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, 12, 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, 15, 20, 25, 26, 29].

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) [16, 30, 31]. 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, 13, 14, 19, 21, 22, 33]).

However, the pairwise connections in many real-world networks possess a natural weighting structure [7, 24]. For example, in social networks, information may be available quantifying the strength of a tie, such as the frequency of interactions between the individuals [28]; 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 [27]. 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 [24] 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 [33] and Gao et al [10], 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.

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 Section 4 we provide the statements of our main results concerning the behavior of our algorithm in terms of misclassification error rates and exact recovery. Section 5 highlights the key technical components employed in the analysis of our algorithm. We close in Section 6 with further implications of our work and open questions related to our results.

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:

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

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}:

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

The edge weight random variable is then:

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

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 $q = 1 - Q_0$.

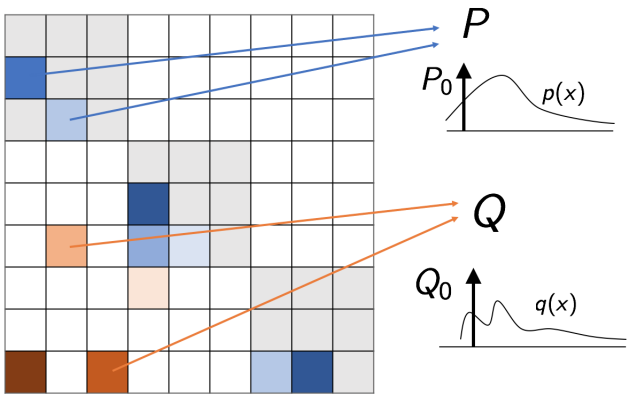

Figure 1: Weighted stochastic block model

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

In this paper we aim to find a tractable community recovery algorithm whose misclustering error can be shown to converge to zero at an optimal rate.

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 $\text{Hamming}(\cdot, \cdot)$ 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 [33] and Gao et al [10] show that the minimax optimal rate of convergence for the unweighted stochastic block model is of the order $\exp\left(-(1+o(1))\frac{nI_{\text{Ber}}}{K}\right)$. $I_{\text{Ber}} = -2\log\sqrt{P_0Q_0}+\sqrt{(1-P_0)(1-Q_0)}$ is the Renyi divergence of order $1/2$ between $\text{Ber}(P_0)$ and $\text{Ber}(Q_0)$, 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 [32] 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 Other notions of 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 [33] beyond the $\log n$ degree setting where $\frac{nI_{\text{Ber}}}{K\log n}$ was shown to be the threshold. 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, 33]. 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: $\widehat{\sigma}^{SBM}_{MLE} = \arg\max_\sigma \sum_{\substack{(u,v)\in E\\ \sigma(u)=\sigma(v)}} \log\frac{p(1-q)}{q(1-p)}$. Since $\log\frac{p(1-q)}{q(1-p)} > 0$, the estimator $\widehat{\sigma}^{SBM}_{MLE}$ 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, whose adjacency matrix we continue to call A.

3. **Add noise.** For a fixed constant $c > 0$, let $\delta = \frac{cL}{n}$. We perform the following process on every edge of the labeled graph, independently of other edges: With probability $1 - \delta$, keep an edge as it is, and with probability δ erase the edge and replace it with an edge with label uniformly drawn from the set of L labels. We continue calling the modified adjacency matrix as A.

4. **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.

5. **Initialization Part 2.** Let A_{l^*} be the adjacency matrix for color l^*. For each $u \in \{1, \ldots, n\}$, we perform spectral clustering on $A_{l^*} \setminus \{u\}$, which is adjacency matrix with vertex u removed. We output n clusters $\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n$, where $\widetilde{\sigma}_u$ is a clustering on $\{1, 2, \ldots, n\} \setminus \{u\}$ for $1 \leq u \leq n$.

6. **Refinement.** From each $\widetilde{\sigma}_u$, we generate a $\widehat{\sigma}_u$ which is a clustering on $\{1, 2, \ldots, n\}$ which retains the assignments specified by $\widetilde{\sigma}_u$ for $\{1, 2, \ldots, n\} \setminus \{u\}$, and assigns $\widehat{\sigma}_u(u)$ by maximizing the likelihood looking at only the neighborhood around u.

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

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.

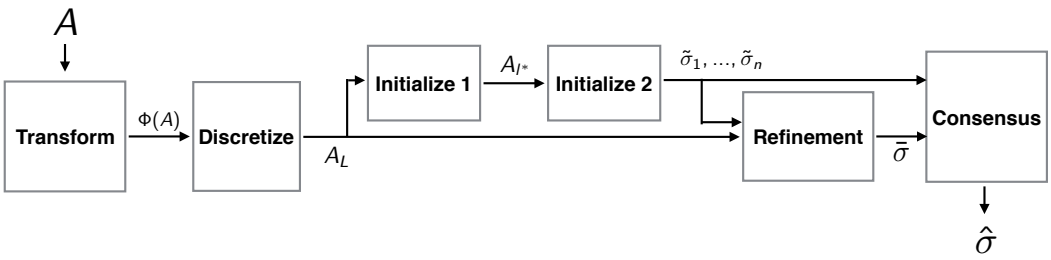

Figure 2: Add a box indicating the add noise step. Pipeline for the our proposed algorithm

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 with L labels

Divide $[0, 1]$ into L bins, labeled $Bin_1, \ldots, \text{bin}_l$.
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
end for
Output A

3.3 Add noise

This part of the algorithm is primarily for technical reasons. As detailed in the proof of Proposition 5.1 in Appendix A, deliberately forming a noisy version of the graph has a negligible effect on the separation between the distributions specifying within and across community edge labels, but it has the desirable effect of ensuring that all the edge labels occur with probability at least c/n. This lower bound is crucial to our analysis in the subsequent steps of the recovery algorithm.

Algorithm 3.2 Add noise

Input: A labeled network with L labels, a constant c
Output: A labeled network A with L labels

for every edge (u, v) **do**
  With probability $1 - \frac{cL}{n}$, do nothing. With probability $\frac{cL}{n}$ replace edge label with a label drawn uniformly at random from $\{1, 2, \ldots, L\}$
end for
Output A

3.4 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 3.3, 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.

