$$

where η is some function of x that satisfies $|\eta| \leq 16$. Thus, we have that

$$\left| \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx - B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 \right|$$
$$\leq B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 L^{\tau-1} + B_l q(x_l)\frac{\gamma(x_l)}{q(x_l)} L^{\tau-1} + B_l q(x_l) L^{2(\tau-1)}$$

Since $q(x_l) \leq C$, the $q(x_l)L^{2(\tau-1)}$ term is bounded by $CL^{2(\tau-1)}$. To bound the $q(x_l)\frac{\gamma(x_l)}{q(x_l)}L^{\tau-1}$ term, we perform case analysis.

Case 1: if $\frac{\gamma(x_l)}{q(x_l)} \geq 1$. Then $q(x_l)\frac{\gamma(x_l)}{q(x_l)}L^{\tau-1} \leq q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 L^{\tau-1}$.

Case 2: if $\frac{\gamma(x_l)}{q(x_l)} \leq 1$. Then $q(x_l)\frac{\gamma(x_l)}{q(x_l)}L^{\tau-1} \leq CL^{\tau-1}$.

Thus, in any case, we have that, for a single good bin:

$$\left| \int_{Bin_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx - B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 \right|$$
$$\leq B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 L^{\tau-1} + B_l C L^{\tau-1}$$

Therefore, we have that

$$|d - d_R| \leq d_R L^{\tau-1} + CL^{\tau-1} + C_{M,M',C,K} L^{\frac{2}{r}-\frac{\tau}{t}}$$

So that $|d - d_R| = o(d_R) + o(1)$.

Since $d = 1$, we have that $|d - d_L| = o(1)$ and that $\left|\frac{d_L}{d} - 1\right| = o(1)$.

□

Proposition A.7. *Suppose assumptions A1-4 hold. Let $n \to \infty$, then, we have that, for any sequence $L_n \to \infty, \alpha_n \to 0$,*

$$\lim_{n\to\infty} \left| \frac{\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2}{\int_R (\sqrt{p(x)} - \sqrt{q(x)})^2 dx} - 1 \right| \to 0$$

Proof. By proposition A.4 and proposition A.5, we have that, for all α,

$$
\begin{aligned}
|H_L^R - H^R| &\leq \left|d_L\alpha^2(1+\eta_L) - d\alpha^2(1+\eta)\right| \\
&\leq d\alpha^2 \left|\frac{d_L}{d}(1+\eta_L) - (1+\eta)\right| \\
&\leq H^R \left|\frac{d_L}{d}\frac{(1+\eta_L)}{(1+\eta)} - 1\right| \\
&\Rightarrow \\
\left|\frac{H_L^R}{H^R} - 1\right| &\leq \left|\frac{d_L}{d}\frac{(1+\eta_L)}{1+\eta} - 1\right|
\end{aligned}
$$

where $|\eta|, |\eta_L| \leq 2(2\alpha+\alpha^2)C'M$ for all L. Thus, it is clear that

$$\lim_{\alpha_n \to 0} \sup_L \left|\frac{1+\eta_L}{1+\eta} - 1\right| = 0$$

Furthermore, it has been shown that

$$\lim_{L_n \to \infty} \sup_\alpha \left|\frac{d_L}{d} - 1\right| = 0$$

Let $\epsilon > 0$ be arbitrarily fixed. Choose α such that $\sup_L \left|\frac{1+\eta_L}{1+\eta} - 1\right| \leq \epsilon/3$ and choose L such that $\sup_\alpha \left|\frac{d_L}{d} - 1\right| \leq \epsilon/3$.

Choose n_0 such that $\alpha_n < \alpha$ and $L_n < L$ for all $n > n_0$.

Then, we have that, for all $n > n_0$:

$$
\begin{aligned}
&\left|\frac{d_L}{d}\frac{1+\eta_L}{1+\eta} - 1\right| \\
&= \left|\frac{d_L}{d}\frac{1+\eta_L}{1+\eta} - \frac{d_L}{d} + \frac{d_L}{d} - 1\right| \\
&= \left|\frac{d_L}{d}\left(\frac{1+\eta_L}{1+\eta} - 1\right) + \left(\frac{d_L}{d} - 1\right)\right| \\
&= \left|\left(\frac{1+\eta_L}{1+\eta} - 1\right) + \left(\frac{d_L}{d} - 1\right)\left(\frac{1+\eta_L}{1+\eta} - 1\right) + \left(\frac{d_L}{d} - 1\right)\right| \\
&\leq \left|\frac{1+\eta_L}{1+\eta} - 1\right| + \left|\left(\frac{d_L}{d} - 1\right)\left(\frac{1+\eta_L}{1+\eta} - 1\right)\right| + \left|\frac{d_L}{d} - 1\right| \\
&\leq \epsilon
\end{aligned}
$$

The claim thus follows.

□

A.2.2 Going from R to $[0,1]$

Proposition A.8. *Suppose assumptions C2 and C5 are satisfied. Suppose also that* $L \leq \frac{2}{H}$.

Define $P_l = \int_{a_l}^{b_l} p(x)dx$ *and* $Q_l = \int_{a_l}^{b_l} q(x)dx$. *Then, we have that* $\frac{1}{2\rho} \leq \frac{P_l}{Q_l} \leq 2\rho$ *for all* l.

Proof. Let Bin_l denote the bin $[a_l, b_l]$. Let us consider an l such that $Bin_l \cap R^c = \emptyset$. Then, for all $x \in Bin_l$, we have that $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$.

Therefore, $\frac{P_l}{Q_l} \leq \int \frac{p(x)}{q(x)}\frac{q(x)}{Q_l}dx \leq 2\rho$. We can upper bound $\frac{Q_l}{P_l}$ similarly.

Now suppose $Bin_l \cap R^c \neq \emptyset$. By the fact that R is an interval and that $\mu\{R^c\} = o(H) \leq \frac{1}{2L}$, only the bins $[0, \frac{1}{L}]$ and $[1 - \frac{1}{L}, 1]$ can potentially satisfy $Bin_l \cap R^c \neq 0$.

Define $P'_l = \int_{Bin_l \cap R} p(x)dx$ and $Q'_l = \int_{Bin_l \cap R} q(x)dx$. Define $P''_l = \int_{Bin_l \cap R^c} p(x)dx$ and $Q''_l = \int_{Bin_l \cap R^c} q(x)dx$.

$$P'_l \geq \min_{x \in Bin_l \cap R} p(x)\frac{1}{2L} \geq \max_{x \in Bin_l \cap R^c} p(x)\frac{1}{2L} \geq P''_l$$

where the first inequality follows because $\mu(R^c) \leq \frac{1}{2L}$ and the second inequality follows from the first derivative conditions. Similarly, we can derive that $Q'_l \geq Q''_l$.

We also notice that $\frac{1}{\rho} \leq \frac{P'_l}{Q'_l} \leq \rho$ by the above discussion.

Thus,

$$\frac{P_l}{Q_l} \leq \frac{2P'_l}{Q'_l} \leq 2\rho$$
$$\frac{P_l}{Q_l} \geq \frac{P'_l}{2Q'_l} \geq \frac{1}{2\rho}$$

□

Proposition A.9. *Suppose assumptions A1-4 hold. Then we have that*

$$\left| \frac{\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2}{\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx} - 1 \right| \to 0$$

Proof. Let $H = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$.

Let a_L be an $o(1)$ sequence such that $\mu(R^c) \leq a_L H$.

We divide the set of bins into three sets L_1, L_2, L_3.

$$\begin{aligned}
L_1 &= \{l \,:\, Bin_l \cap R^c = \emptyset\} \\
L_2 &= \{l \,:\, Bin_l \cap R^c \neq \emptyset,\, P_l \vee Q_l \geq 2Ca_L H\} \\
L_3 &= \{l \,:\, Bin_l \cap R^c \neq \emptyset,\, P_l \vee Q_l \leq 2Ca_L H\}
\end{aligned}$$

For each bin l, define $P'_l = \int_{Bin_l \cap R} p(x)dx$ and $P''_l = \int_{Bin_l \cap R^c} p(x)dx$. Likewise for Q'_l and Q''_l.

We now proceed in two steps:

Step 1: We first claim that for all $l \in L_2$,

$$\left| (\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P'_l} - \sqrt{Q'_l})^2 \right| \leq a_L H$$

Since $\mu(R^c) \leq a_L H$, we have that $P''_l = \int_{Bin_l \cap R^c} p(x)dx \leq Ca_L H$ and likewise for Q''_l.

$$\begin{aligned}
(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P'_l} - \sqrt{Q'_l})^2 &= P_l + Q_l - P'_l - Q'_l - 2\sqrt{P_l Q_l} + 2\sqrt{P'_l Q'_l} \\
&\leq P''_l + Q''_l - 2\sqrt{P''_l Q''_l} \\
&\leq P''_l + Q''_l \\
&\leq 2Ca_L H
\end{aligned}$$

The first inequality follows because of the following reason. First, by AM-GM inequality, we have that $2\sqrt{P'_l Q'_l P''_l Q''_l} \leq P'_l Q''_l + P''_l Q'_l$. Thus:

$$\begin{aligned}
&P_l'Q_l' + P_l''Q_l'' + 2\sqrt{P_l'Q_l'P_l''Q_l''} \leq (P_l' + P_l'')(Q_l' + Q_l'') \\
(\Rightarrow)&\sqrt{P_l'Q_l'} + \sqrt{P_l''Q_l''} \leq \sqrt{(P_l' + P_l'')(Q_l' + Q_l'')} \\
(\Rightarrow)&\sqrt{P_l''Q_l''} \leq \sqrt{P_lQ_l} - \sqrt{P_l'Q_l'}
\end{aligned}$$

On the other hand, we have that

$$\begin{aligned}
\sqrt{P_lQ_l} - \sqrt{P_l'Q_l'} &= \frac{(\sqrt{P_lQ_l} - \sqrt{P_l'Q_l'})(\sqrt{P_lQ_l} + \sqrt{P_l'Q_l'})}{\sqrt{P_lQ_l} + \sqrt{P_l'Q_l'}} \\
&= \frac{P_lQ_l - P_l'Q_l'}{\sqrt{P_lQ_l} + \sqrt{P_l'Q_l'}} \\
&= \frac{P_l'Q_l'' + P_l''Q_l' + P_l''Q_l''}{\sqrt{P_lQ_l} + \sqrt{P_l'Q_l'}} \\
&\leq \frac{P_l'Q_l'' + P_l''Q_l' + P_l''Q_l''}{2\sqrt{P_l'Q_l'}} \\
&\leq Q_l''\frac{P_l'}{2\sqrt{P_l'Q_l'}} + P_l''\frac{Q_l'}{2\sqrt{P_l'Q_l'}} + Q_l''\frac{P_l''}{2\sqrt{P_l'Q_l'}}
\end{aligned}$$

Note that because P_l' and Q_l' are defined on R, we have that

$$\begin{aligned}
\left|\frac{P_l'}{Q_l'}\right| &= \left|\int_{Bin_l \cap R} \frac{p(x)}{Q_l'} dx\right| \\
&\leq \int_{Bin_l \cap R} \left|\frac{p(x)}{q(x)}\right| \frac{q(x)}{Q_l'} dx \\
&\leq \rho
\end{aligned}$$

Thus, $\sqrt{\frac{P_l'}{Q_l'}} \vee \sqrt{\frac{Q_l'}{P_l'}} \leq \sqrt{\rho}$. This bounds the terms $Q_l''\frac{P_l'}{2\sqrt{P_l'Q_l'}} + P_l''\frac{Q_l'}{2\sqrt{P_l'Q_l'}} \leq \sqrt{\rho}(Q_l'' + P_l'')$. We still need to bound the last term $\frac{Q_l''P_l''}{2\sqrt{P_l'Q_l'}}$.

Since $l \in L_2$, we have that either $P_l \geq 2Ca_LH$ or that $Q_l \geq 2Ca_LH$. Let us suppose the former; the latter case can be handled in an identical manner.

Since $P_l'' \leq Ca_LH$ and $P_l \geq 2Ca_LH$, we have that $P_l'' \leq P_l'$ and thus, $\frac{Q_l''P_l''}{2\sqrt{P_l'Q_l'}} \leq Q_l''\frac{P_l'}{2\sqrt{P_l'Q_l'}} \leq \sqrt{\rho}Q_l''$.

Putting all this together, we have that

$$\sqrt{P_lQ_l} - \sqrt{P_l'Q_l'} \leq 2\sqrt{\rho}(Q_l'' + P_l'')$$

Thus,

$$\begin{aligned}
(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2 &= P_l + Q_l - P_l' - Q_l' - 2\sqrt{P_lQ_l} + 2\sqrt{P_l'Q_l'} \\
&\geq P_l'' + Q_l'' - \sqrt{\rho}(Q_l'' + P_l'') \\
&\geq (1 - \sqrt{\rho})(P_l'' + Q_l'') \\
&\geq -(\sqrt{\rho} + 1)Ca_LH
\end{aligned}$$

Combining these two bounds, we have

$$\left|(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2\right| \leq C_{C,\rho}a_LH$$

This finishes step 1.

Step 2: In step 2, we verify that $\{Bin_l\}_{l \in L_1} \cup \{Bin_l \cap R\}_{l \in L_2} \cup \{Bin_l \cap R\}_{l \in L_3}$ constitute a valid approximately uniform binning of R.

First, because R is an interval, it is easy to see that $Bin_l \cap R$ is an interval as well. Secondly, $|Bin_l \cap R^c| \leq \mu\{R^c\} \leq a_L H$. Since $\frac{1}{H} \leq L$ by assumption, we have that $\mu\{R^c\} \leq a_L \frac{1}{L}$ and so, there exists constants c_{bin} such that $\frac{c_{bin}}{L} \leq |Bin_l \cap R| \leq \frac{1}{L}$.

Step 3:

$$\begin{aligned}
&\left|\sum_{l=1}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| \\
&= \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_3}(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| \\
&\leq \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + 8Ca_L H
\end{aligned}$$

The inequality follows because $P_l \vee Q_l \leq 2Ca_L H$ for all $l \in L_3$ and because $|L_3| \leq 2$. Continuing on, we have:

$$\leq \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P'_l} - \sqrt{Q'_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + C_{C,\rho} a_L H$$

This is because of our bound in step 1.

$$\begin{aligned}
&\leq \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P'_l} - \sqrt{Q'_l})^2 + \sum_{l \in L_3}(\sqrt{P'_l} - \sqrt{Q'_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + C_{C,\rho} a_L H \\
&\leq \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P'_l} - \sqrt{Q'_l})^2 + \sum_{l \in L_3}(\sqrt{P'_l} - \sqrt{Q'_l})^2 - \int_R(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + C_{C,\rho} a_L H \\
&\leq C_{C,\rho} a_L H
\end{aligned}$$

The first inequality follows because $P'_l \leq P_l$ and thus $\sum_{l \in L_3}(\sqrt{P'_l} - \sqrt{Q'_l})^2 \leq 2Ca_L H$ as well. The second inequality follows because $\int_{R^c}(\sqrt{p(x)} - \sqrt{q(x)})^2 \leq C\mu\{R^c\} = Ca_L H$

The last inequality follows by proposition A.7.

Since $a_L \to 0$, the conclusion follows.