Algorithm 3.3 Initialization

Input: A labeled network A with L labels

Output: A set of clusterings $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$, where $\widetilde{\sigma}_u$ is a clustering on $\{1,2,\dots,n\}\setminus\{u\}$

Separate A_L into L networks $\{A_l\}_{l=1,\dots,L}$ 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 $\tau = \overline{d}$ and $\mu \geq C\beta$ to get $\widetilde{\sigma}_l$, where C is an appropriately chosen large constant
    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 $A_{l^*}\setminus\{u\}$ by removing node u from A_{l^*}
    Perform spectral clustering on $A_{l^*}\setminus\{u\}$ to get $\widetilde{\sigma}_u$
end for
Output the set of clusterings $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$.

For technical reasons, we will actually create n separate rough clusterings $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$ 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 3.3 involves several applications of spectral clustering. We describe the spectral clustering algorithm used as a subroutine in Algorithm 3.4 below:

Algorithm 3.4 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 $i = 2,\dots,K$ **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 $u = 1,\dots,n$ **do**
    Take $\arg\min_i \|\widehat{A}_u - \widehat{A}_{S[i]}\|_2$ and assign $\sigma(u) = i$
end for

Importantly, note that we may always choose the parameter μ sufficiently large such that Algorithm 3.4 generates a set S with $|S| = K$.

3.5 Refinement and consensus

Our refinement and consensus step closely follow the method described by Gao et al [10]. In the refinement step, we use the set of initial clusterings $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$ to generate a more accurate clustering for the labeled network. We do this by locally maximizing an approximate log-likelihood expression for each of the nodes $u = 1, \dots, n$. The consensus step is to resolve a technical cluster label consistency problem that arises after the refinement stage.

Algorithm 3.5 Refinement

Input: A labeled network A and a set of rough clusterings $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$, where $\widetilde{\sigma}_u$ is a clustering on the set $\{1, 2, \dots, n\} \setminus \{u\}$ for each u

Output: a clustering $\widehat{\sigma}$ over the whole network

1: **for** each node u **do**

2: Estimate $\{\widehat{P}_l, \widehat{Q}_l\}_{l=0,\dots,L}$ from $\widetilde{\sigma}_u$

3: Let $\widehat{\sigma}_u : [n] \to [K]$ where $\widehat{\sigma}_u(v) = \widetilde{\sigma}_u(v)$ for all $v \neq u$ and

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

4: **end for**

5: Let $\widehat{\sigma}(1) = \widehat{\sigma}_1(1)$ ▷ Consensus Stage

6: **for** each node $u \neq 1$ **do**

$$\widehat{\sigma}(u) = \arg\max_k |\{v \,:\, \widehat{\sigma}_1(v) = k\} \cap \{v \,:\, \widehat{\sigma}_u(v) = \widehat{\sigma}_u(u)\}|$$

7: **end for**

8: 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

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

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

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

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

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

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 $\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx = \Theta(1)$, that K is fixed, and that $I = o(1)$. Then, we have that*

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

For the proof of Theorem 4.2, see Appendix F.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, $f'_\theta(x) \ge (\log \phi)'(x) > 0$ for all $x \le -c'$ and $f'_\theta(x) \le (\log \phi)'(x) < 0$ for all $x \ge c'$.

Note that B4 translates to a moment condition on the transformation distribution Φ. We then have the following result, proved in Appendix G:

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 uniformly at random in the sense that $\sigma_0 = \sigma'_0 \circ \pi$ where $\sigma'_0 : [n] \to [K]$ is any fixed clustering and $\pi \in S_n$ is a random permutation on $[n]$. We take 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.

Theorem 4.3. *Suppose we have K clusters at least two of which are of size $\frac{n}{\beta K}$ for some constant $\beta \geq 1$. Let the true clustering σ_0 be drawn uniformly at random. Suppose $I \to 0$. 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. Then, we have that, for any community recovery algorithm $\widehat{\sigma}$,*

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

and if $\frac{nI}{K} \to c < \infty$ for some constant c, then $\mathbb{E}l(\widehat{\sigma}, \sigma_0) \geq c' > 0$ for some constant c'.

The proof of Theorem 4.3 is provided in Appendix H. 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 [32]. 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 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.

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

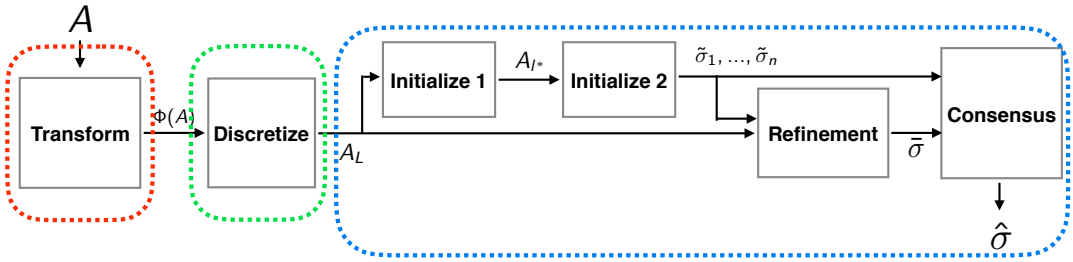

Figure 3: The add noise component also goes into the blue section. Analysis of the right-most blue region is in subsection 5.1, of the middle green region in subsection 5.2, and of the left-most red region in subsection 5.3

5.1 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 $l = 1, ...L$. 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 5.1. *Suppose we have $l = 1, ..., L$ 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 $\frac{nI_L}{L\rho_L^4} \to \infty$. 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 [32] have proposed an algorithm for the labeled SBM that achieves the same rate of convergence. Proposition 5.1 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.

5.2 Discretization of the Renyi divergence

The rate of Proposition 5.1 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 5.2 and 5.3 shows exactly that.

The following proposition is useful for proving Theorem 4.1:

Proposition 5.2. *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_h intervals for all large enough κ, and for a constant K_h.

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 $I = -2\log\left(\sqrt{P_0 Q_0} + \int\sqrt{(1 - P_0)(1 - Q_0)p(z)q(z)}dz\right)$ and $I_L = -2\log\left(\sqrt{P_0 Q_0} + \sum_{l=1}^{L}\sqrt{P_l Q_l}\right)$. Then, we have that

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

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

The following proposition is useful for proving Theorem 4.2:

Proposition 5.3. *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_h 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.

5.3 Analysis on the transformation function

Proposition 5.2 and 5.3 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 5.2 (or Proposition 5.3). This is done through Proposition F.1 and F.2.

6 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 [32] 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 [11, 34] 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] Christopher Aicher, Abigail Z Jacobs, and Aaron 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] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

[11] Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.

[12] 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.

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

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

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

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

[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*.

[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. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.

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

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

[27] 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.

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

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

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

[31] 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.

[32] Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

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

[34] Yunpeng Zhao, Elizaveta Levina, Ji Zhu, et al. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

A Proof of Proposition 5.1

The proof of Proposition 5.1 is quite involved. We structure the proof to follow the flow of our algorithm. Algorithm 3.2 is where we deliberately add noise to the graph by changing edge colors at random. This may seem like a suboptimal step to take: after all, adding such random noise is only going to destroy information and make it harder to cluster the communities. However, we show in Lemma B.2 that this process does not significantly affect the Renyi divergence term I_L, and ensures that the new probabilities of edge labels are at least c/n for a constant c. This lower bound is crucial to the rest of our analysis. The proof of Lemma B.2 may be found in Appendix B.6. To simplify notation, we continue to refer the new edge label probabilities as P_l and Q_l throughout the proof.

Next, our algorithm performs a spectral clustering using only the edges with label l, and calculates $\widehat{I}_l := \frac{(\widehat{P}_l - \widehat{Q}_l)^2}{\widehat{P}_l \vee \widehat{Q}_l}$, where $\widehat{P}_l$ and $\widehat{Q}_l$ are the estimated probabilities obtained by using the output of spectral clustering, defined as follows:

$$\widehat{P}_l = \frac{\sum_{u \neq v \,:\, \widetilde{\sigma}_l(u) = \widetilde{\sigma}_l(v)} (A_l)_{uv}}{|u \neq v \,:\, \widetilde{\sigma}_l(u) = \widetilde{\sigma}_l(v)|}, \quad \text{and}$$
$$\widehat{Q}_l = \frac{\sum_{u \neq v \,:\, \widetilde{\sigma}_l(u) \neq \widetilde{\sigma}_l(v)} (A_l)_{uv}}{|u \neq v \,:\, \widetilde{\sigma}_l(u) \neq \widetilde{\sigma}_l(v)|},$$

where A_l is the adjacency matrix for label l. We then select l^* to be the color that has the largest value of $\widehat{I}_l$. Notice that if $\widehat{P}_l$ and $\widehat{Q}_l$ closely approximate the true P_l and Q_l, then $\widehat{I}_l$ is a measure of how good a color is for clustering: if $\widehat{I}_l$ is large, then the true probabilities are well-separated and therefore provide more information about the community structure than if $\widehat{I}_l$ is small. Naturally, the estimated edge probabilities are reasonably good only if the spectral clustering is reasonably good. Our first proposition makes this statement rigorous. Before stating the proposition, we classify the colors into two sets L_1 and L_1^c as follows:

$$L_1 = \left\{ l : \frac{n(P_l - Q_l)^2}{P_l \vee Q_l} := \frac{\Delta_l^2}{P_l \vee Q_l} \geq 1 \right\}.$$

Notice that a color is "good" or "bad" for clustering depending on the value $\frac{n(P_l - Q_l)^2}{P_l \vee Q_l}$, since it captures how well-separated the edge probabilities are. We may now bound difference between the estimated probabilities and the true probabilities for good and bad colors using Proposition B.1, the formal statement and proof of which may be found in Appendix B.1.

Proposition B.1. *Suppose σ is a clustering with error rate at most γ i.e., $l(\sigma, \sigma_0) \leq \gamma$. Then with probability at least $1 - Ln^{-(3+\delta_p)}$ for a small $\delta_p > 0$, the following event happens for all small enough γ:*

1. *For $l \in L_1$, we have $|\widehat{P}_l - P_l| \leq \eta \Delta_l$ and $|\widehat{Q}_l - Q_l| \leq \eta \Delta_l$.*
2. *For $l \in L_1^c$, we have $|\widehat{P}_l - P_l| \leq \eta \sqrt{\frac{P_l \vee Q_l}{n}}$ and $|\widehat{Q}_l - Q_l| \leq \eta \sqrt{\frac{P_l \vee Q_l}{n}}$.*

In both cases, $\eta = C\sqrt{\gamma \log \frac{1}{\gamma}}$ for an absolute constant C.

We will now work towards getting a good initial clustering with a small error rate γ. In Proposition B.2, we show that if the edge probabilities for a particular label are well-separated, then the spectral clustering output of Algorithm 3.4 is indeed reasonably good. We give a rough statement of the proposition here, and refer to Appendix B.2 for the precise statement and its proof.

Proposition B.2. *If P_l and Q_l satisfy $C_1 \frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2} \leq 1$ for an absolute constant C_1, the output σ^l of the spectral clustering Algorithm 3.4 satisfies the inequality*

$$l(\sigma^l, \sigma_0) \leq C_2 \frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2},$$

for a constant C_2 with probability at least $1 - n^{-4}$.

Thus, if we want to cluster with arbitrarily small error rates γ, we need $\frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2} \to 0$ for at least one well-separated color l. This is precisely what Algorithm 3.3 does when it chooses l^*. We show that l^* satisfies $\frac{(P_{l^*} \vee Q_{l^*})}{n(P_{l^*} - Q_{l^*})^2} \to 0$ in two steps.