□

A.2.3 Lemmas

Lemma A.6. *Let a, b be positive scalars such that $\log \frac{a}{b}$ is bounded away from $-\infty$ and ∞. Suppose that $\sum_l(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = o\left(\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right)$.*

Then, we have that

$$\sum_l(\sqrt{aP_l} - \sqrt{bQ_l})^2 - \int(\sqrt{ap(x)} - \sqrt{bq(x)})^2 dx = o\left(\int(\sqrt{ap(x)} - \sqrt{bq(x)})^2 dx\right)$$

Proof. First, note that

$$\sum_l (\sqrt{aP_l} - \sqrt{bQ_l})^2 = b\sum_l (\sqrt{\frac{a}{b}P_l} - \sqrt{Q_l})^2$$
$$\int (\sqrt{ap(x)} - \sqrt{bq(x)})^2 dx = b\int (\sqrt{\frac{a}{b}p(x)} - \sqrt{q(x)})^2 dx$$

Therefore, we can, without loss of generality, rename $a \leftarrow \frac{a}{b}$, assume that a is bounded away from 0 and ∞, and prove only that

$$\sum_l (\sqrt{aP_l} - \sqrt{Q_l})^2 - \int (\sqrt{ap(x)} - \sqrt{q(x)})^2 dx = o\left(\int (\sqrt{ap(x)} - \sqrt{q(x)})^2 dx\right)$$

To show this, we use the following identity:

$$\begin{aligned}&(\sqrt{aP_l} - \sqrt{Q_l})^2 - (\sqrt{P_l} - \sqrt{Q_l})^2\\ &= aP_l + Q_l - 2\sqrt{aP_lQ_l} - P_l - Q_l + 2\sqrt{P_lQ_l}\\ &= (a-1)P_l - (\sqrt{a}-1)2\sqrt{P_lQ_l}\end{aligned}$$

Therefore,

$$\begin{aligned}&\sum_l (\sqrt{aP_l} - \sqrt{Q_l})^2 - \sum_l (\sqrt{P_l} - \sqrt{Q_l})^2\\ &= (a-1) - (\sqrt{a}-1)\sum_l 2\sqrt{P_lQ_l}\\ &= (a-1) - (\sqrt{a}-1)\left(2 - \sum_l (\sqrt{P_l} - \sqrt{Q_l})^2\right)\\ &= (a-1) - 2(\sqrt{a}-1) + (\sqrt{a}-1)\left\{\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2\right\}\end{aligned}$$

where we used the fact that $\sum_l 2\sqrt{P_lQ_l} = 2 - \sum_l (\sqrt{P_l} - \sqrt{Q_l})^2$.

Therefore,

$$\sum_l (\sqrt{aP_l} - \sqrt{Q_l})^2 = (a-1) - 2(\sqrt{a}-1) + \sqrt{a}\left\{\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2\right\}$$

By the exact same calculation, we would get

$$\int (\sqrt{ap(x)} - \sqrt{q(x)})^2 dx = (a-1) - 2(\sqrt{a}-1) + \sqrt{a}\left\{\int (\sqrt{p(x)} - \sqrt{q(x)})^2\right\}$$

Now we are ready to prove the proposition.

$$\begin{aligned}\sum_l (\sqrt{aP_l} - \sqrt{Q_l})^2 - \int (\sqrt{ap(x)} - \sqrt{q(x)})^2 dx &= \sqrt{a}\left\{\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 - \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right\}\\ &= o\left(\sqrt{a}\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right)\\ &= o\left((a-1) - 2(\sqrt{a}-1) + \sqrt{a}\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right)\\ &= o\left(\int (\sqrt{ap(x)} - \sqrt{q(x)})^2 dx\right)\end{aligned}$$

where the third inequality follows because a is bounded away from 0 and because $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \leq 2$.

□

A.3 Proof of Proposition 6.3

Proposition A.10. *Suppose $p(x), q(x)$ are supported on $[0,1]$.*

C1' Suppose $p(x), q(x) \leq C$ on $[0,1]$ and are absolutely continuous.

C2' There exists R a subinterval of $[0,1]$ such that $\exp(-L^{1/r}) \leq \frac{p(x)}{q(x)} \leq \exp(L^{1/r})$ and $\mu\{R^c\} \leq \frac{1}{2L}$.

C3' Let $h(x) \geq \sup_n \max\left\{\left|\frac{p'(x)}{p(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|\right\}$. Suppose $\int |h(x)|^t dx \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{x : |h(x)| \geq \kappa\}$ is a union of at most K intervals for all large enough κ.

C4' $p'(x), q'(x) \geq 0$ for all $x < \frac{1}{L}$ and $p'(x), q'(x) \leq 0$ for all $x > 1 - \frac{1}{L}$.

Suppose $\frac{1}{c_0} \leq \frac{1-P_0}{1-Q_0} \leq c_0$. Let $Bin_l = [a_l, b_l]$ for $l = 1, ..., L$ be a uniformly spaced binning of the interval $[0,1]$ and let $P_l = (1-P_0)\int_{a_l}^{b_l} p(x)dx$ and $Q_l = (1-Q_0)\int_{a_l}^{b_l} q(x)dx$.

Define $I = -2\log\left(\sqrt{P_0 Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right)$ and $I_L = -2\log\left(\sqrt{P_0Q_0} + \sum_{l=1}^L \sqrt{P_l Q_l}\right)$.

Then, we have that

$$\left|\frac{I - I_L}{I}\right| = o(1)$$

and that $\frac{1}{4\rho_L c_0} \leq \frac{P_l}{Q_l} \leq 4\rho_L c_0$ for all l.

Proof. The proof is identical to that of proposition 6.2 except that we use proposition A.11 instead of proposition A.9.

□

Proposition A.11. *Let assumptions C1, C3 be satisfied. Let $Bin_l = [a_l, b_l]$ be a uniform binning of $[0,1]$ for $l = 1, .., L$ and let $P_l = \int_{Bin_l} p(x)dx$ and $Q_l = \int_{Bin_l} q(x)dx$.*

Then, we have that

$$\left|\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx - \sum_l (\sqrt{P_l} - \sqrt{Q_l})^2\right| \to 0$$

Proof. The proof will be similar to that of Proposition A.6.

First, we observe that $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \int p(x)dx + \int q(x)dx - 2\int \sqrt{p(x)q(x)}dx = 2 - 2\int \sqrt{p(x)q(x)}$. And, that $\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 = \sum_l P_l + \sum_l Q_l - 2\sum_l \sqrt{P_l Q_l}$.

Thus, we need only show that

$$\left|\int \sqrt{p(x)q(x)}dx - \sum_l \sqrt{P_l Q_l}\right| \to 0$$

We have that $|h(x)| \geq \left|\frac{p'(x)}{p(x)}\right| \vee \left|\frac{q'(x)}{q(x)}\right|$.

Let $0 < \tau < 1$. We say that a bin l is good if

$$\sup_{x \in Bin_l} |h(x)| \leq L^\tau$$

the exponent τ will be chosen later to balance two error terms. We will now argue that the proportion of bad bins goes to 0 as $L \to \infty$.

For all large enough L, $\{x : |h(x)| \geq L^\tau\}$ is a union of at most K_h intervals, thus, we have that

$$\begin{aligned}\sum_{l \in \left\{l \,:\, \sup_{x \in Bin_l} |h(x)| \geq L^\tau\right\}} B_l &\leq \mu\left(\left\{x : \sup_{x \in Bin_l} |h(x)| \geq L^\tau\right\}\right) + 2KC_{bin}L^{-1} \\ &\leq M'L^{-\tau t} + 2KC_{bin}L^{-1} \\ &\leq C_{M',K}L^{-\tau t}\end{aligned}$$

The last inequality follows because $t \leq 1$ and thus $\tau t < 1$ and so the first term dominates. The second inequality follows from Lemma A.7.

We can now bound the number of bad bins:

$$\#\{l : |h(x)| \geq L^\tau\} \leq \frac{C_{M',K}L^{-\tau t}L}{c_{bin}} \leq C_{M',K}L^{1-\tau t}$$

For a bad bin, we can bound $P_l, Q_l \leq C_{bin}CL^{-1}$ and $\int_{Bin_l}(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \leq 2L^{-1}CC_{bin}$.

Now we consider a good bin l. Let x_l be $\arg\max_{x \in Bin_l} p(x)$. The argmax is attainable since p is continuous and $p(x_l) < \infty$ since p is bounded.

$$\begin{aligned}P_l &= \int_{a_l}^{b_l} p(x)dx \\ &= \int_{a_l}^{b_l} p(x_l) + p'(c_x)(x - x_l)dx \\ &= B_l p(x_l) + B_l^2 \xi_l\end{aligned}$$

where $\xi_l = \frac{1}{B_l^2}\int_{a_l}^{b_l} p'(c_x)(x - x_l)dx$. We can bound $B_l \left|\frac{\xi_l}{p(x_l)}\right|$:

$$B_l \left|\frac{\xi_l}{p(x_l)}\right| \leq \frac{1}{B_l}\int_{a_l}^{b_l}\left|\frac{p'(c_x)}{p(x_l)}\right||x{-}x_l|dx \leq \frac{1}{B_l}\int_{a_l}^{b_l}\left|\frac{p'(c_x)}{p(c_x)}\right||x{-}x_l|dx \leq \frac{1}{B_l}\int_{a_l}^{b_l} L^\tau|x{-}x_l|dx \leq \frac{1}{2}C_{bin}L^{\tau-1}$$

Likewise, define $x'_l = \arg\max_{x \in Bin_l} q(x)$. We have that

$$Q_l = B_l q(x'_l) + B_l^2 \xi'_l$$

where $\xi'_l = \frac{1}{B_l}\int_{a_l}^{b_l} q'(c_x)(x - x'_l)dx$. We can also bound

$$B_l \left|\frac{\xi'_l}{q(x'_l)}\right| \leq \frac{1}{2}C_{bin}L^{\tau-1}$$

Thus, we have that

$$
\begin{aligned}
\sqrt{P_l Q_l} &= \sqrt{(B_l p(x_l) + B_l^2 \xi_l)(B_l q(x_l') + B_l^2 \xi_l')} \\
&= \sqrt{p(x_l) q(x_l')} \sqrt{(B_l + B_l^2 \frac{\xi_l}{p(x_l)})(B_l + B_l^2 \frac{\xi_l'}{q(x_l')})} \\
&= \sqrt{p(x_l) q(x_l')} B_l \sqrt{(1 + B_l \frac{\xi_l}{p(x_l)})(1 + B_l \frac{\xi_l'}{q(x_l')})}
\end{aligned}
$$

By our bounds on $B_l \frac{\xi_l}{p(x_l)}$ and $B_l \frac{\xi_l'}{q(x_l')}$, we can bound the nuissance term

$$
\begin{aligned}
\sqrt{(1 + B_l \frac{\xi_l}{p(x_l)})(1 + B_l \frac{\xi_l'}{q(x_l')})} &\leq \sqrt{1 + C_{bin} L^{\tau-1}(1 + o(1))} \\
&\leq 1 + \frac{1}{2} C_{bin} L^{\tau-1}(1 + o(1))
\end{aligned}
$$

It is clear that $B_l \sqrt{p(x_l) q(x_l')} \leq B_l C$. Therefore, we have that

$$
\left| \sqrt{P_l Q_l} - \sqrt{p(x_l) q(x_l')} B_l \right| \leq B_l C C_{bin} L^{\tau-1}(1 + o(1)) \tag{A.8}
$$

Likewise, we have that

$$
\begin{aligned}
\int_{a_l}^{b_l} \sqrt{p(x) q(x)} dx &= \int_{a_l}^{b_l} \sqrt{p(x) q(x)} dx \\
&= \int_{a_l}^{b_l} \sqrt{(p(x_l) + p'(c_x)(x - x_l))(q(x_l') + q'(c_x')(x - x_l'))} dx \\
&= \int_{a_l}^{b_l} \sqrt{p(x_l) q(x_l')} \left(\sqrt{1 + (x - x_l) \frac{p'(c_x)}{p(x_l)} + (x - x_l') \frac{q'(c_x')}{q(x_l')} + (x - x_l)(x - x_l') \frac{p'(c_x)}{p(x_l)} \frac{q'(c_x')}{q(x_l')}} \right) dx
\end{aligned}
$$

We have that

$$
\begin{aligned}
\left| (x - x_l) \frac{p'(c_x)}{p(x_l)} \right| &\leq B_l \left| \frac{p'(c_x)}{p(c_x)} \right| \\
&\leq C_{bin} L^{\tau-1}
\end{aligned}
$$

$$
\begin{aligned}
\left| (x - x_l) \frac{q'(c_x')}{q(x_l)} \right| &\leq B_l \left| \frac{q'(c_x')}{q(c_x')} \right| \\
&\leq C_{bin} L^{\tau-1}
\end{aligned}
$$

Therefore, we can bound the nuissance term:

$$
\begin{aligned}
\sqrt{1 + (x - x_l) \frac{p'(c_x)}{p(x_l)} + (x - x_l') \frac{q'(c_x')}{q(x_l')} + (x - x_l)(x - x_l') \frac{p'(c_x)}{p(x_l)} \frac{q'(c_x')}{q(x_l')}} &\leq \sqrt{1 + C_{bin} L^{\tau-1}(1 + o(1))} \\
&\leq 1 + \frac{1}{2} C_{bin} L^{\tau-1}(1 + o(1))
\end{aligned}
$$

The term $B_l\sqrt{p(x_l)q(x'_l)}$ is bounded by B_lC. Hence, we have

$$\left|\int_{a_l}^{b_l}\sqrt{p(x)q(x)}dx - B_l\sqrt{p(x_l)q(x'_l)}\right| \leq B_lCC_{bin}L^{\tau-1} \tag{A.9}$$

By combining inequalities (A.8) and (A.9), we have that

$$\left|\sqrt{P_lQ_l} - \int_{a_l}^{b_l}\sqrt{p(x)q(x)}dx\right| \leq B_lCC_{bin}L^{\tau-1}$$

We can then complete the proof:

$$\begin{aligned}\left|\sum_l\sqrt{P_lQ_l} - \int\sqrt{p(x)q(x)}dx\right| &\leq \sum_{l:\ l\text{ bad}}B_lC + \sum_{l:\ l\text{ good}}\left|\sqrt{P_lQ_l} - \int_{a_l}^{b_l}\sqrt{p(x)q(x)}dx\right| \\ &\leq C_{M',K}L^{-\tau t} + \sum_{l:\ l\text{ good}}B_lCC_{bin}L^{\tau-1} \\ &\leq C_{M',K}L^{-\tau t} + CC_{bin}L^{\tau-1}\end{aligned}$$

By setting $\tau = \frac{1}{1+t}$, we get that

$$\left|\sum_l\sqrt{P_lQ_l} - \int\sqrt{p(x)q(x)}dx\right| \to 0$$

□

Proposition A.12. *Suppose assumption C2 holds. Define* $P_l = \int_{a_l}^{b_l}p(x)dx$ *and* $Q_l = \int_{a_l}^{b_l}q(x)dx$. *Then, we have that, for all* l*:*

$$\frac{1}{4\rho_L} \leq \frac{P_l}{Q_l} \leq 4\rho_L$$

Proof. The proof is identical to that of proposition A.8.