First, we combine the results of Propositions B.1 and B.2 and show that for sufficiently well-separated colors, the estimated $\widehat{I}_l$ is close to $\frac{(P_l - Q_l)^2}{P_l \vee Q_l}$. If the probabilities are not well-separated, then we claim that $\widehat{I}_l$ is negligibly small. We give a rough statement of Proposition B.3 here, and refer to Appendix B.3 for the precise statement and its proof.

Proposition B.3. *There is a positive constant C_{test} such that with probability at least $1 - Ln^{-3+\delta_p}$, for a small $\delta_p > 0$, we have*

1. *If $\Delta_l \geq \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$, then $\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} = \Theta\left(\frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}\right)$.*
2. *If $\Delta_l < \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$, then $\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq O\left(\sqrt{\frac{1}{n}}\right)$.*

Next, we then argue that

$$I_L = \Theta\left(\sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l}\right),$$

and use this fact to conclude the existence of a color l for which $\frac{n(P_l - Q_l)^2}{P_l \vee Q_l}$ is arbitrarily large. Using Proposition B.3, we then conclude that Algorithm 3.3 succeeds in choosing a color l^* for which $\frac{n(P_{l^*} - Q_{l^*})^2}{P_{l^*} \vee Q_{l^*}}$ is arbitrarily large. This is stated and proved in the following result:

Proposition B.4. *With probability at least $1 - 2Ln^{-(3+\delta_p)}$, we have that $\frac{n(P_{l^*} - Q_{l^*})^2}{\rho_L^4 (P_{l^*} \vee Q_{l^*})} \geq a_n$ for suitable sequence $a_n \to \infty$.*

Again, we refer to Appendix B.4 for the precise statement and its proof. Let event E_1 denote the successful selection of l^* satisfying the inequality in Proposition B.4. Having selected l^*, we now perform spectral clustering n times, by leaving one vertex out and clustering on the remaining graph. Denote the community assignments obtained in this manner by $\widetilde{\sigma}_u$, for $1 \leq u \leq n$. Note that Proposition B.2 combined with a simple union bound gives that with probability at least $1 - n^{-3}$, all of these clusterings have error rate at most γ, where γ satisfies

$$\gamma \leq C \frac{P_{l^*} \vee Q_{l^*}}{n(P_{l^*} - Q_{l^*})^2},$$

for some constant C. Let us denote this event by E_2. Conditioned on E_1 and E_2 described above, we have that $\gamma \rho_L^4 \to 0$. Thus, we can apply Proposition B.1 on each of the rough clustering $\widetilde{\sigma}_u$'s and show that the conclusion of Proposition B.1 holds simultaneously for all $\widetilde{\sigma}_u$ with probability at least $1 - Ln^{2+\delta_p}$. Furthermore, the η that appears in Proposition B.1 is $\Theta\left(\sqrt{\gamma \log \frac{1}{\gamma}}\right)$, and thus it satisfies satisfies $|\eta \rho_L| \to 0$. Let us denote this event by E_3. Having obtained the n clusterings $\widetilde{\sigma}_u$, we now create $\widehat{\sigma}_u$ by assigning vertex u to an appropriate community in $\widetilde{\sigma}_u$, using the relation from Algorithm 3.5

$$\widehat{\sigma}_u(u) = \arg\max_k \sum_{v\,:\,\widetilde{\sigma}_u(v) = k,\, v \neq u} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l).$$

In Proposition B.5, we show that with high probability, the assignment $\widehat{\sigma}_u(u)$ is "correct". We give a rough statement of the proposition here, and defer the exact statement and its proof to Appendix B.5. We may check that all that all the necessary pre-conditions for applying Proposition B.5, in particular $|\eta \rho_L| \to 0$ are satisfied when event E_3 occurs.

Proposition B.5. *Let $\pi_u \in S_K$ be a permutation such that $l(\sigma_0, \widetilde{\sigma}_u) = d(\sigma_0, \pi_u(\widetilde{\sigma}_u))$. Then with probability at least $1 - \exp\left(-(1-o(1))\frac{n}{\beta K} I_L\right)$, we have $\pi_u^{-1}(\sigma_0(u)) = \widehat{\sigma}_u(u)$.*

Since we haven't yet addressed what π_u is and whether it is unique, we elaborate a bit more on the topic. 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 a permutation such that $d(\pi_u(\widehat{\sigma}_u), \sigma_0) < \frac{n}{8\beta K}$. We argue that π_u is the unique permutation that satisfies $l(\widehat{\sigma}_u, \sigma_0) = d(\pi_u(\widehat{\sigma}_u), \sigma_0)$. We show this by first observing that for any u, the minimum cluster size of the clustering $\widehat{\sigma}_u$ is at least $\frac{n}{\beta K} - (n\gamma + 1) \geq \frac{n(1-\beta K\gamma - \beta K/n)}{\beta K} \geq \frac{n}{2\beta K}$ for small enough γ. Now we use a simple argument in Lemma B.8 which shows that if the Hamming distance $d(\pi(\sigma'), \sigma_0)$ for some permutation $\pi \in S_K$ and some assignment σ' is at most half the size of the smallest cluster ($n/2\beta K$ in this case), then π is the unique permutation that satisfies $l(\sigma', \sigma_0) = d(\pi(\sigma'), \sigma_0)$.

Restating Proposition B.5, we have

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

The final stage of our algorithm is the consensus stage, where we combine the $\widehat{\sigma}_u$'s to produce a $\widehat{\sigma}$ according to the rule

$$\widehat{\sigma}(u) = \arg\max_k |\{v : \widehat{\sigma}_1(v) = k\} \cap \{v : \widehat{\sigma}_u(v) = \widehat{\sigma}_u(u)\}|.$$

This means that if the cluster containing u in $\widehat{\sigma}_u$ has the largest overlap with some cluster k in $\widehat{\sigma}_1$, then u is assigned to cluster k; i.e. we put $\widehat{\sigma}(u) = k$. Note that $\widehat{\sigma}_1$ and $\widehat{\sigma}_u$ both have an error rate of at most $\gamma + 1/n$, and since γ is small, we may conclude that these two assignments are close to each other. Intuitively, one may guess that the permutation $\xi_u \in S_K$ that minimizes $d(\widehat{\sigma}_1, \xi_u(\widehat{\sigma}_u))$ is the consensus function; i.e. $\widehat{\sigma}(u) = \xi_u(\widehat{\sigma}_u(u))$. This is indeed the case, and we show this rigorously. Using the triangle inequality, we have that $l(\widehat{\sigma}_u, \widehat{\sigma}_1) \leq 2\gamma + 2/n < \frac{1}{4\beta K}$ for small enough γ. Thus, the assignments $\widehat{\sigma}_1$ and $\widehat{\sigma}_u$ are very close to each other. We then apply Lemma B.7 on the pair $(\widehat{\sigma}_1, \widehat{\sigma}_u)$ to show that the consensus function ξ_u is the permutation that minimizes $d(\widehat{\sigma}_1, \xi_u(\widehat{\sigma}_u))$.

Observe that $\sigma_0, \widehat{\sigma}_1$, and $\widehat{\sigma}_u$ are all close to each each other, and therefore it is plausible that the permutations π_1 and π_u that minimize $d(\sigma_0, \pi_1(\widehat{\sigma}_1))$ and $d(\sigma_0, \pi_u(\widehat{\sigma}_u))$ respectively, must be related by ξ_u, the permutation that minimizes $d(\widehat{\sigma}_1, \xi_u(\widehat{\sigma}_u))$, by the relation $\xi_u = \pi_1^{-1} \circ \pi_u$. We prove this relation rigorously as well. We know that $d(\sigma_0, \pi_1(\widehat{\sigma}_1)) < \frac{n}{8\beta K}$ and that $d(\sigma_0, \pi_u(\widehat{\sigma}_u)) < \frac{n}{8\beta K}$. Therefore, $d(\widehat{\sigma}_1, \pi_1^{-1}(\pi_u(\widehat{\sigma}_u))) < \frac{n}{4\beta K}$. Since the minimum cluster size of both $\widehat{\sigma}_1$ and $\widehat{\sigma}_u$ is $\frac{n}{2\beta K}$, we may apply Lemma B.8 to conclude that $\pi_1^{-1} \circ \pi_u = \xi_u$. Thus, we have that

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

Using Proposition B.5,

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

where η' is some $o(1)$ sequence. We take then $\eta'' = \eta' + \beta\sqrt{\frac{K}{nI_L}} = o(1)$. First, suppose that

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

In this case, we argue as follows:

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

On the other hand, if we have

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

then

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

This completes the proof of Proposition 5.1.

B Appendix for Proposition 5.1

B.1 Analysis of estimation error of $\widehat{P}_l$ and $\widehat{Q}_l$

Proposition B.1. *Let A_L be a labeled network with true clustering σ_0. Suppose σ is a random initial clustering with error rate at most γ i.e., $l(\sigma,\sigma_0) \leq \gamma$. 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 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 small enough γ:*

1. *For all l such that $P_l \vee Q_l \geq \frac{c}{n}$, if $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq 1$, then*

$$\begin{aligned}
|\widehat{P}_l - P_l| &\leq \eta\Delta_l, \quad \text{and} \\
|\widehat{Q}_l - Q_l| &\leq \eta\Delta_l.
\end{aligned}$$

2. For all l such that $P_l \vee Q_l \geq \frac{c}{n}$, if $\frac{n\Delta_l^2}{P_l \vee Q_l} \leq 1$, then

$$|\widehat{P}_l - P_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}, \quad \text{and}$$
$$|\widehat{Q}_l - Q_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}.$$

In both cases, $\eta = C\sqrt{\gamma \log \frac{1}{\gamma}}$ for an absolute constant C.

Proof. Our proof strategy proceeds by showing that for almost all random graphs generated according to the labeled SBM model, the event described in Proposition B.1 holds for every single assignment in set of all possible assignments with error rate at most γ. For a fixed assignment σ with error rate γ, we call a random graph as "bad graph for σ" if the event in Proposition B.1 does not hold for that graph and that σ. For such a fixed σ, we upper-bound the probability of the set of bad graphs for σ. We then use the union bound over all possible choices of σ and show that the set of bad graphs has probability at most $Ln^{-(3+\delta_p)}$, and thus conclude the proof.

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

$$\begin{aligned}
\log \binom{n}{\gamma n} K^{\gamma n} &= \log\left(\frac{n(n-1)...(n-\gamma n+1)}{(\gamma n)!}\right) + \gamma n \log K \\
&\leq \log\left(\frac{n^{\gamma n} e^{\gamma n}}{(\gamma n)^{\gamma n}} \frac{1}{\sqrt{2\pi\gamma n}}\right) + \gamma n \log K \\
&\leq \log\left(\frac{e^{\gamma n}}{\gamma^{\gamma n}}\right) - \frac{1}{2}\log 2\pi\gamma n + \gamma n \log K \\
&\leq \gamma n \log \frac{e}{\gamma} + \gamma n \log K \\
&\leq \gamma n \log \frac{Ke}{\gamma} \\
&\leq C_1 n\gamma \log \frac{1}{\gamma},
\end{aligned}$$

for a suitable constant C_1 and for all small enough γ. Next, we bound the bias of $\widehat{P}_l$. Our estimator of P_l is

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

The expected value $\mathbb{E}\widehat{P}_l$ is a convex combination of P_l, Q_l, given by

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

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

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

Observe that λ may be bounded from above as follows:

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

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

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 $\Delta_l = |Q_l - P_l|$. A similar calculation may also be performed for the bias of $\widehat{Q}_l$, and taking the worse of the two bounds for λ obtained in each case, we conclude that there exists a constant C_2 such that

$$
\begin{aligned}
|\mathbb{E}\widehat{P}_l - P_l| &\leq C_2 \gamma \Delta_l \quad \text{and} \\
|\mathbb{E}\widehat{Q}_l - Q_l| &\leq C_2 \gamma \Delta_l.
\end{aligned}
$$