□

Lemma A.7. *Let* $h : [0,1] \to \mathbb{R}$ *be a measurable function. Let* $t > 0$, *suppose* $\int|h(x)|^t dx \leq M'$, *then, we have that* $\mu\{x : |h(x)| \geq \kappa\} \leq \frac{M'}{\kappa^t}$ *for any* $\kappa > 0$.

Proof. By definition of Lebesgue integral:

$$\kappa\mu\{x : |h(x)|^t \geq \kappa\} \leq \int|h(x)|^t dx \leq M'$$

Thus,

$$\begin{aligned}\mu\{x : |h(x)|^t \geq \kappa\} &\leq \frac{M'}{\kappa} \\ (\Leftrightarrow) \quad \mu\{x : |h(x)| \geq \kappa^{1/t}\} &\leq \frac{M'}{\kappa}\end{aligned}$$

A change of variable completes the proof.

□

B Proofs of Theorems 4.1 and Theorem 4.2

We now outline the proofs of Theorems 4.1 and 4.2, with proofs of supporting propositions in the succeeding subsections.

B.1 Main argument: Proof of Theorem 4.1

First, we claim that the divergence I and H between $p(x), q(x)$ does not change after we apply the transformation Φ. To see this, note that the transformed density $p_\Phi(z)$ and $q_\Phi(z)$, now supported over $[0, 1]$, have the following form:

$$p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))} \qquad q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$$

Therefore, we have, by a change of variable $z = \Phi^{-1}(x)$, that

$$\int_{\mathbb{R}} \sqrt{p(x)q(x)}dx = \int_0^1 \sqrt{p_\Phi(z)q_\Phi(z)}dz$$
$$\int_{\mathbb{R}} (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \int_0^1 (\sqrt{p_\Phi(z)} - \sqrt{q_\Phi(z)})^2 dz$$

Under A1-A5, Proposition B.1 shows that conditions C1-C5 are also satisfied. Also, under our assumption that $L = o(\frac{1}{H})$, it must be that $L \leq \frac{2}{H}$ for large enough L. Therefore, Proposition 6.2 applies and we can conclude that after transformation and discretization, the label probabilities P_l, Q_l's satisfy

$$\frac{1}{2c_0\rho} \leq \frac{P_l}{Q_l} \leq 2c_0\rho$$

for all l.

Under our assumption that $L = o(nI)$ and the conclusion from Proposition 6.2 that $I_L = I(1 + o(1))$, we know that $L = o(nI_L)$ as well and thus, we can use Proposition 6.1 (with $\rho_L = 2c_0\rho$) to get that

$$\lim_{n\to\infty} P\left(l(\widehat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)\right) \to 1$$

The theorem then follows from the fact that $I_L = I(1 + o(1))$.

B.2 Main argument: Proof of Theorem 4.2

The proof mirrors that of Theorem 4.2. First we note again that the divergence I and H does not change after we transform the densities $p(x), q(x)$ into $p_\Phi(z)$ and $q_\Phi(z)$ with Φ.

Proposition B.2 then shows that assumptions A1'-A4' implies C1'-C4'. Therefore, Proposition 6.2 applies and we can conclude that after transformation and discretization, the label probabilities P_l, Q_l's satisfy

$$\frac{1}{2c_0 \exp(L^{1/r})} \leq \frac{P_l}{Q_l} \leq 2c_0 \exp(L^{1/r})$$

for all l and that $I_L = I(1 + o(1))$.

Therefore, we can again use Proposition 6.1 (with $\rho_L = 2c_0 \exp(L^{1/r})$) to conclude that

$$\lim_{n\to\infty} P\left(l(\widehat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)\right) \to 1$$

The theorem follows from the fact that $I_L = I(1 + o(1))$.

B.3 Transformation Analysis

Proposition B.1. *Let $p(x), q(x)$ be densities over $\mathbb{R}$ and let $\Phi : \mathbb{R} \to [0,1]$ be a CDF. Suppose $p(x), q(x), \Phi$ satisfy the following conditions:*

A1 Suppose $p(x), q(x) \leq C$ are absolutely continuous. $\lim_{|x|\to\infty} \sup_n \frac{p(x) \vee q(x)}{\phi(x)} < \infty$

A2 There exists R a subinterval of $\mathbb{R}$ such that $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$ and $\Phi\{R^c\} = o(H)$.

A3 Define $\alpha^2 = \int_R q(x) \left(\frac{p(x)-q(x)}{q(x)}\right)^2 dx$ and $\gamma(x) = \frac{q(x)-p(x)}{\alpha}$. Suppose $\int_R q(x) \left|\frac{\gamma(x)}{q(x)}\right|^r dx \leq M$ for constants $M, r \geq 4$.

A4 Let $h(x) \geq \sup_n \max\left\{\left|\frac{\gamma'(x)}{q(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{\phi'(x)}{\phi(x)}\right|\right\}$. Let $\int_R |h(x)|^{4t/(1-t)} \phi(x) dx \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{x : |h(x)| \geq \kappa\}$ is a union of at most K_h intervals for all large enough κ. Suppose $\int \phi(x)^{\frac{1-t}{1+t}} dx < \infty$.

A5 $(\log p)'(x), (\log q)'(x) \geq (\log \phi)'(x) \geq 0$ for all $x < -c'$ and $(\log p)'(x), (\log q)'(x) \leq (\log \phi)'(x) \leq 0$ for all $x > c'$ for a constant $c' > 0$.

Now we let $p(z), q(z)$ be the Φ-transformed densities over $[0,1]$. Then, we have that the following conditions are satisfied for $p(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$:

C1 Suppose $p(z), q(z) \leq C$ on $[0,1]$ and are absolutely continuous.

C2 There exists R a subinterval of $[0,1]$ such that $\frac{1}{\rho} \leq \left|\frac{p(z)}{q(z)}\right| \leq \rho$ and $\mu\{R^c\} = o(H)$ where μ is the Lebesgue measure.

C3 Define $\alpha^2 = \int_R \frac{(p(z)-q(z))^2}{q(z)} dz$ and $\gamma(z) = \frac{q(z)-p(z)}{\alpha}$. Suppose $\int_R q(z) \left|\frac{\gamma(z)}{q(z)}\right|^r dz \leq M$ for constants $M, r \geq 4$.

C4 Let $h(z) \geq \sup_n \max\left\{\left|\frac{\gamma'(z)}{q(z)}\right|, \left|\frac{q'(z)}{q(z)}\right|\right\}$. Suppose $\int_R |h(z)|^t dz \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{z : |h(z)| \geq \kappa\}$ is a union of at most K intervals for all large enough κ.

C5 For all large enough L, we have that for all $z \leq \frac{1}{L}$, $p'(z), q'(z) \geq 0$ and for all $z \geq \Phi(1 - \frac{1}{L})$, we have that $p'(z), q'(z) \leq 0$.

Proof. The first and second claim directly follow from A1 and A2 respectively. The third claim follows from A3 with a change of variable. Therefore, we need only prove the fourth and fifth claim.

Note that

$$p'(z) = \frac{p'(\Phi^{-1}(z)) - p(\Phi^{-1}(z)) \frac{\phi'(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}}{\phi^2(\Phi^{-1}(z))}$$

Therefore, $p'(z) \geq 0$ if and only if $p'(x) \geq p(x)\frac{\phi'(x)}{\phi(x)}$. Likewise for $q'(z)$. The fifth claim follows. For the fourth claim, note that

$$\frac{q'(z)}{q(z)} = \frac{q'(\Phi^{-1}(z))}{q(\Phi^{-1}(z))} \frac{1}{\phi(\Phi^{-1}(z))} - \frac{\phi'(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))} \frac{1}{\phi(\Phi^{-1}(z))}$$

By a change of variables, we have that

$$\int_R \left|\frac{q'(z)}{q(z)}\right|^t dz = \int_R \left|\frac{q'(x)}{q(x)}\frac{1}{\phi(x)} - \frac{\phi'(x)}{\phi(x)}\frac{1}{\phi(x)}\right|^t \phi(x)dx$$
$$\leq \int_R \left|\frac{q'(x)}{q(x)}\frac{1}{\phi(x)}\right|^t \phi(x)dx + \int_R \left|\frac{\phi'(x)}{\phi(x)}\frac{1}{\phi(x)}\right|^t \phi(x)dx$$

This is finite since

$$\int_R \left|\frac{q'(x)}{q(x)}\frac{1}{\phi(x)}\right|^t \phi(x)dx \leq \left\{\int_R \left|\frac{q'(x)}{q(x)}\right|^{\frac{2t}{1-t}} \phi(x)dx\right\}^{(1-t)/2} \left\{\int_R \phi(x)^{\frac{1-t}{1+t}} dx\right\}^{(1+t)/2} dx$$

Likewise for the second term.

Finally, we also have

$$\int_R \left|\frac{\gamma'(z)}{q(z)}\right|^t dz = \int_R \left|\frac{1}{\alpha}\frac{p'(x) - p(x)\frac{\phi'(x)}{\phi(x)} - q'(x) + q(x)\frac{\phi'(x)}{\phi(x)}}{q(x)\phi(x)}\right|^t \phi(x)dx$$
$$= \int_R \left|\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)} - \frac{1}{\alpha}\frac{p(x) - q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right|^t \left|\frac{1}{\phi(x)}\right|^t \phi(x)dx$$

By Holder's inequality again, we have

$$\leq \left\{\int_R \left|\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)} - \frac{1}{\alpha}\frac{p(x) - q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right|^{2t/(1-t)} \phi(x)dx\right\}^{(1-t)/2} \left\{\int_R \phi(x)^{\frac{1-t}{1+t}} dx\right\}^{(1+t)/2}$$

The latter quantity is finite by assumption. To show that the former quantity is finite, we need only show that

$$\int_R \left|\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)}\right|^{2t/(1-t)} \phi(x)dx$$

is finite, which is known, and that

$$\int_R \frac{1}{\alpha}\left|\frac{p(x) - q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right|^{2t/(1-t)} \phi(x)dx$$

is finite, which follows from an application of Cauchy-Schwartz.

□

Proposition B.2. *Suppose that the following assumptions hold:*

A1' Suppose $p(x), q(x) \leq C$ are absolutely continuous. $\lim_{|x|\to\infty} \sup_n \frac{p(x)\vee q(x)}{\phi(x)} < \infty$

A2' For all large enough $\kappa > 0$, there exists R a subinterval of $\mathbb{R}$ that satisfies $\exp(-\kappa^{1/r}) \leq \frac{p(x)}{q(x)} \leq \exp(\kappa^{1/r})$ and $\Phi\{R^c\} \leq \frac{1}{2\kappa}$, where $r > 2$ is a constant.

A3' Let $h(x) \geq \sup_n \max\left\{\left|\frac{p'(x)}{p(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{\phi'(x)}{\phi(x)}\right|\right\}$. Suppose $\int_R |h(x)|^{2t/(1-t)}\phi(x)dx \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{x : |h(x)| \geq \kappa\}$ is a union of at most K intervals for all large enough κ. Suppose $\int \phi(x)^{\frac{1-t}{1+t}} dx < \infty$.

A4' $(\log p)'(x), (\log q)'(x) \geq (\log\phi)'(x) \geq 0$ for all $x < -c'$ and $(\log p)'(x), (\log q)'(x) \leq (\log\phi)'(x) \leq 0$ for all $x > c'$ for a constant $c' > 0$.

Now we let $p(z), q(z)$ be the Φ-transformed densities over $[0,1]$. Then, we have that the following conditions are satisfied for $p(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$:

Let L be a sequence such that $L \to \infty$.

C1' Suppose $p(z), q(z) \leq C$ on $[0,1]$ and are absolutely continuous.

C2' There exists R a subinterval of $\mathbb{R}$ that satisfies $\exp(-L^{1/r}) \leq \frac{p(x)}{q(x)} \leq \exp(L^{1/r})$ and $\mu\{R^c\} \leq \frac{1}{2\kappa}$, where $r > 2$ is a constant.

C3' Let $h(z) \geq \sup_n \max\left\{\left|\frac{p'(z)}{p(z)}\right|, \left|\frac{q'(z)}{q(z)}\right|\right\}$. Suppose $\int |h(z)|^t dz \leq M'$ for some constant M' and $1 \geq t \geq 2/r$. Suppose also that the level set $\{z : |h(z)| \geq \kappa\}$ is a union of at most K intervals for all large enough κ.

C4' $p'(z), q'(z) \geq 0$ for all $z < \frac{1}{L}$ and $p'(z), q'(z) \leq 0$ for all $z > 1 - \frac{1}{L}$.

Proof. The proof is identical to that of Proposition B.1. □

B.4 Analysis of the Initialization Scheme

Proposition B.3. *Let K be fixed. Suppose that $\frac{1}{\rho_L} \leq \frac{P_l}{Q_l} \leq \rho_L$ for all colors l.*

Let σ^l be a spectral clustering of the graph based on $\widetilde{A}_{ij} = \mathbf{1}(A_{ij} = l)$ and let $\widehat{P}_l, \widehat{Q}_l$ be estimates of P_l, Q_l constructed from σ^l.

Then, there is a positive constant C_{test} such that, with probability at least $1 - Ln^{-3+\delta_p}$,

1. *for all colors l satisfying $\Delta_l \geq C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$, we have that*

$$\frac{1}{\sqrt{5}} \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}} \leq \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq \frac{3}{\sqrt{3}} \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}} \tag{B.1}$$

2. *for all colors satisfying $\Delta_l < C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$, we have that*

$$\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq C_{test}^{1/2} C_{\beta,K,\delta_p} \sqrt{\frac{1}{n}} \tag{B.2}$$

Proof. Define

$$\gamma = 32C^2\beta \frac{K^2(P_l \vee Q_l)}{n(P_l - Q_l)^2}.$$

From Proposition B.5, we have that$l(\sigma^l, \sigma_0) \leq \gamma$ with probability at least $1 - n^{-4}$. We let $C_{test} = 10 \cdot 15^3 \cdot 32C^2\beta K^4 \vee C_{thresh}$.

Suppose $\Delta_l^2 \geq C_{test}\frac{P_l \vee Q_l}{n}$. Then, $C_{test}\frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \leq 1$ and thus, $\gamma \leq \frac{1}{10 \cdot 15^3 K^2}$. By Proposition A.1, we have that $|\widehat{P}_l - P_l| \leq \eta\Delta_l$ where $|\eta| \leq 15\sqrt{\gamma K \log\frac{K}{\gamma}} \leq \frac{1}{4}$. Likewise, we have that $|\widehat{Q}_l - Q_l| \leq \eta\Delta_l$.

To bound $\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}}$, we bound the numerator and denominator separately. First, we bound the numerator:

$$\begin{aligned}
|\widehat{P}_l - \widehat{Q}_l| &\leq |\widehat{P}_l - P_l| + |P_l - Q_l| + |\widehat{Q}_l - Q_l| \\
&\leq 2|\eta|\Delta_l + \Delta_l \\
&\leq \frac{3}{2}\Delta.
\end{aligned}$$

Furthermore,

$$\begin{aligned}|\widehat{P}_l - \widehat{Q}_l| &\geq |P_l - Q_l| - |\widehat{Q}_l - Q_l| - |\widehat{P}_l - P_l| \\ &\geq \Delta_l - 2|\eta|\Delta_l \\ &\geq \frac{1}{2}\Delta_l.\end{aligned}$$

Next, we bound the denominator:

$$\begin{aligned}\widehat{P}_l &\leq (P_l \vee Q_l) + |\eta|\Delta_l \\ &\leq (P_l \vee Q_l) + |\eta|(P_l \vee Q_l) \\ &\leq \frac{5}{4}(P_l \vee Q_l),\end{aligned}$$

where we have used the fact that $|\eta| \leq \frac{1}{4}$. Similar reasoning applies to give an upper bound on $\widehat{Q}_l$. For the lower bound on the denominator, we first observe that $\widehat{P}_l \geq P_l - |\eta|\Delta_l$ and that $\widehat{Q}_l \geq Q_l - |\eta|\Delta_l$.