To simplify presentation, we define $\eta_1 = C_2 \gamma$ so that

$$
|\mathbb{E}\widehat{P}_l - P_l| \leq \eta_1 \Delta_l. \tag{B.2}
$$

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

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

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 $2\exp(-C_1\gamma n\log\frac{1}{\gamma}-(3+\delta_p)\log n)$. We choose the following t

$$t^2=4\left\{\left(2(P_l\vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k-1)\right)\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)\right\}\vee 4\left\{\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)^2\right\}$$

We now verify that regardless of which term among $\{2(P_l\vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k-1),\, C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\}$ is larger, the probability term is at most $2\exp\left(-\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)\right)$. Let $A=2(P_l\vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k-1)$ and $B=C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n$. We verify our claim by a direct calculation in each of the two cases:

1. Suppose $A\geq B$, then $t^2=4AB$ and the probability term is at most $2\exp\left(-\frac{4AB}{A+\frac{4}{3}\sqrt{AB}}\right)\leq 2\exp\left(-\frac{4AB}{A+\frac{4}{3}A}\right)\leq 2\exp(-B)$.
2. 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 least $1-2\exp\left(-\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)\right)$,

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

Now we derive a more manageable upper bound for t. Note that $t^2=\max(4AB,4B^2)\leq 4(\sqrt{AB}+B)^2$; i.e.,

$$t^2\leq 4\left\{\sqrt{\left(2(P_l\vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k-1)\right)\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)}+\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)\right\}^2.$$

Note that we can without loss of generality assume $\gamma\geq\frac{1}{n}$, since if $\gamma=0$ the result follows trivially. We then have that $\gamma n\log\frac{1}{\gamma}\geq\log n$. Thus, $C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\leq\widetilde{C}_1\gamma n\log\frac{1}{\gamma}$.

$$\begin{aligned}\frac{t}{\sum_{u,v}\mathbf{1}(\sigma(u)=\sigma(v))}&\leq 2\frac{\sqrt{(2(P_l\vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k-1))\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)}+\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)}{\sum_{u,v}\mathbf{1}(\sigma(u)=\sigma(v))}\\&\leq 2\frac{\sqrt{2(P_l\vee Q_l)\widetilde{C}_1\gamma n\log\frac{1}{\gamma}}}{\sqrt{\sum_k \widehat{n}_k(\widehat{n}_k-1)}}+2\frac{\widetilde{C}_1\gamma n\log\frac{1}{\gamma}}{\sum_k \widehat{n}_k(\widehat{n}_k-1)}\\&\overset{(a)}{\leq} 2\frac{\sqrt{2(P_l\vee Q_l)}\sqrt{\widetilde{C}_1\gamma K\log\frac{1}{\gamma}}}{\sqrt{n-K}}+2\frac{\widetilde{C}_1\gamma K\log\frac{1}{\gamma}}{n-K}\end{aligned}$$

$$\overset{(b)}{\leq} 4\sqrt{\frac{P_l \vee Q_l}{n}}\sqrt{\widetilde{C}_1 K\gamma \log\frac{1}{\gamma}} + 4\frac{\widetilde{C}_1 K\gamma\log\frac{1}{\gamma}}{n}.$$

For (a) we used the fact that $\sum_k(\widehat{n}_k - 1)\frac{\widehat{n}_k}{n} = n\sum_k\left(\frac{\widehat{n}_k}{n}\right)^2 - 1 \geq \frac{n}{K} - 1$ because $\frac{\widehat{n}_k}{n}$ sums to 1. For (b), we used the assumption that $n - K \geq \frac{n}{2}$. To further simplify the expression, we have that $P_l \vee Q_l \geq \frac{c}{n}$ and thus, $\frac{1}{n} \leq \frac{1}{\sqrt{c}}\sqrt{\frac{P_l\vee Q_l}{n}}$. Therefore, we have that, with probability at least $1 - \exp(-C_1\gamma n\log\frac{1}{\gamma} - (3+\delta_p)\log n)$,

$$|\widehat{P_l} - \mathbb{E}\widehat{P_l}| \leq \sqrt{\frac{P_l\vee Q_l}{n}}\left(C_1'\sqrt{\gamma\log\frac{1}{\gamma}} + C_2'\gamma\log\frac{1}{\gamma}\right) \tag{B.3}$$

for suitable constants C_1' and C_2'. Using a similar calculation, we may also show that there exist suitable constants C_3' and C_4' such that

$$|\widehat{Q_l} - \mathbb{E}\widehat{Q_l}| \leq \sqrt{\frac{P_l\vee Q_l}{n}}\left(C_3'\sqrt{\gamma\log\frac{1}{\gamma}} + C_4'\gamma\log\frac{1}{\gamma}\right). \tag{B.4}$$

Note that for small enough γ, the term $\sqrt{\gamma\log\frac{1}{\gamma}}$ dominates, and thus we may take choose the right hand side to be

$$\eta_2\sqrt{\frac{P_l\vee Q_l}{n}}$$

where $\eta_2 = C_3\sqrt{\gamma\log\frac{1}{\gamma}}$ for a constant C_3. It is clear that η_2 can be made arbitrarily small by taking γ to be arbitrarily small. Taking the union bound across all clusterings with error γ and across all colors, we have that the probability of (B.3) holding simultaneously for all colors l is at least $1 - Ln^{-(3+\delta_p)}$. Combining equations (B.2) and (B.3), we arrive at the bound

$$|P_l - \widehat{P_l}| \leq \eta_1\Delta_l + \eta_2\sqrt{\frac{P_l\vee Q_l}{n}}. \tag{B.5}$$

If $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq 1$, then

$$|P_l - \widehat{P_l}| \leq \eta_1\Delta_l + \eta_2\Delta_l = (\eta_1+\eta_2)\Delta_l. \tag{B.6}$$

If $\frac{n\Delta_l^2}{P_l\vee Q_l} < 1$, then

$$|P_l - \widehat{P_l}| \leq \eta_1\sqrt{\frac{P_l\vee Q_l}{n}} + \eta_2\sqrt{\frac{P_l\vee Q_l}{n}} = (\eta_1+\eta_2)\sqrt{\frac{P_l\vee Q_l}{n}}. \tag{B.7}$$