Let us suppose without loss of generality that $P_l \geq Q_l$. Then, we have that

$$\widehat{P}_l \vee \widehat{Q}_l \geq (P_l \vee Q_l) - |\eta|\Delta_l \geq \frac{3}{4}(P_l \vee Q_l)$$

Therefore, we have that

$$\frac{1}{\sqrt{5}} \frac{\Delta_l}{P_l \vee Q_l} \leq \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq \frac{3}{\sqrt{3}} \frac{\Delta_l}{P_l \vee Q_l}$$

Now we move on to the second claim where we suppose $\Delta_l^2 \leq C_{test} \frac{P_l \vee Q_l}{n}$.

From proposition A.1, we have

$$|\widehat{P}_l - P_l| \leq \eta \left(\Delta_l \vee \sqrt{\frac{P_l \vee Q_l}{n}} \right)$$

where $|\eta| \leq C_{\beta,K,\delta_p}$. $|\eta|$ is bounded by a constant here because the best upper bound we have on γ is that $\gamma \leq 1$.

$$|\widehat{P}_l - P_l| \leq C_{test}^{1/2} C_{\beta,K,\delta_p} \sqrt{\frac{P_l \vee Q_l}{n}}$$

and likewise with $|\widehat{Q}_l - Q_l|$, both uniformly for all l, with probability $1 - Ln^{-(3+\delta_p)}$.

Putting these together, we have that

$$\begin{aligned}|\widehat{P}_l - \widehat{Q}_l| &= |\widehat{P}_l - P_l + P_l - Q_l + Q_l - \widehat{Q}_l| \\ &\leq \Delta_l + |\widehat{P}_l - P_l| + |\widehat{Q}_l - Q_l| \\ &\leq \Delta_l + C_{test} C_{\beta,k,\delta} \sqrt{\frac{P_l \vee Q_l}{n}} \\ &\leq 2 C_{test} C_{\beta,k,\delta} \sqrt{\frac{P_l \vee Q_l}{n}}\end{aligned}$$

To bound the denominator term $\sqrt{\widehat{P}_l \vee \widehat{Q}_l}$, we have that $\widehat{P}_l \geq P_l - C_{test}^{1/2} C_{\beta,K,\delta_p} \sqrt{\frac{P_l \vee Q_l}{n}}$ and $\widehat{Q}_l \geq Q_l - C_{test}^{1/2} C_{\beta,K,\delta_p} \sqrt{\frac{P_l \vee Q_l}{n}}$. Suppose without loss of generality that $P_l \geq Q_l$ so that $P_l = (P_l \vee Q_l)$.

$$\widehat{P_l} \vee \widehat{Q_l} \geq (P_l \vee Q_l) - C_{test}^{1/2} C_{\beta,K,\delta_p} \sqrt{\frac{P_l \vee Q_l}{n}} \geq \frac{1}{2} P_l \vee Q_l$$

Where we used the assumption that $P_l \vee Q_l \geq \frac{c}{n}$.
Thus, we have that

$$\frac{|\widehat{P_l} - \widehat{Q_l}|}{\sqrt{\widehat{P_l} \vee \widehat{Q_l}}} \leq 2C_{test}^{1/2} C_{\beta,K,\delta} \sqrt{\frac{1}{n}}$$

□

Let l^* be the color chosen by the initialization algorithm. We will show that $\frac{nI_{l^*}}{\rho_L} \to \infty$.

Proposition B.4. *Let $a_n = \frac{nI_L}{L\rho_L^2}$ and assume that $a_n \to \infty$.*

For large enough n, we have that, with probability at least $1-2Ln^{-(3+\delta_p)}$, we have that $\frac{n(P_{l^*}-Q_{l^*})^2}{(P_{l^*} \vee Q_{l^*})\rho_L^2} \geq c \cdot a_n$ for some constant c.

Proof. Throughout this proof, we let C denote a $\Theta(1)$ sequence whose value may change from line to line.

Let C_{test} be the constant in proposition B.3.

Note that $I_L = C \sum_{l=1}^L \frac{\Delta_l^2}{P_l \vee Q_l}$, therefore, there must exist some color l_n (we leave the dependency on n implicit and denote it just by l) such that $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq C\frac{nI_L}{L} = Ca_n\rho_L^2$. The same color l must also satisfy $\Delta_l \geq C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$.

Suppose that the probability event of proposition B.3 holds, which happens with probability at least $1 - Ln^{-(3+\delta)}$.

Step 1. We claim that l^* satisfies $\Delta_{l^*} \geq C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$. Let l be a color such that $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq Ca_n\rho_L^2$ and suppose l^* does not satisfy $\Delta_{l^*} \geq C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$.

Then, we have that, by proposition B.3,

$$\frac{|\widehat{P_l} - \widehat{Q_l}|}{\sqrt{\widehat{P_l} \vee \widehat{Q_l}}} \leq \frac{|\widehat{P_{l^*}} - \widehat{Q_{l^*}}|}{\sqrt{\widehat{P_{l^*}} \vee \widehat{Q_{l^*}}}} \leq C_{test}^{1/2} C_{\beta,K,\delta_p} \sqrt{\frac{1}{n}}$$

But, because l satisfies $\Delta_l \geq C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$, we also have

$$\frac{|\widehat{P_l} - \widehat{Q_l}|}{\sqrt{\widehat{P_l} \vee \widehat{Q_l}}} \geq \frac{1}{\sqrt{5}} \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}} \geq C\sqrt{a_n \rho_L^2 \frac{1}{n}}$$

Since $a_n \to \infty$ and $\rho_L \geq 1$, we have a contradiction.

Step 2:

Again, let l be a color such that $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq Ca_n\rho_L^2$.

$$
\begin{aligned}
\frac{|P_{l^*} - Q_{l^*}|}{\sqrt{P_{l^*} \vee Q_{l^*}}} &\geq \frac{\sqrt{3}}{3} \frac{|\widehat{P}_{l^*} - \widehat{Q}_{l^*}|}{\sqrt{\widehat{P}_{l^*} \vee \widehat{Q}_{l^*}}} \\
&\geq \frac{\sqrt{3}}{3} \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \\
&\geq \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}} \frac{1}{\sqrt{15}} \\
&\geq C\sqrt{a_n \rho_L^2 \frac{1}{n}}
\end{aligned}
$$

where the first inequality follows from step 1 and proposition B.3. The third inequality again follows from proposition B.3.

The conclusion thus follows. □

B.5 Analysis of the spectral clustering algorithm

Define $\overline{d} = \frac{1}{n}\sum_{u=1}^n d_u$ be the average degree.

Proposition B.5. *Suppose that an unweighted A is drawn from a homogeneous stochastic block model with probabilities p, q and cluster imbalance factor β, with the number of communities K fixed. Suppose $p, q \geq \frac{c}{n}$ for some absolute constant c.*

Suppose we run spectral clustering (algorithm 3.3) with tuning parameter $\mu \geq 16C^2\beta$ and trim threshold $\tau = \widetilde{C}\overline{d}$, where $\widetilde{C}$ is a constant depending on K. Let σ be the output.

Suppose that $128\mu\beta C^2 K^3 \frac{(p\vee q)}{n(p-q)^2} \leq 1$.

Then, we have that, probability at least $1 - n^{-C'}$ where $C' \geq 4$,

$$
l(\sigma, \sigma_0) \leq 32C^2\beta \frac{K^2(p \vee q)}{n(p-q)^2}
$$

for some constant C.

Proof. First, we see that for an appropriate choice of $\widetilde{C}$ and large enough n, the parameter τ lies in the range described by Lemma B.1 with probability at least $1 - \exp(-cn)$. Hence,

$$
\|T_\tau(A) - P\|_2 \leq C\sqrt{n(p \vee q) + 1}.
$$

Thus, we have that,

$$
\begin{aligned}
\|\widehat{A} - P\|_2 &\leq \|T_\tau(A) - P\|_2 + \|\widehat{A} - T_\tau(A)\|_2 \\
&\leq 2\|T_\tau(A) - P\|_2 \\
&\leq 2C\sqrt{n(p \vee q) + 1}
\end{aligned}
$$

The second inequality follows because $\widehat{A}$ is the best rank-K approximation of $T_\tau(A)$ and $\text{rank}(P) = K$, so $\|T_\tau(A) - \widehat{A}\|_2 \leq \|T_\tau(A) - P\|_2$ by the Eckart-Young-Mirsky Theorem.

Thus, we have that

$$\begin{aligned}\sum_{u=1}^{n} \|\widehat{A}_u - P_u\|_2^2 &= \|\widehat{A} - P\|_F^2 \\ &\leq 2KC^2 n(p \vee q)\end{aligned}$$

We also know that, if u, v are in different clusters,

$$\|P_u - P_v\|_2^2 \geq \frac{2}{\beta K}(p-q)^2 n$$

Suppose that the P_u's are known, then we can cluster v by matching $\widehat{A}_v$ to the closest P_u. We would make a mistake only if $\|\widehat{A}_u - P_u\|_2^2 \geq \frac{1}{\beta K}(p-q)^2 n$. Thus, the number of mistakes we make cannot be larger than

$$\frac{\sum_{u=1}^{n} \|\widehat{A}_u - P_u\|_2^2}{\frac{1}{\beta K}(p-q)^2 n} \leq \frac{2KC^2 n(p \vee q)}{\frac{1}{\beta K}(p-q)^2 n} \leq \frac{2\beta K^2 C^2 (p \vee q)}{(p-q)^2}$$

But because we do not know the true P_u's, we use the set S as a surrogate.

First, we define a point u as valid if $\|\widehat{A}_u - P_v\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$ for some v, not necessarily in the same cluster as u. u is declared invalid if the condition is not fulfilled.

Secondly, for a node u, define $u^* = \arg\min_{v \in S} \|\widehat{A}_u - P_v\|_2^2$, so P_{u^*} is the row of the P matrix closest to $\widehat{A}_u$. Notice that if u is valid, then $\|\widehat{A}_u - P_{u^*}\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$.

We then claim the following:

Claim 1: S contains only valid points.
Claim 2: For every pair of distinct nodes $u, v \in S$, we have $P_{u^*} \neq P_{v^*}$.

First, let us suppose that these two claims are true and see that the proposition follows. The number of mistakes we make is bounded by the number of invalid points plus the number of misclassified valid points. Note that if u is invalid, we have $\|\widehat{A}_u - P_u\|_2^2 \geq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. We claim that the same inequality holds for any valid point u incorrectly assigned to a point in S. Suppose for a contradiction that $\|\widehat{A}_u - P_u\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$.

Let w be a point in S such that $P_u = P_{w^*}$. We then have $\|\widehat{A}_u - P_{w^*}\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. By claim 1, w is a valid point and thus, $\|\widehat{A}_w - P_{w^*}\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. By combining these two bounds with the triangle inequality, we have that $\|\widehat{A}_u - \widehat{A}_w\|_2^2 \leq \frac{1}{4}\frac{1}{\beta K}(p-q)^2 n$.

Since u was assigned to $w' \in S$, $\|\widehat{A}_u - \widehat{A}_{w'}\|_2^2 \leq \frac{1}{4}\frac{1}{\beta K}(p-q)^2 n$, and since w' is valid, $\|\widehat{A}_u - P_{w'^*}\|_2^2 \leq \frac{1}{\beta K}(p-q)^2 n$.

By applying triangle inequality between $\|\widehat{A}_u - P_{w^*}\|$ and $\|\widehat{A}_u - P_{w'^*}\|$, we conclude that:

$$\|P_{w^*} - P_{w'^*}\|_2^2 < 2\frac{1}{\beta K}(p-q)^2 n$$

Since $P_{w^*} \neq P_{w'^*}$ by **claim 2**, $\|P_{w^*} - P_{w'^*}\|_2^2 \geq 2\frac{1}{\beta K}(p-q)^2 n$. Thus, we have a contradiction. Hence, it cannot be that $\|\widehat{A}_u - P_u\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$ to start out with.

Thus, the number of mistakes is bounded by

$$\frac{\sum_{u=1}^{n} \|\widehat{A}_u - P_u\|_2^2}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{2KC^2 n(p \vee q)}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{32\beta K^2 C^2 (p \vee q)}{(p-q)^2},$$

as wanted.

Proof of Claim 1: Recall that given a point u, the neighbors of u are $N(u) = \{v : \|\widehat{A}_u - \widehat{A}_v\|_2^2 \leq \mu K^2 \frac{\overline{d}}{n}\}$. Furthermore, $\overline{d} \leq 2(p \vee q)n$ with probability $1 - e^{-n}$. We condition on this event so that if $v \in N(u)$, then $\|\widehat{A}_u - \widehat{A}_v\| \leq 2\mu K^2(p \vee q)$.

We prove the claim by showing that an invalid point u cannot have $\frac{1}{\mu}\frac{n}{K}$ neighbors. We have that by definition of invalidity, $\|\widehat{A}_u - P_w\|_2^2 \geq \frac{1}{16\beta K}(p-q)^2 n$ for any w.

Let v be a neighbor of u. Since $\frac{2\mu K^2(p\vee q)}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{1}{64}$ by assumption, we have that $\|\widehat{A}_v - P_w\|_2^2 \geq \frac{1}{64\beta K}(p-q)^2 n$ for any w. Therefore, it follows that $\|\widehat{A}_v - P_v\|_2^2 \geq \frac{1}{64\beta K}(p-q)^2 n$.

Because we have a bound on the total error, the number of neighbors of u is bounded by

$$\frac{\sum_{v=1}^n \|\widehat{A}_v - P_v\|_2^2}{\frac{1}{64\beta K}(p-q)^2 n} \leq \frac{2KC^2 n(p \vee q)}{\frac{1}{64\beta K}(p-q)^2 n} \leq \frac{128\beta K^2 C^2 (p \vee q)}{(p-q)^2}$$

This quantity is less than $\frac{1}{\mu}\frac{n}{K}$ by assumption.

Proof of Claim 2: We first claim that in every cluster, at least half the points u satisfy $\|\widehat{A}_u - P_u\|_2^2 \leq \frac{1}{4}\mu K^2(p\vee q)$. This is because the total error is bounded by $\sum_{u=1}^n \|\widehat{A}_u - P_u\|_2^2 \leq 2KC^2 n(p \vee q)$ and thus, the total number of points that violate the condition is at most $\frac{2KC^2 n(p\vee q)}{\frac{1}{4}\mu K^2 (p \vee q)} \leq \frac{n}{2\beta K}$ by the assumption that $\mu \geq 16C^2\beta$.

If $\|\widehat{A}_w - P_w\|_2^2 \leq \frac{1}{4}\mu K^2(p \vee q)$ for $w \in \{u, v\}$, we also have $\|\widehat{A}_u - \widehat{A}_v\|_2^2 \leq \mu K^2 (p \vee q)$. This is because

$$\|\widehat{A}_u - \widehat{A}_v\|_2 \leq \|\widehat{A}_u - P_u\|_2 + \|\widehat{A}_v - P_v\|_2,$$

since $P_u = P_v$. Thus, in every cluster, there exists a point u such that $N(u) \geq \frac{1}{\mu}\frac{n}{K}$.

Now we are ready to prove claim 3. Suppose for contradiction that the algorithm is placing w' into S but that $P_{w'^*} = P_{w^*}$ for some w that already exists in S. Then, because w, w' are both valid, we have bounds for both $\|\widehat{A}_w - P_{w^*}\|$ and $\|\widehat{A}_{w'} - P_{w^*}\|$. Applying triangle inequality then gives $\|\widehat{A}_w - \widehat{A}_{w'}\|_2^2 \leq \frac{1}{4\beta K}(p-q)^2 n$.

On the other hand, because S does not yet have K nodes, there must be a row of P not equal to P_{v^*} for any $v \in S$. The cluster that corresponds to this missing row must, by our neighborhood size analysis, contain a node u such that $N(u) \geq \frac{1}{\mu}\frac{n}{K}$ and that

$$\|\widehat{A}_u - P_{u^*}\|_2^2 \leq \frac{1}{4}\mu K^2 (p \vee q) \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$$

the second inequality follows from the assumption of the proposition statement.

Since $P_{u^*} \neq P_{v^*}$ for any $v \in S$, $\|P_{u^*} - P_{v^*}\|_2^2 \geq 2\frac{1}{\beta K}(p-q)^2 n$ for all $v \in S$. v is valid by claim 1 and thus, $\|\widehat{A}_v - P_{v^*}\| \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. So we have that, by triangle inequality, that $\|\widehat{A}_u - \widehat{A}_v\|_2^2 \geq \frac{1}{\beta K}(p-q)^2 n$ for all $v \in S$.

This is a contradiction because u is farther away from every $v \in S$ than w' is from w. So, the algorithm would have put u into the set S rather than w'.

□

B.6 Supporting lemmas

Lemma B.1. *(Lemma 5 of [**?**])*

Let $P \in [0,1]^{n \times n}$ be a symmetric matrix. Let A be an adjacency matrix such that $A_{uu} = 0$, $A_{uv} \sim Ber(P_{uv})$ for the lower triangular part $u < v$. For any $C' > 0$, there exists some $C > 0$ such that

$$\|T_\tau(A) - P\|_2 \leq C\sqrt{np_{max} + 1}$$

with probability at least $1 - n^{-C'}$, *uniformly over* $\tau \in [C_1(np_{max}+1), C_2(np_{max}+1)]$, *for some sufficiently large constants* C_1, C_2, *where* $p_{max} = \max_{u \geq v} P_{uv}$.

C Proof of Proposition 4.1

First suppose $\|\theta_1 - \theta_0\| \to 0$. Then $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$ by Lemma C.2. Note that condition A1 follows directly from condition B1 and condition A5 follows directly from condition B5. In Propositions C.1, C.2, C.3 below, we establish conditons A2, A3, and A4, respectively.

Now suppose that $\|\theta_1 - \theta_0\| = \Theta(1)$. Then $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \Theta(1)$ by Lemma C.2. Condition A1' follows directdly from B1, while condition A4' follows directly from condition B5. In Propositions C.1 with $\rho = \exp(L^{1/r})$, we derive condition A2'. In Proposition C.3, we derive condition A3'.

C.1 Supporting propositions

Proposition C.1. *Suppose assumption B3, B4 holds. There exists an interval* $R \subset \{x : \frac{1}{\rho} \leq \left|\frac{p(x)}{q(x)}\right| \leq \rho\}$ *such that*

$$\Phi\{R^c\} \leq C \frac{\|\theta_1 - \theta_0\|^r}{(\log \rho)^r} \leq \frac{H^r}{(\log \rho)^r}$$

where $H \equiv \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$.

Proof. We start with the observation that

$$\begin{aligned}
\log \frac{p(x)}{q(x)} &= f_{\theta_1}(x) - f_{\theta_0}(x) \\
&= (\theta_1 - \theta_0)^{\mathsf{T}} \nabla_\theta f_{\overline{\theta}}(x) \\
|\log \frac{p(x)}{q(x)}| &\leq \|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\| \\
&\leq \|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\|
\end{aligned}$$

where $\overline{\theta}$ is some convex combination of θ_0, θ_1.

Therefore, $\{x : \frac{1}{\rho} \leq \left|\frac{p(x)}{q(x)}\right| \leq \rho\} \supset \left\{x : \sup_\theta \|\nabla_\theta f_\theta(x)\| \leq \frac{\log \rho}{\|\theta_1 - \theta_0\|}\right\}$ and we take the latter quantity to be R.

R is an interval by B3 for small enough $\|\theta_1 - \theta_0\|$ or large enough ρ.

Since we have that

$$\int_{-\infty}^{\infty} \sup_\theta \|\nabla_\theta f_\theta(x)\|^r \phi(x) dx < \infty$$

By Markov's inequality, we have

$$\Phi\left\{x : \sup_\theta \|\nabla_\theta f_\theta(x)\| > \frac{\log \rho}{\|\theta_1 - \theta_0\|}\right\} \leq C \frac{\|\theta_1 - \theta_0\|^r}{(\log \rho)^r}$$

An application of lemma C.2 finishes the proof.

□

Proposition C.2. *Suppose assumptions B1-B4 are satisfied. Let* $R \subset \mathbb{R}$ *be the interval in proposition C.1.*

Then, we have that,

$$\int_R \frac{1}{\alpha}\left|\frac{p(x)}{q(x)}-1\right|^r q(x)dx \le \infty$$

Proof. By lemma C.1, we have that $\alpha \asymp \|\theta_1-\theta_0\|$. Thus, we need only show that

$$\int_R \frac{1}{\|\theta_0-\theta_1\|}\left|\frac{p(x)}{q(x)}-1\right|^r q(x)dx \le \infty$$

$$\begin{aligned}
\frac{1}{\|\theta_1-\theta_0\|}\left|\frac{p(x)}{q(x)}-1\right| &= \frac{1}{\|\theta_1-\theta_0\|}\left|\exp(f_{\theta_1}(x)-f_{\theta_0}(x))-1\right| \\
&= \frac{1}{\|\theta_1-\theta_0\|}\left|(\theta_1-\theta_0)^\mathsf{T}\nabla_\theta f_{\overline{\theta}}(x)\right|\exp(f_{\overline{\theta}}(x)-f_{\theta_0}(x)) \\
&\le \|\nabla_\theta f_{\overline{\theta}}(x)\|\exp(f_{\overline{\theta}}(x)-f_{\theta_0}(x)) \\
&= \|\nabla_\theta f_{\overline{\theta}}(x)\|\exp\left((\overline{\theta}-\theta_1)^\mathsf{T}\nabla_\theta f_{\widetilde{\theta}}(x)\right) \\
&\le \|\nabla_\theta f_{\overline{\theta}}(x)\|\exp\left(\|\theta_1-\theta_0\|\|\nabla_\theta f_{\widetilde{\theta}}(x)\|\right)
\end{aligned}$$

Where $\overline{\theta}, \widetilde{\theta}$ are some convex combinations of θ_0, θ_1.

Therefore,

$$\begin{aligned}
\int_R\left(\frac{1}{\|\theta_1-\theta_0\|}\left|\frac{p(x)}{q(x)}-1\right|\right)^r q(x)dx &\le \int_R \|\nabla_\theta f_{\overline{\theta}}(x)\|^r \exp\left(r\|\theta_1-\theta_0\|\|\nabla_\theta f_{\widetilde{\theta}}(x)\|\right)q(x)dx \\
&\le \int_R \|\nabla_\theta f_{\overline{\theta}}(x)\|^r e^{r\log\rho}q(x)dx \\
&= \int_R \|\nabla_\theta f_{\overline{\theta}}(x)\|^r \rho^r q(x)dx \\
&\le \rho^r\int_{-\infty}^{\infty} \|\nabla_\theta f_{\overline{\theta}}(x)\|^r q(x)dx < \infty
\end{aligned}$$

□

Proposition C.3. *Suppose assumptions B1-B4 are satisfied. Let $R \subset \mathbb{R}$ be the interval in proposition C.1.*

Then, we have that,

$$\int_R\left|\frac{1}{\alpha}\frac{p'(x)-q'(x)}{q(x)}\right|^{t/(1-t)}\phi(x)dx < \infty$$

Proof. Again, using the fact that $\alpha \asymp \|\theta_1-\theta_0\|$, we need only prove that

$$\int_R\left|\frac{1}{\|\theta_1-\theta_0\|}\frac{p'(x)-q'(x)}{q(x)}\right|^{t/(1-t)}\phi(x)dx < \infty$$

Note that

$$\begin{aligned}
\frac{1}{\|\theta_0-\theta_1\|}\frac{p'(x)-q'(x)}{q(x)} &= \frac{1}{\|\theta_0-\theta_1\|}\left[f'_{\theta_1}\frac{p(x)}{q(x)}-f'_{\theta_0}(x)\right] \\
&= \frac{1}{\|\theta_0-\theta_1\|}\left\{(f'_{\theta_1}(x)-f'_{\theta_0}(x))\frac{p(x)}{q(x)}+f'_{\theta_0}\left(\frac{p(x)}{q(x)}-1\right)\right\} \\
&= \nabla_\theta f'_{\overline{\theta}}(x)\frac{p(x)}{q(x)}+\frac{1}{\|\theta_1-\theta_0\|}f'_{\theta_0}(x)\left(\frac{p(x)}{q(x)}-1\right)
\end{aligned}$$

where $\overline{\theta}$ is some convex combination of θ_1, θ_0.

Therefore, we have:

$$\begin{aligned}
&\int_R \left| \frac{1}{\|\theta_1 - \theta_0\|} \frac{p'(x) - q'(x)}{q(x)} \right|^{t/(1-t)} \phi(x) dx \\
&= \int_R \left| \nabla_\theta f'_{\overline{\theta}}(x) \frac{p(x)}{q(x)} + \frac{1}{\|\theta_1 - \theta_0\|} f'_{\theta_0}(x) \left(\frac{p(x)}{q(x)} - 1 \right) \right|^{t/(1-t)} \phi(x) dx
\end{aligned}$$

To show that this integral is finite, we need only show, regardless of the value of $t/(1-t)$, that the two components have finite integrals:

$$\int_R \left| \nabla_\theta f'_{\overline{\theta}}(x) \frac{p(x)}{q(x)} \right|^{t/(1-t)} \phi(x) dx \quad \text{and} \quad \int_R \left| \frac{1}{\|\theta_1 - \theta_0\|} f'_{\theta_0}(x) \left(\frac{p(x)}{q(x)} - 1 \right) \right|^{t/(1-t)} \phi(x) dx$$

We bound the first integral.

$$\int_R \left| \nabla_\theta f'_{\overline{\theta}}(x) \frac{p(x)}{q(x)} \right|^{t/(1-t)} \phi(x) dx \leq \int_R \left| \nabla_\theta f'_{\overline{\theta}}(x) \right|^{t/(1-t)} \rho \phi(x) dx < \infty$$

And then the second.

$$\begin{aligned}
&\int_R \left| \frac{1}{\|\theta_1 - \theta_0\|} f'_{\theta_0}(x) \left(\frac{p(x)}{q(x)} - 1 \right) \right|^{t/(1-t)} \phi(x) dx \\
&\leq \int_R |f'_{\theta_0}(x)| \left| \frac{1}{\|\theta_1 - \theta_0\|} \left(\frac{p(x)}{q(x)} - 1 \right) \right|^{t/(1-t)} \phi(x) dx \\
&\leq \left\{ \int_R |f'_{\theta_0}(x)|^{2t/(1-t)} \phi(x) dx \right\}^{1/2} \left\{ \int_R \left| \frac{1}{\|\theta_1 - \theta_0\|} \left(\frac{p(x)}{q(x)} - 1 \right) \right|^{2t/(1-t)} \phi(x) dx \right\}^{1/2}
\end{aligned}$$

The first quantity is finite by assumption. A bound for the second quantity follows from proposition C.2.

□

C.2 Supporting lemmas

Lemma C.1. *Suppose assumptions B1-B4 hold. Let R be the interval in Proposition C.1. Define* $\alpha = \int_R q(x) \left(\frac{p(x)}{q(x)} - 1 \right)^2 dx$. *Then we have that*

$$\alpha \asymp \|\theta_1 - \theta_0\|$$

Proof.

$$\begin{aligned}
\alpha^2 &= \int_R \left(\frac{p(x)}{q(x)} - 1 \right)^2 q(x) dx \\
&= \int_R \left| \exp\left(f_{\theta_1}(x) - f_{\theta_0}(x) \right) - 1 \right| q(xdx \\
&= \int_R \left((\theta_1 - \theta_0)^\top \nabla_\theta f_{\overline{\theta}}(x) \exp\left(f_{\overline{\theta}}(x) - f_{\theta_0}(x) \right) \right)^2 q(x) dx
\end{aligned}$$

First for the upper bound:

$$\begin{aligned}
&\leq \int_R \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp\Big(f_{\overline{\theta}}(x) - f_{\theta_0}(x)\Big) \exp(f_{\overline{\theta}}(x))dx \\
&\leq \int_R \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp\Big(\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\widetilde{\theta}}(x)\|\Big) \exp(f_{\overline{\theta}}(x))dx
\end{aligned}$$

Since we are in R, we have that $\|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\| \leq \log\rho$. We can thus continue the bound:

$$\begin{aligned}
&\leq \|\theta_1 - \theta_0\|^2 \int_R \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 e^{\log\rho} \exp(f_{\overline{\theta}}(x))dx \\
&\leq \|\theta_1 - \theta_0\|^2 \rho \int_{-\infty}^{\infty} \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp(f_{\overline{\theta}}(x))dx \\
&\lesssim \|\theta_1 - \theta_0\|^2
\end{aligned}$$

Now for the lower bound:

$$\begin{aligned}
\alpha^2 &\geq \int_R \big((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x)\big)^2 \exp\Big(-|f_{\overline{\theta}}(x) - f_{\theta_0}(x)|\Big) \exp(f_{\overline{\theta}}(x))dx \\
&\geq \int_R \big((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x)\big)^2 \exp\Big(-\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\|\Big) \exp(f_{\overline{\theta}}(x))dx \\
&\geq e^{-C_R} (\theta_1 - \theta_0)^\mathsf{T} \left(\int_R (\nabla_\theta f_{\overline{\theta}}(x))(\nabla_\theta f_{\overline{\theta}}(x))^\mathsf{T} \exp(f_{\overline{\theta}}(x))dx\right) (\theta_1 - \theta_0)
\end{aligned}$$

Define

$$\widetilde{G}_\theta = \int_R (\nabla_\theta f_{\overline{\theta}}(x))(\nabla_\theta f_{\overline{\theta}}(x))^\mathsf{T} \exp(f_{\overline{\theta}}(x))dx$$

For increasing ρ or as $\|\theta_1 - \theta_0\| \to 0$, $R \to \mathbb{R}$, therefore, $\lambda_{min}(\widetilde{G}_\theta) \to \lambda_{min}(G_\theta) > 0$. Hence, $\alpha^2 \gtrsim \|\theta_1 - \theta_0\|^2$

□

Lemma C.2. *Under assumption D2, we have that*

$$\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = c\|\theta_0 - \theta_1\|_2^2$$

where $c_{\min} \leq c \leq \frac{1}{4} c_{\max} d_\Theta$.