Note that $\eta_2 = C_3\sqrt{\gamma\log\frac{1}{\gamma}}$ dominates $\eta_1 = C_1\gamma$ for all small enough γ, and thus we may substitute $\eta_1 + \eta_2$ by $C_4\sqrt{\gamma\log\frac{1}{\gamma}}$ for all small enough γ. □

B.2 Analysis of the spectral clustering algorithm

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

Proposition B.2. *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. Then there exist a constant C such that the output σ of the spectral clustering (algorithm 3.4) with tuning parameter $\mu \geq 32C^2\beta$ and trim threshold $\tau = \overline{d}$ satisfies, under the assumption*

$$256\mu\beta C^2K^3\frac{(p\vee q)}{n(p-q)^2} \leq 1,$$

the inequality

$$l(\sigma, \sigma_0) \leq 64C^2\beta\frac{K^2(p \vee q)}{n(p-q)^2},$$

with probability at least $1 - n^{-C'}$ *for a* $C' > 4$.

Proof. Note that the trimming parameter τ is a random variable, since $\overline{d}$ is random. Since the community sizes are bounded by $n/K\beta$, it is easy to see that we can find constants $C_{d_1} < C_{d_2} = 1$ which depend only on K and β, such that

$$C_{d_1} n(p \vee q) \leq \mathbb{E}\overline{d} \leq C_{d_2} n(p \vee q).$$

Using the multiplicative form of Chernoff's bound, we conclude that with probability at least $1 - \exp(-nC_{\overline{d}})$ for some constant $C_{\overline{d}}$, we must have

$$\frac{C_{d_1}}{2} n(p \vee q) \leq \overline{d} \leq 2C_{d_2} n(p \vee q).$$

We now use state and use the following lemma:

Lemma B.1. *(Lemma 5 of [10]) 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* $C' > 0$ *and* $0 < C_1 < C_2$ *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* $\tau \in [C_1(np_{max} + 1), C_2(np_{max} + 1)]$ *where* $p_{max} = \max_{u \geq v} P_{uv}$.

Remark B.1. Lemma 5 from [10] is stated slightly differently — for any $C' > 0$ there exist constants c, C_1 and C_2 such that the result holds with probability $1 - n^{-C'}$. However, our restatement follows immediately by examining the proof of Lemma 5 in [10].

Using the above with a fixed $C' > 4$, and $C_1 = \frac{C_{d_1}}{2}$ and $C_2 = 2C_{d_2}$ and conclude that there exists a constant C such that the inequality

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

holds for any choice of $\tau \in [C_1, C_2]$ with a probability of $1 - n^{-C'}$. In particular, this inequality will also hold for $\tau = \overline{d}$ with probability $1 - n^{-C'}$. (Note that we may replace $\sqrt{n(p \vee q) + 1}$ by $\sqrt{n(p \vee q)}$, since we assume that $(p \vee q) > c/n$.) We assume that $C \geq 1$, since the inequality in Lemma B.1 holds with C replaced by $\max(1, C)$. Thus, we have that,

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

where (a) 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

$$\sum_{u=1}^{n} \|\widehat{A}_u - P_u\|_2^2 = \|\widehat{A} - P\|_F^2$$

$$\leq K\|\widehat{A} - P\|_2^2$$
$$\leq 4KC^2 n(p \vee q).$$

For simplicity of notation, denote the K possible distinct rows of P by $\mathcal{Z}_i$, for $1 \leq i \leq K$. For a vertex u, denote the row P_u by $\mathcal{Z}(u)$. Note that for $i \neq j$, we have the inequality

$$\|\mathcal{Z}_i - \mathcal{Z}_j\|_2^2 \geq \frac{2}{\beta K}(p-q)^2 n,$$

obtained by considering the lower bound $n/\beta K$ on the cluster sizes. If the $\mathcal{Z}_i$'s are known, then we can cluster v by matching $\widehat{A}_v$ to the closest $\mathcal{Z}_i$. We would make a mistake only if $\|\widehat{A}_u - P_u\|_2^2 = \|\widehat{A}_u - \mathcal{Z}(u)\|_2^2 \geq \frac{1}{\beta K}(p-q)^2 n$. Thus, the number of mistakes we make cannot be larger than

$$\frac{\sum_{u=1}^n \|\widehat{A}_u - P_u\|_2^2}{\frac{1}{\beta K}(p-q)^2 n} \leq \frac{4KC^2 n(p \vee q)}{\frac{1}{\beta K}(p-q)^2 n} \leq \frac{4\beta K^2 C^2 (p \vee q)}{(p-q)^2}.$$

But because we do not know the true $\mathcal{Z}_i$'s, we will construct a set S as a surrogate. This set shall have K points, such that every $\mathcal{Z}_i$ has a point in S which is close to $\mathcal{Z}_i$ and acts as a surrogate for $\mathcal{Z}_i$.

Define a point u as valid if $\|\widehat{A}_u - \mathcal{Z}_i\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$ for some $\mathcal{Z}_i$, not necessarily $\mathcal{Z}(u)$. A point u is declared invalid if the condition is not fulfilled. For a node u, define $\mathcal{Z}^*(u) = \arg\min_{\mathcal{Z}_i} \|\widehat{A}_u - \mathcal{Z}_i\|_2^2$, so $\mathcal{Z}^*(u)$ is the row of the P matrix closest to $\widehat{A}_u$. Notice that if u is valid, then $\|\widehat{A}_u - \mathcal{Z}^*(u)\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. We break up our goal of showing that every $\mathcal{Z}_i$ has a surrogate element in S into the following two claims:

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

First, let us suppose that these two claims are true and see that the proposition follows. Denote the rows of $\widehat{A}$ corresponding to the members of the set S by $\mathcal{S}_i$, where $\mathcal{S}_i$ is the surrogate for $\mathcal{Z}_i$. Define $\mathcal{S}(u)$ to be the surrogate of $\mathcal{Z}(u)$, and denote $\mathcal{S}^*(u) = \arg\min_{\mathcal{S}_i} \|\widehat{A}_u - \mathcal{S}_i\|^2$, that is, the member of S that is closest to $\widehat{A}_u$. We say that a point u is misclassified if $\mathcal{S}^*(u) \neq \mathcal{S}(u)$. 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 misclassified any valid point u.