Proof.

$$\begin{aligned}
&\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \\
&= \int q(x) \left(\sqrt{\frac{p(x)}{q(x)}} - 1\right)^2 dx \\
&= \int q(x) \Big(\exp\Big(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2\Big) - 1\Big)^2 dx
\end{aligned}$$

Now, let's look at the exponential term $\exp(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2) - 1$.

Define $h(\theta) = \exp(f_\theta(x)/2 - f_{\theta_0}(x)/2)$. It is clear that $h(\theta_0) = 1$ and that we wish to bound $h(\theta_1) - h(\theta_0)$.

$$\begin{aligned} h(\theta_1) - h(\theta_0) &= (\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta h(\overline{\theta}) \\ &= \frac{1}{2}(\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x) \exp(f_{\overline{\theta}}(x)/2 - f_{\theta_0}(x)/2) \\ |h(\theta_1) - h(\theta_0)| &\leq \frac{1}{2}\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\| \exp(f_{\overline{\theta}}(x)/2 - f_{\theta_0}(x)/2) \end{aligned}$$

where $\overline{\theta} \in \Theta$ is some convex combination of θ_1, θ_0.

Thus, we have that

$$\begin{aligned} \int q(x)\Big(\exp\Big(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2\Big) - 1\Big)^2 dx &\leq \int q(x)\frac{1}{4}\|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp(f_{\overline{\theta}}(x) - f_{\theta_0}(x))dx \\ &\leq \frac{1}{4}\|\theta_1 - \theta_0\|^2 \int \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp(f_{\overline{\theta}}(x))dx \\ &\leq \frac{1}{4}\|\theta_1 - \theta_0\|^2 \operatorname{tr}(G_{\overline{\theta}}) \\ &\leq \frac{1}{4}\|\theta_1 - \theta_0\|^2 c_{\max} d_\Theta \end{aligned}$$

$$\begin{aligned} \int q(x)\Big(\exp\Big(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2\Big) - 1\Big)^2 dx &= \int \Big((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x)\Big)^2 \exp(f_{\overline{\theta}}(x))dx \\ &= (\theta_1 - \theta_0)^\mathsf{T} G_{\overline{\theta}}(\theta_1 - \theta_0) \\ &\geq c_{\min}\|\theta_1 - \theta_0\|^2 \end{aligned}$$

□

D Proof of Theorem 4.3

We first state a lemma that gives an alternative characterization of the Renyi divergence.

Lemma D.1. *Let P, Q be two probability measures on $\mathbb{R}$ absolutely continuous with respect to each other. Suppose that part of P, Q singular to the Lebesgue measure is a point mass at zero, denoted P_0, Q_0.*

Define

$$I = -2\log \int \left(\frac{dP}{dQ}\right)^{1/2} dQ \qquad D = \inf_{Y \in \mathcal{P}} \max\left\{\int \log\frac{dY}{dP}dY, \int \log\frac{dY}{dQ}dY\right\}$$

where we use $\mathcal{P}$ to denote probability measures absolutely continuous to P (and thus Q).

Then, we have that

$$I = 2D$$

Proof. First, we note that D must be finite since we can substitute $Y = P$ or $Y = Q$. We claim that D is equivalent to the following:

$$\inf_{Y \in \mathcal{P}} \int \log\frac{dY}{dP}dY$$

$$\int \log\frac{dP}{dQ}dY = 0$$

This is because for any $Y \in \mathcal{P}$ such that $\int \log \frac{dP}{dQ} dY \neq 0$, we have that $\int \log \frac{dY}{dP} dY \neq \int \log \frac{dY}{dQ} dY$. Suppose without loss of generality that the former quantity is larger. Therefore, it is possible to take $\widetilde{Y} = (1-\epsilon)Y + \epsilon P$ for ϵ small enough such that $\max \left\{ \int \log \frac{d\widetilde{Y}}{dP} d\widetilde{Y},\, \int \log \frac{d\widetilde{Y}}{dQ} d\widetilde{Y} \right\}$ strictly decreases.

Since the new optimization is convex in Y, we can solve and get $Y_0 = \frac{P_0^{1/2} Q_0^{1/2}}{Z}$ and $(1-Y_0)y(x) = \frac{((1-P_0)\cdot p(x))^{1/2}((1-Q_0)\cdot q(x))^{1/2}}{Z}$. Here, we denote by $(1-Y_0)y(x), (1-P_0)p(x), (1-Q_0)q(x)$ the Radon-Nikodym derivative of the continuous part of Y, P, Q with respect to the Lebesgue measure. Z is the normalization term: $Z = P_0^{1/2} Q_0^{1/2} + \int \sqrt{(1-P_0)p(x)(1-Q_0)q(x)} dx$.

$$\begin{aligned}
\int \log \frac{dY}{dP} dY &= \log \frac{1}{Z} \left\{ \left(\frac{Q_0}{P_0} \right)^{1/2} Y_0 + \int \left(\frac{(1-P_0)p(x)}{(1-Q_0)q(x)} \right)^{1/2} (1-Y_0)y(x)dx \right\} \\
&= \log \frac{dP}{dQ} dY - \log Z \\
&= -\log Z
\end{aligned}$$

It is straightforward to verify that $-2 \log Z = I$.

□

Proof. (of Theorem 4.3)
Throughout this proof, we let C denote a constant whose value may change from line to line.

Without loss of generality, let us suppose that node 1 was placed in cluster 1, i.e, $\sigma_0(1) = 1$.

Let Φ denote the measure on the graph described by the colored SBM. Let Ψ denote a measure on the graph defined as follows:

1. If $u, v \neq 1$, then A_{uv} is distributed just as in Φ.
2. If $u = 1$ and $v \notin C_1 \cup C_2$, then A_{1v} is distributed just as in Φ.
3. If $u = 1$ and $v \in C_1 \cup C_2$, then A_{1v} is distributed as Y, where Y is the distribution that minimizes D in lemma D.1. ($Y_0 \propto (P_0 Q_0)^{1/2}$ and $(1-Y_0)y(x) \propto \sqrt{(1-P_0)p(x)(1-Q_0)q(x)}$)

The log-likelihood ratio $\log \frac{dP_\Psi}{dP_\Phi}$ is

$$\mathcal{Q} = \sum_{v \neq 1,\, v \in C_1} \log \frac{Y(A_{1v})}{P(A_{1v})} + \sum_{v \neq 1,\, v \in C_2} \log \frac{Y(A_{1v})}{Q(A_{1v})}$$

where we use the notation $P(A_{1v}) = P_0$ if $A_{1v} = 0$ and $P(A_{1v}) = (1-P_0)p(A_{1v})$ if $A_{1v} \neq 0$.

Let $f(n)$ be an arbitrary function and $\widehat{\sigma}$ be an arbitrary clustering algorithm. Let $\mathcal{E} = \mathcal{E}(\widehat{\sigma}(G))$ and let v_1 denote node 1.

$$P_\Psi(\mathcal{Q} \leq f(n)) = P_\Psi(\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}) + P_\Psi(\mathcal{Q} \leq f(n), v_1 \notin \mathcal{E})$$

The first term can be bounded.

$$\begin{aligned}
P_\Psi(\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}) &= \int_{\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}} dP_\Psi \\
&= \int_{\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}} \exp(\mathcal{Q}) dP_\Phi \\
&\leq \exp(f(n)) P_\Phi(\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}) \\
&\leq \exp(f(n)) P_\Phi(v_1 \in \mathcal{E}) \\
&\leq \exp(f(n)) \mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0)
\end{aligned}$$

The last inequality follows because $\mathbb{E}l(\widehat{\sigma}(G), \sigma_0) = \frac{1}{n}\sum_{v=1}^n P(v \in \mathcal{E}(\widehat{\sigma}(G))) = P(v_1 \in \mathcal{E}(\widehat{\sigma}(G))$ since the nodes are exchangeable when σ_0 is drawn uniformly at random.

To bound the second term, note that under P_Ψ, any clustering algorithm has at most $\frac{1}{K}$ chance of clustering v_1 correctly. Thus

$$P_\Psi(\mathcal{Q} \leq f(n), v_1 \notin \mathcal{E}) \leq P_\Psi(v_1 \notin \mathcal{E}) \leq \frac{1}{2}$$

Combining these two bounds, we have that

$$P_\Psi(\mathcal{Q} \leq f(n)) \leq \exp(f(n))\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0) + \frac{1}{2}$$

Let $f(n) = \log \frac{1}{4\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0)}$, then

$$P_\Psi\left(\mathcal{Q} \leq \log \frac{1}{4\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0)}\right) \leq \frac{3}{4}$$

From Chebyshev's inequality, we also have that

$$P_\Psi\left(\mathcal{Q} \leq \mathbb{E}_\Psi\mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}\right) \geq 4/5$$

Hence, we get that $\log \frac{1}{4\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0)} \leq \mathbb{E}_\Psi\mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}$.

$$\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0) \geq \frac{1}{4}\exp\Big(-(\mathbb{E}_\Psi\mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})})\Big)$$

We now just need to compute $\mathbb{E}_\Psi\mathcal{Q}$ and $V_\Psi(\mathcal{Q})$.

$$\begin{aligned}
\mathbb{E}_\Psi\mathcal{Q} &= \mathbb{E}_\Psi \sum_{v\neq 1,\, v\in C_1} \log\frac{Y(A_{1v})}{P(A_{1v})} + \sum_{v\neq 1,\, v\in C_2} \log\frac{Y(A_{1v})}{Q(A_{1v})} \\
&= (n/K - 1)\int \log\frac{dY}{dP}dY + (n/K)\int \log\frac{dY}{dQ}dY \\
&= (n/K - 1/2)2D \\
&= (n/K - 1/2)I \\
&= \frac{nI}{K}(1 + o(1))
\end{aligned}$$

Moving onto the variance term:

$$\begin{aligned}
V_\Psi(\mathcal{Q}) &= \sum_{v\neq 1, v\in C_1} V_\Psi\left(\log\frac{Y(A_{1v})}{P(A_{1v})}\right) + \sum_{v\neq 1, v\in C_2} V_\Psi\left(\log\frac{Y(A_{1v})}{Q(A_{1v})}\right) \\
&\leq (n/K - 1)\mathbb{E}_\Psi\left(\log\frac{Y(A_{1v})}{P(A_{1v})}\right)^2 + (n/K)\mathbb{E}_\Psi\left(\log\frac{Y(A_{1v})}{Q(A_{1v})}\right)^2
\end{aligned}$$

Here, we have that

$$\begin{aligned}
\mathbb{E}_\Psi\left(\log\frac{Y(A_{1v})}{P(A_{1v})}\right)^2 &= \int\left(\log\frac{dY}{dP}\right)^2 dY \\
&= Y_0\log^2\frac{Y_0}{P_0} + (1 - Y_0)\int y(x)\log^2\frac{(1 - Y_0)y(x)}{(1 - P_0)p(x)}dx
\end{aligned}$$

For the first term, observe that $Y_0 = \frac{\sqrt{P_0 Q_0}}{Z} \to 1$ because $Z \to 1$ (we assumed that $I = -2\log Z$ is going to 0) and because $P_0, Q_0 \to 1$.

$$\begin{aligned}
\left|\log \frac{Y_0}{P_0}\right| &= \left|\frac{1}{2}\log\frac{Q_0}{P_0} - \log Z\right| \\
&\le \frac{1}{2}\left|\log\left(1 - \frac{P_0 - Q_0}{P_0}\right)\right| + I/2 \\
&\le \frac{1}{2}\Delta_0(1 + o(1)) + I/2
\end{aligned}$$

Here, $\Delta_0 = |P_0 - Q_0|$. The last inequality follows because $P_0 \to 1$ and $P_0 - Q_0 \to 0$.

Therefore,

$$Y_0 \log^2 \frac{Y_0}{Q_0} \le \frac{1}{4}(\Delta_0 + I)^2(1 + o(1))$$

Suppose $I \le \Delta_0$, then $(\Delta_0 + I)^2 \le 4\Delta_0^2 \le 4(\sqrt{P_0} - \sqrt{Q_0})^2(1 + o(1)) \le 4I(1 + o(1))$. The last inequality follows because $I = (\sqrt{P_0} - \sqrt{Q_0})^2 + (\sqrt{1-P_0} - \sqrt{1-Q_0})^2 + \sqrt{(1-P_0)(1-Q_0)} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$. Suppose $I \ge \Delta_0$, then $(\Delta_0 + I)^2 \le 4I^2 \le 4I(1 + o(1))$ because $I \to 0$.

Thus, the first term is always bounded by $4I(1 + o(1))$.

Now we turn our attention to the second term:

$$\begin{aligned}
(1 - Y_0)\int y(x)\log^2\frac{(1-Y_0)y(x)}{(1-P_0)p(x)}dx &\le (1 - Y_0)C \\
&\le (1 - \sqrt{P_0 Q_0})C(1 + o(1)) \\
&\le (1 - P_0)C(1 + o(1))
\end{aligned}$$

The first inequality follows because $\left|\log\frac{(1-Y_0)y(x)}{(1-P_0)p(x)}\right|$ is bounded by a constant. The second inequality follows because $Y_0 = \frac{\sqrt{P_0 Q_0}}{Z}$ and $Z \to 1$. The third inequality follows from the assumption that $1 - P_0 \asymp 1 - Q_0$.

Therefore, we have that

$$\mathbb{E}_\Psi \left(\log\frac{Y(A_{1v})}{P(A_{1v})}\right)^2 \le C(I + (1 - P_0))(1 + o(1))$$

$$\sqrt{V_\Psi(\mathcal{Q})} \le \left(\sqrt{\frac{nI}{K}} + \sqrt{(1 - P_0)\frac{n}{K}}\right) C(1 + o(1))$$

Suppose $nI/K \to 0$, then $\sqrt{nI/K} = o(nI/K)$. Under the assumption in the theorem statement that $H = \omega(\sqrt{K/n})$ and the fact that $I = (1 + o(1))\{(\sqrt{P_0} - \sqrt{Q_0})^2 + (\sqrt{1-P_0} - \sqrt{1-Q_0})^2 + \sqrt{(1-P_0)(1-Q_0)}H\}$, we have that $\sqrt{\frac{(1-P_0)n}{K}} = o(nI/K)$ as well. Thus, $\sqrt{5V_\Psi(\mathcal{Q})}$ is $o(nI/K)$.

If $nI/K \to c < \infty$, then $\mathbb{E}_\Psi l(\widehat{\sigma}, \sigma_0) > c' > 0$.

□

E Proof of Theorem 5.2

We will follow the proof strategy of Abbe et al. [2].

E.1 Main argument

We will show that if

$$(\sqrt{a}-\sqrt{b})^2+\sqrt{ab}\int(\sqrt{p(x)}-\sqrt{q(x)})^2<K$$

there with a probability of at least 1/3, we can find nodes $i \in A$ and $j \in B$ such that exchanging their community assignments has a larger likelihood than the ground truth. This would establish that the maximum likelihood estimator fails with probability at least 1/3. Although we will establish the proof for the case of two communities, we note that the proof below trivially extends to $K > 2$ communities each of size n, simply by taking A and B to be any two fixed communities from the K communities. For the sake of convenience, we assume n/K is an integer.

Let $A = \{1, 2, \dots, n/K\}$ and $B = \{n/K + 1, n + 2, \dots, 2n/K\}$. For $u \neq v$, let A_{uv} be the weight of the edge (u, v) and let $A_{uv} = 0$ if there exists no edge between u and v. Just as in the case of unlabeled edges, maximizing the likelihood in the labeled case may be interpreted as finding the min-cut for the stochastic block model, where the weight of an edge is $d_n(A_{uv})$ where we define $d_n(0) = \log \frac{P_0}{Q_0}$ and $d_n(x)$ for $x \neq 0$ to be

$$d_n(x) = \log\left(\frac{ap(x)}{bq(x)}\right).$$

We may describe $d_n(0)$ explicitly as

$$d_n(0) = \log\left\{\frac{1-a\log n/n}{1-b\log n/n}\right\} \tag{E.1}$$

Note that since $d_n(0) \to 0$ as $n \to \infty$, and since the likelihood ratio $p(x)/q(x)$ is assumed to be bounded, we may find a constant $\mathcal{M} > 0$ that upper-bounds d_n for all n. Thus,

$$\mathcal{M} \geq \max_x d_n(x), \qquad \text{for all } n.$$

For any node u and any subset of nodes H, denote

$$\mathcal{S}(u, H) = \sum_{v\in H, v\neq u} d_n(A_{uv}).$$

Using an argument along the lines of Lemma **??**, it is easy to check that if there exist nodes $u \in A$ and $v \in B$ such that

$$\mathcal{S}(u, A\setminus\{u\}) + \mathcal{S}(v, B\setminus\{v\}) < \mathcal{S}(u, B\setminus\{v\}) + \mathcal{S}(v, A\setminus\{u\}), \tag{E.2}$$

then the community assignment where $\sigma(u) = B$ and $\sigma(v) = A$ and every other assignment remains the same is more likely than the truth. Thus, the maximum likelihood estimator will fail if this happens. Define the following events:

$$\begin{aligned} F &= \text{ maximum likelihood fails,}\\ F_A &= \exists u \in A \ : \ \mathcal{S}(u, A\setminus\{i\}) < \mathcal{S}(u, B) - \mathcal{M},\\ F_B &= \exists v \in B \ : \ \mathcal{S}(v, B\setminus\{j\}) < \mathcal{S}(u, A) - \mathcal{M}. \end{aligned}$$

We have the following simple lemma:

Lemma E.1. *If $\mathbb{P}(F_A) \geq 2/3$, then $\mathbb{P}(F) \geq 1/3$.*

Proof. By symmetry, we have $\mathbb{P}(F_B) \geq 2/3$, so by a union bound, $\mathbb{P}(F_A \cap F_B) \geq 1/3$. Thus, with probability at least $1/3$, there exist nodes $i \in A$ and $j \in B$ such that

$$\begin{aligned} \mathcal{S}(i, A \setminus \{i\}) < \mathcal{S}(i, B) - \mathcal{M} \leq \mathcal{S}(i, B) - \mathcal{S}(i, j) = \mathcal{S}(i, B \setminus \{j\}), \qquad \text{and} \\ \mathcal{S}(j, B \setminus \{j\}) < \mathcal{S}(j, A) - \mathcal{M} \leq \mathcal{S}(j, A) - \mathcal{S}(i, j) = \mathcal{S}(j, A \setminus \{j\}). \end{aligned}$$

This implies

$$\mathcal{S}(i, A \setminus \{i\}) + \mathcal{S}(j, B \setminus \{j\}) < \mathcal{S}(i, B \setminus \{j\}) + \mathcal{S}(j, A \setminus \{j\}),$$

which from expression (E.2), implies that the maximum likelihood estimator fails. □

We now define $\gamma(n)$ and $\delta(n)$ as follows:

$$\gamma(n) = (\log n)^{\log^{\frac{2}{3}} n}, \qquad \text{and} \qquad \delta(n) = \frac{\sqrt{\log n}}{\log \log n}.$$

Let H be a fixed subset of A of size $\frac{n}{\gamma(n)}$. We will take $\gamma(n) \asymp (\log n)^{\log^{\frac{2}{3}} n}$, such that $\frac{n}{\gamma(n)}$ is an integer. Define the event Δ as follows:

$$\Delta = \text{ for all nodes } i \in H, \quad \mathcal{S}(i, H) < \delta(n).$$

We then have the following lemma:

Lemma E.2. $\mathbb{P}(\Delta) \geq \frac{9}{10}$.

Proof. Let Δ_i be the event $\mathcal{S}(i, H) < \delta(n)$. By a simple union bound calculation, we have

$$\mathbb{P}(\Delta) = 1 - \mathbb{P}(\Delta^c) = 1 - \mathbb{P}\left(\cup_{i \in H} \Delta_i^c\right) \geq 1 - |H| \cdot \mathbb{P}(\Delta_i^c).$$

We will show that

$$|H| \cdot \mathbb{P}(\Delta_i^c) = o(1),$$

by showing that

$$\log |H| + \log \mathbb{P}(\Delta_i^c) \to -\infty,$$

as $n \to \infty$. Let the weights of the edges from i to nodes within H be the random variables $\{X_1, \dots, X_{|H|-1}\}$. Note that the X_i's are independent and identically distributed according to p_n. We have

$$\mathbb{P}(\Delta_i^c) = \mathbb{P}\left(\mathcal{S}(i, H) \geq \frac{\sqrt{\log n}}{\log \log n}\right) = \mathbb{P}\left(\sum_{j=1}^{|H|-1} d_n(X_i) \geq \frac{\sqrt{\log n}}{\log \log n}\right) \leq \inf_{t > 0} \left\{ \frac{\mathbb{E}\left[e^{t d_n(X_1)}\right]^{|H|-1}}{e^{\frac{t\sqrt{\log n}}{\log \log n}}} \right\},$$

using a Chernoff bound in the last inequality. Thus, for $t > 0$, we have

$$\begin{aligned} \log |H| + \log \mathbb{P}(\Delta_i^c) &\leq \log \frac{n}{\gamma(n)} + \log \frac{\mathbb{E}\left[e^{t d_n(X_1)}\right]^{\frac{n}{\gamma(n)} - 1}}{e^{\frac{t\sqrt{\log n}}{\log \log n}}} \\ &= \log \frac{n}{\gamma(n)} + \left(\frac{n}{\gamma(n)} - 1\right) \log \mathbb{E}\left[e^{t d_n(X_1)}\right] - \frac{t\sqrt{\log n}}{\log \log n}. \end{aligned}$$

Picking $t = \sqrt{\log n} \log \log n$, the last expression simplifies to

$$-\log \gamma(n) + \left(\frac{n}{\gamma(n)} - 1\right) \log \mathbb{E}\left[e^{\sqrt{\log n}(\log \log n) d_n(X_1)}\right]. \tag{E.3}$$

We now analyze $\log \mathbb{E}\left[e^{\sqrt{\log n}(\log\log n) d_n(X_1)}\right]$ carefully. Note that

$$\begin{aligned}\log \mathbb{E}\left[e^{\sqrt{\log n}(\log\log n) d_n(X_1)}\right] &= \log\Bigg[\left(\frac{1-a\log n/n}{1-b\log n/n}\right)^{\sqrt{\log n}\log\log n}\left(1-\frac{a\log n}{n}\right) \\ &\qquad + \int_{\mathbb{R}}\left(\frac{ap(x)}{bq(x)}\right)^{\sqrt{\log n}\log\log n}\frac{a\log n}{n}p(x)dx\Bigg] \\ &:= \log(1+\mu_n+\nu_n),\end{aligned}$$

where

$$\begin{aligned}1+\mu_n &= \left(\frac{1-a\log n/n}{1-b\log n/n}\right)^{\sqrt{\log n}\log\log n}\left(1-\frac{a\log n}{n}\right), \quad \text{and} \\ \nu_n &= \int_{\mathbb{R}}\left(\frac{ap(x)}{bq(x)}\right)^{\sqrt{\log n}\log\log n}\frac{a\log n}{n}p(x)dx.\end{aligned}$$

The following bound holds for ν_n:

$$\nu_n \le C_1\frac{(\log n)^{C_2\sqrt{\log n}}}{n},$$

for suitable constants C_1, C_2. For μ_n, we have

$$\begin{aligned}\mu_n &= \left(\frac{1-a\log n/n}{1-b\log n/n}\right)^{\sqrt{\log n}\log\log n}\left(1-\frac{a\log n}{n}\right)-1 \\ &= \left(\left(\frac{1-a\log n/n}{1-b\log n/n}\right)^{n/\log n}\right)^{\frac{(\log n)^{3/2}\log\log n}{n}}\left(1-\frac{a\log n}{n}\right)-1.\end{aligned}$$

The term $\left(\frac{1-a\log n/n}{1-b\log n/n}\right)^{n/\log n}$ tends to a constant, $\exp(b-a)$. Thus, for large enough n, we may find constants $0<c_1<c_2$ such that $\left(\frac{1-a\log n/n}{1-b\log n/n}\right)^{n/\log n} \in (c_1, c_2)$. Using the Taylor series approximation of c_i^x near 0, we have

$$c_i^{\frac{(\log n)^{3/2}\log\log n}{n}} = 1+\frac{(\log n)^{3/2}\log\log n}{n}\log c_i + O\left(\left(\frac{(\log n)^{3/2}\log\log n}{n}\right)^2\right),$$

so

$$\begin{aligned}c_i^{\frac{(\log n)^{3/2}\log\log n}{n}}\left(1-\frac{a\log n}{n}\right)-1 &= \frac{(\log n)^{3/2}\log\log n}{n}\log c_i + O\left(\left(\frac{(\log n)^{3/2}\log\log n}{n}\right)^2\right) \\ &\qquad - \frac{a\log n}{n}\left(1+\frac{(\log n)^{3/2}\log\log n}{n}\log c_i\right).\end{aligned}$$

Thus, for large enough n, there exists a constant C_3 that satisfies

$$|\mu_n| \le \frac{C_3\log^2 n}{n}.$$

Using the bound

$$\log(1+\mu_n+\nu_n) \le |\mu_n|+|\nu_n|,$$

we conclude that

$$\log \mathbb{E}\left[e^{\sqrt{\log n}(\log\log n)d_n(X_1)}\right] \le C_1' \frac{(\log n)^{C_2'\sqrt{\log n}}}{n},$$

for a suitable constants C_1' and C_2'. Returning to the expression (E.3), we conclude that

$$-\log\gamma(n) + \left(\frac{n}{\gamma(n)} - 1\right)\log \mathbb{E}\left[e^{\sqrt{\log n}(\log\log n)d_n(X_1)}\right]$$

$$\le -\log\gamma(n) + \left(\frac{n}{\gamma(n)} - 1\right) C_1' \frac{(\log n)^{C_2'\sqrt{\log n}}}{n}.$$

Substituting $\gamma(n) = (\log n)^{\log^{\frac{2}{3}} n}$, we arrive at the upper bound

$$-\log^{\frac{2}{3}} n(\log\log n) + \left(\frac{n}{(\log n)^{\log^{\frac{2}{3}} n}} - 1\right) C_1' \frac{(\log n)^{C_2'\sqrt{\log n}}}{n}.$$

It is easy to check that as $n \to \infty$, we have

$$\left(\frac{n}{(\log n)^{\log^{\frac{2}{3}} n}} - 1\right) C_1' \frac{(\log n)^{C_2'\sqrt{\log n}}}{n} \to 0,$$

and

$$-\log^{\frac{2}{3}} n(\log\log n) \to -\infty.$$

This concludes the proof. □

Finally, define the events $F_H^{(i)}$ and F_H as follows:

$$\begin{aligned} F_H^{(i)} &= \text{ node } i \in H \text{ satisfies } \mathcal{S}(i, A\setminus H) + \delta(n) < \mathcal{S}(i,B) - \mathcal{M}, \\ F_H &= \cup_{i\in H} F_H^{(i)}, \end{aligned}$$

and define

$$\rho(n) = \mathbb{P}\Big(F_H^{(i)}\Big). \tag{E.4}$$

We have the following result:

Lemma E.3. *If $\rho(n) > \frac{\gamma(n)\log 10}{n}$, then $\mathbb{P}(F) > 1/3$ for sufficiently large n.*

Proof. We first show that $\mathbb{P}(F_H) > \frac{9}{10}$ for large enough n. Since the events $F_H^{(i)}$ are independent, we have

$$\mathbb{P}(F_H) = \mathbb{P}\left(\cup_{i\in H} F_H^{(i)}\right) = 1 - \mathbb{P}\left(\cap_{i\in H}\left(F_H^{(i)}\right)^c\right) = 1 - (1-\rho(n))^{\frac{n}{\gamma(n)}}.$$

Clearly, if $\rho(n)$ is not $o(1)$, then $\mathbb{P}(F)$ tends to 1 and we are done. If $\rho(n)$ is $o(1)$, then

$$\lim_{n\to\infty}(1-\rho(n))^{\frac{n}{\gamma(n)}} = \lim_{n\to\infty}(1-\rho(n))^{\frac{1}{\rho(n)}\frac{\rho(n)n}{\gamma(n)}} = \lim_{n\to\infty}\exp\left(-\frac{\rho(n)n}{\gamma(n)}\right) < \frac{1}{10},$$

where the last inequality used the fact that $\rho(n) > \frac{\gamma(n)\log 10}{n}$. Hence, $\mathbb{P}(F_H) > \frac{9}{10}$, as claimed.

Now note that $\Delta \cap F_H \subseteq F_A$. By Lemma E.2, we also have $\mathbb{P}(\Delta) \ge \frac{9}{10}$. Hence,

$$\mathbb{P}(F_A) \ge \mathbb{P}(\Delta) + \mathbb{P}(F_H) - 1 \ge \frac{8}{10} > \frac{2}{3},$$

which combined with Lemma E.1 implies the desired result. □

Let $\{X_i\}_{i\geq 1}$ be a sequence of i.i.d. random variables distributed according to p_n, and let $\{Y_i\}_{i\geq 1}$ be a sequence of i.i.d. random variables distributed according to q_n. From the definition (E.4) of $\rho(n)$, and using independence, we have

$$\begin{aligned}\rho(n) &= \mathbb{P}\left(\sum_{i=1}^{\frac{n}{K}} d_n(Y_i) - \sum_{i=1}^{\frac{n}{K}-\frac{n}{\gamma(n)}} d_n(X_i) > \delta(n) + \mathcal{M}\right)\\ &\geq P\left(\sum_{i=1}^{\frac{n}{K}-\frac{n}{\gamma(n)}} d_n(Y_i) - \sum_{i=1}^{\frac{n}{K}-\frac{n}{\gamma(n)}} d_n(X_i) > \delta(n) + \mathcal{M} - \widehat{\delta}(n)\right) \times \mathbb{P}\left(\sum_{i=\frac{n}{K}-\frac{n}{\gamma(n)}+1}^{\frac{n}{K}} d_n(Y_i) \geq \widehat{\delta}(n)\right),\end{aligned} \tag{E.5}$$

for any $\widehat{\delta}(n)$. We will choose a suitable $\widehat{\delta}(n)$ so that

$$\mathbb{P}\left(\sum_{i=\frac{n}{K}-\frac{n}{\gamma(n)}+1}^{\frac{n}{K}} d_n(Y_i) \geq \widehat{\delta}(n)\right) \longrightarrow 1. \tag{E.6}$$

Note that $d_n(Y_i)$ is a random variable satisfying

$$\mathbb{P}\left(d_n(Y_i) = \log\left\{\frac{1-a\log n/n}{1-b\log n/n}\right\}\right) = 1 - \frac{b\log n}{n}.$$

Thus,

$$\mathbb{P}\left(d_n(Y_i) = \log\left\{\frac{1-a\log n/n}{1-b\log n/n}\right\},\ \text{for all}\ \frac{n}{K} - \frac{n}{\gamma(n)} - 1 \leq i \leq \frac{n}{K}\right) = \left(1 - \frac{b\log n}{n}\right)^{\frac{n}{\gamma(n)}}.$$

We may check that

$$\left(1 - \frac{b\log n}{n}\right)^{\frac{n}{\gamma(n)}} \longrightarrow 1,$$

implying that

$$\mathbb{P}\left(\sum_{i=\frac{n}{K}-\frac{n}{\gamma(n)}+1}^{\frac{n}{K}} d_n(Y_i) = \frac{n}{\gamma(n)}\cdot\log\left\{\frac{1-a\log n/n}{1-b\log n/n}\right\}\right) \longrightarrow 1.$$

Thus, equation (E.6) holds with

$$\widehat{\delta}(n) = \left|\frac{n}{\gamma(n)}\cdot\log\left\{\frac{1-a\log n/n}{1-b\log n/n}\right\}\right|. \tag{E.7}$$

Since

$$\widehat{\delta}(n) = O\left(\frac{\log n}{\gamma(n)}\right) = o(\sqrt{\log n}),$$

we have $\delta(n) + \mathcal{M} - \widehat{\delta}(n) = o(\sqrt{\log n})$. Define

$$T(N, p_n, q_n, \epsilon) = \mathbb{P}\left(\sum_{i=1}^{N}\Big(d_n(Y_i) - d_n(X_i)\Big) \geq \epsilon\right). \tag{E.8}$$

Using this notation gives

$$T\left(\frac{n}{K} - \frac{n}{\gamma(n)}, p_n, q_n, \delta(n) + \mathcal{M} - \widehat{\delta}(n)\right) = \mathbb{P}\left(\sum_{i=1}^{\frac{n}{K}-\frac{n}{\gamma(n)}} d_n(Y_i) - \sum_{i=1}^{\frac{n}{K}-\frac{n}{\gamma(n)}} d_n(X_i) \geq \delta(n) + \mathcal{M} - \widehat{\delta}(n)\right),$$

and using Lemma E.4, we conclude that

$$-\log T\left(\frac{n}{K}-\frac{n}{\gamma(n)}, p_n, q_n, \delta(n)+\mathcal{M}-\widehat{\delta}(n)\right) \leq \left(\frac{1}{K}\int_{\mathbb{R}}\left(\sqrt{ap(x)}-\sqrt{bq(x)}\right)^2 dx\right)\log n + o(\log n). \tag{E.9}$$

Substituting the bounds (E.6) and (E.9) into equation (E.5), we then conclude that

$$-\log \rho(n) \leq \left(\frac{1}{K}\int_{\mathbb{R}}\left(\sqrt{ap(x)}-\sqrt{bq(x)}\right)^2 dx\right)\log n + o(\log n).$$

In particular, when

$$\left(\frac{1}{K}\int_{\mathbb{R}}\left(\sqrt{ap(x)}-\sqrt{bq(x)}\right)^2 dx\right) < 1,$$

we have

$$-\log \rho(n) \leq \log n - \log \gamma(n) - \log\log 10,$$

for sufficiently large n. Lemma E.3 then implies that maximum likelihood fails with probability at least $\frac{1}{3}$, completing the proof of the theorem.

E.2 Supporting lemmas

Lemma E.4. *Let $\omega(n) = o(\sqrt{\log n})$ and $N(n) = \frac{n}{K}(1+o(1))$. Then*

$$-\log T\left(N(n), p_n, q_n, \omega(n)\right) \leq \left(\frac{1}{K}\int_{\mathbb{R}}\left(\sqrt{ap(x)}-\sqrt{bq(x)}\right)^2 dx\right)\log n + o(\log n).$$

Proof. We will use the proof strategy found in Zhang and Zhou [32]. Let

$$Z = d_n(Y) - d_n(X),$$

where $X \sim p_n$ and $Y \sim q_n$. Let $M(t) = \mathbb{E}e^{tZ}$, and note that

$$\begin{aligned} t^\star &= \arg\min_{t>0} M(t) = \frac{1}{2}, \\ M(t^\star) &= \left(\int_{\mathbb{R}}\sqrt{p_n(x)q_n(x)}dx\right)^2, \\ I &= -\log M(t^\star) = -2\log\left(\int_{\mathbb{R}}\sqrt{p_n(x)q_n(x)}dx\right). \end{aligned}$$

Let $S_N = \sum_{i=1}^{N(n)} Z_i$, where the Z_i's are i.i.d. and distributed according to Z, and denote the distribution of Z by p_Z. Define

$$\eta(n) = \log^{\frac{3}{4}} n.$$

Then

$$\begin{aligned} \mathbb{P}\left(S_N \geq \omega(n)\right) &\geq \sum_{z:S_N\in[\omega(n),\eta(n))}\prod_{i=1}^{N(n)} p_Z(z_i) \\ &\geq \frac{M^{N(n)}(t^\star)}{e^{t^\star\eta(n)}}\sum_{z:S_N\in[\omega(n),\eta(n))}\prod_{i=1}^{N(n)}\frac{e^{t^\star z_i}p_Z(z_i)}{M(t^\star)} \\ &= \exp\left(-N(n)I-\frac{\eta(n)}{2}\right)\sum_{z:S_N\in[\omega(n),\eta(n))}\prod_{i=1}^{N(n)}\frac{e^{t^\star z_i}p_Z(z_i)}{M(t^\star)}, \end{aligned} \tag{E.10}$$

where the second inequality uses the fact that $e^{t^\star \eta(n)} \geq e^{t^\star \sum_i z_i}$ when $\sum_{i=1}^{N(n)} z_i < \eta(n)$.

Now denote $r(w) = \frac{e^{t^\star w} p_Z(w)}{M(t^\star)}$, and note that r defines a probability distribution. Defining $W_1, W_2, \ldots, W_n$ to be i.i.d. random variables with probability mass function $r(w)$, we then have

$$\sum_{z: S_N \in [\omega(n), \eta(n))} \prod_{i=1}^{N(n)} \frac{e^{t^* z_i} p_Z(z_i)}{M(t^*)} = \mathbb{P}\left(\omega(n) \leq \sum_{i=1}^{N(n)} W_i < \eta(n) \right). \tag{E.11}$$

By Lemma E.5, it follows that

$$\frac{1}{\sqrt{\log N(n)}} \sum_{i=1}^{N(n)} W_i \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

for some constant $\nu > 0$. Furthermore, by our choices of $\omega(n)$, $N(n)$, and $\eta(n)$, we have

$$\frac{\omega(n)}{\sqrt{\log N(n)}} \to 0, \qquad \text{and} \qquad \frac{\eta(n)}{\sqrt{\log N(n)}} \to +\infty.$$

Thus,

$$\mathbb{P}\left(\frac{\omega(n)}{\sqrt{\log N(n)}} \leq \frac{1}{\sqrt{\log N(n)}} \sum_{i=1}^{N(n)} W_i < \frac{\eta(n)}{\sqrt{\log N(n)}} \right) \to 1/2,$$

implying that the left-hand probability expression becomes larger that $1/4$ for all large enough n. Combining this with the bounds (E.10) and (E.11), we then obtain

$$\mathbb{P}(S_N \geq \omega(n)) \geq \exp\left(-N(n) I - \frac{\log^{\frac{3}{4}} n}{2} - \log 4 \right).$$

Using $N = \frac{n}{K}(1 + o(1))$, we arrive at

$$-\log T\left(N(n), p_n, q_n, \omega(n)\right) = -\log \mathbb{P}(S_N \geq \omega(n)) \leq \left(\frac{1}{K} \int_{\mathbb{R}} \left(\sqrt{ap(x)} - \sqrt{bq(x)} \right)^2 dx \right) \log n + o(\log n).$$

This concludes the proof.

□

Lemma E.5. *Let $\{W_i\}_{i \geq 1}$ be i.i.d. random variables distributed as $r(w)$. Then*

$$\frac{\sum_{i=1}^n W_i}{\sqrt{\log n}} \xrightarrow{d} \mathcal{N}(0, \nu^2),$$

as $n \to \infty$, where $\nu > 0$ is a constant.

Proof. We show that the moment generating function of $\frac{\sum_{i=1}^n W_i}{\sqrt{\log n}}$ converges to that of a normal random variable. By a simple computation, we may check that r has the following distribution:

$$r(0) = 1 - \frac{C_0 \log n}{n} + O\left(\frac{\log^2 n}{n^2} \right),$$

for some constant $C_0 > 0$, and $r(x) = \Theta(\log n / n) + O\left(\frac{\log^2 n}{n^2} \right)$ elsewhere. Since W is a bounded and symmetric random variable, it is easy to see that its moment generating function is given by

$$\mathbb{E} e^{tW} = 1 + \int_{\mathbb{R}} r(\widehat{w}) \left(e^{t\widehat{w}/2} - e^{-t\widehat{w}/2} \right)^2 d\widehat{w}, \tag{E.12}$$

using the fact that $r(0) = 1 - \int_{\mathbb{R}} 2r(\widehat{w})d\widehat{w}$. Using the expression (E.12), the moment generating function of W is then given by

$$\mathbb{E}e^{tW} \quad = \quad 1 \quad + \quad \int_{\mathbb{R}} \left(\frac{C(\widehat{w})\log n}{n} + O\left(\frac{\log^2 n}{n^2}\right)\right) \left(e^{t\widehat{w}/2} - e^{-t\widehat{w}/2}\right)^2 d\widehat{w}$$

Substituting $\frac{t}{\sqrt{\log n}}$ in place of t and using the approximation $a^{x/2} - a^{-x/2} = x\log a + O(x^2 \log^2 a)$ for $x = o(1)$, we arrive at

$$\begin{aligned}
\mathbb{E}e^{tW/\sqrt{\log n}} &= 1 + \int_{\mathbb{R}} \left(\frac{C(\widehat{w})\log n}{n} + O\left(\frac{\log^2 n}{n^2}\right)\right) \left(\frac{t\widehat{w}}{\sqrt{\log n}} + O\left(\frac{1}{\log n}\right)\right)^2 d\widehat{w} \\
&\quad + \int_{\mathbb{R}} O\left(\frac{\log^2 n}{n^2}\right) \left(\frac{t\widehat{w}}{\sqrt{\log n}} + O\left(\frac{1}{\log n}\right)\right)^2 d\widehat{w} \\
&= 1 + \frac{Ct^2}{n} + o\left(\frac{1}{n}\right),
\end{aligned}$$

for a suitable constant C. Hence, the moment generating function of $\frac{\sum_{i=1}^n W_i}{\sqrt{\log n}}$ is given by

$$\left(\mathbb{E}e^{tW/\sqrt{\log n}}\right)^n = \left(1 + \frac{Ct^2}{n} + o\left(\frac{1}{n}\right)\right)^n \longrightarrow e^{Ct^2},$$

which is the moment generating function of $\mathcal{N}(0, 2C)$. This completes the proof.

□

F Additional useful lemmas

Lemma F.1. *Let* $I = -2\log\left(\sqrt{P_0 Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right)$ *and let* $H = (\sqrt{P_0} - \sqrt{Q_0})^2 + \int \left(\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)}\right)^2 dx$*. If* $I < 1 - \epsilon$*, then we have that*

$$I = H(1+\eta)$$

where $|\eta| \le \frac{H}{4\epsilon}$*. Therefore, we have that* $I \to 0$ *iff* $H \to 0$ *and that if* $I \to 0$*,* $I = H(1 + o(1))$*.*

Proof.

$$\begin{aligned}
I &= -2\log\left(\sqrt{P_0 Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right) \\
&= -2\log\left(1 - \frac{1}{2}\left((\sqrt{P_0} - \sqrt{Q_0})^2 + \int (\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)})^2 dx\right)\right) \\
&= -2\log\left(1 - \frac{1}{2}H\right) \\
&= 2\frac{1}{2}H(1+\eta)
\end{aligned}$$

where $|\eta| \le \frac{H}{2\epsilon}$.

□

Lemma F.2. *Suppose* $x \ge 0$ *and* $1 \ge \epsilon > 0$*, then we have that, for all* $0 \le x < 1 - \epsilon$*,*

$$\log(1-x) = -(1+\eta)x$$

where $|\eta| \le \frac{x}{2\epsilon}$

Proof. This follows by taking the Taylor expansion of $\log(1-x)$ around $x=0$. □

Lemma F.3. *Define $f(z) = \frac{1-\frac{z}{2}-\sqrt{1-z}}{z}$ for $z \leq 1$ and $z \neq 0$ and define $f(0) = 0$. Then we have that,*

$$|f(z)| \leq |z|$$

for all $z \leq 1$.

Proof. Define $f(z) = \frac{1-\frac{z}{2}-\sqrt{1-z}}{z}$, where we set $f(0) = 0$. Note that f is continuous.

The derivative of f is

$$f'(z) = -\frac{1}{z^2} - \frac{z-2}{2z^2\sqrt{1-z}}$$

It is straight forward to check that $f'(z) \geq 0$ for all $z < 1$ and that we can define $f'(0) = \frac{1}{4}$ such that $f'(z)$ is continuous.

Therefore, $f(z)$ is monotonic and maximized at $z = 1$, yielding $f(1) = 1/2$ and minimized at $\lim_{z\to-\infty} f(z) = -\frac{1}{2}$.

Now we perform case analysis. Suppose $z < -1/2$, then $|f(z)| \leq \frac{1}{2} < |z|$.

Suppose $-1/2 \leq z \leq 1/2$. By Taylor expansion, we have

$$\sqrt{1-z} = 1 - \frac{1}{2}z - \frac{1}{8}z^2 - \frac{1}{16}z^3 - ...\frac{(n+1)!!}{2^n n!}z^n - ...$$

Therefore,

$$\begin{aligned}\left|\sqrt{1-z} - (1-\frac{z}{2})\right| &\leq \frac{1}{8}(|z|^2 + |z|^3 + ...)\\ &\leq \frac{1}{8}|z|^2(1 + |z| + |z|^2 + ...)\\ &\leq \frac{1}{8}|z|^2\frac{1}{1-|z|}\\ &\leq \frac{1}{4}|z|^2\end{aligned}$$

Therefore, $|f(z)| \leq \frac{1}{4}|z|$.

Finally, suppose $z > 1/2$. Then,

$|f(z)| \leq \frac{1}{2} < z$.

□