Since u is a valid point, we know that there is a $\mathcal{Z}_i$ that is close to $\widehat{A}_u$:

$$\|\widehat{A}_u - \mathcal{Z}_i\|^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n.$$

We claim that $\mathcal{S}^*(u) = \mathcal{S}_i$, that is, the point S that is closest to $\widehat{A}_u$ is the surrogate of the row of P that is close to $\widehat{A}_u$. For any $j \neq i$, we have

$$\begin{aligned}\|\widehat{A}_u - \mathcal{S}_j\| &\geq \|Z_i - Z_j\| - \|\mathcal{Z}_j - \mathcal{S}_j\| - \|\mathcal{Z}_i - \widehat{A}_u\| \\ &\geq \sqrt{\frac{2}{\beta K}(p-q)^2 n} - 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} \\ &> 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n}.\end{aligned}$$

Furthermore, we have

$$\|\widehat{A}_u - \mathcal{S}_i\| \leq \|\widehat{A}_u - \mathcal{Z}_i\| + \|\mathcal{S}_i - \mathcal{Z}_i\|$$

$$\leq 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n}.$$

Thus, for any $j \neq i$

$$\|\widehat{A}_u - \mathcal{S}_j\| > \|\widehat{A}_u - \mathcal{S}_i\|,$$

and we must have $\mathcal{S}^*(u) = \mathcal{S}_i$.

Since u has been misclassified, we have that $\mathcal{S}(u) \neq \mathcal{S}^*(u) = \mathcal{S}_i$. Let $\mathcal{S}(u) = \mathcal{S}_j$ and $\mathcal{Z}(u) = \mathcal{Z}_j$. We have the sequence of inequalities:

$$\begin{aligned}
\|\widehat{A}_u - \mathcal{Z}(u)\| &= \|\widehat{A}_u - \mathcal{Z}_j\| \\
&\geq \|\widehat{A}_u - \mathcal{S}_j\| - \|\mathcal{S}_j - \mathcal{Z}_j\| \\
&\geq 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} - \sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} \\
&= \sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n},
\end{aligned}$$

which is the bound we wanted to prove. Thus, the number of mistakes is bounded by

$$\frac{\sum_{u=1}^n \|\widehat{A}_u - P_u\|_2^2}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{4KC^2 n(p \vee q)}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{64\beta K^2 C^2 (p \vee q)}{(p-q)^2},$$

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, by Chernoff's inequality $\overline{d} \leq 2(p \vee q)n$ with probability $1 - \exp(-C_{\overline{d}} n)$. We condition on this event so that if $v \in N(u)$, then $\|\widehat{A}_u - \widehat{A}_v\|^2 \leq 2\mu K^2 (p \vee q)$. 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, $\|\widehat{A}_u - \mathcal{Z}_i\|_2^2 \geq \frac{1}{16\beta K}(p-q)^2 n$ for any $\mathcal{Z}_i$. Let v be a neighbor of u. By triangle inequality, we have

$$\begin{aligned}
\|\widehat{A}_v - \mathcal{Z}(v)\| &\geq \|\widehat{A}_u - \mathcal{Z}(v)\| - \|\widehat{A}_u - \widehat{A}_v\| \\
&\geq \sqrt{\frac{1}{16\beta K}(p-q)^2 n} - \sqrt{2\mu K^2 (p \vee q)} \\
&\overset{(a)}{\geq} \sqrt{\frac{1}{16\beta K}(p-q)^2 n} - \sqrt{\frac{1}{64\beta K}(p-q)^2 n} = \sqrt{\frac{1}{64\beta K}(p-q)^2 n},
\end{aligned}$$

where (a) follows from our assumption coupled with the choice of $C \geq 1$, which essentially states that

$$2\mu K^2 (p \vee q) \leq \frac{1}{128\beta K}(p-q)^2 n < \frac{1}{64\beta K}(p-q)^2 n.$$

Thus, for every neighbor v of u, we must have $\|\widehat{A}_v - P_v\|_2^2 \geq \frac{1}{64\beta K}(p-q)^2 n$. The number of neighbors of u may be bounded by

$$\frac{\sum_{v=1}^n \|\widehat{A}_v - P_v\|_2^2}{\frac{1}{64\beta K}(p-q)^2 n} \leq \frac{4KC^2 n(p \vee q)}{\frac{1}{64\beta K}(p-q)^2 n} \leq \frac{256\beta K^2 C^2 (p \vee q)}{(p-q)^2}.$$

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 $\|\widehat{A}_u - P_u\|_2^2 \le \frac{1}{4}\mu K^2(p\vee q)$. This is because the total error is bounded by $\sum_{u=1}^n \|\widehat{A}_u - P_u\|_2^2 \le 4KC^2 n(p\vee q)$ and thus, the total number of points that violate the condition is at most $\frac{4KC^2n(p\vee q)}{\frac{1}{4}\mu K^2(p\vee q)} \le \frac{n}{2\beta K}$ by the assumption that $\mu \ge 32C^2\beta$.

For two points u and v in the same cluster satisfying $\|\widehat{A}_w - P_w\|_2^2 \le \frac{1}{4}\mu K^2(p\vee q)$ for $w \in \{u, v\}$, we also have $\|\widehat{A}_u - \widehat{A}_v\|_2^2 \le \mu K^2(p\vee q)$ by triangle inequality. Thus, in every cluster, there exists a point u such that $N(u) \ge \frac{n}{2\beta K} \ge \frac{1}{\mu}\frac{n}{K}$, since $\mu \ge 32C^2\beta > 2\beta$ by our choice of $C > 1$.

Suppose at iteration r the set S consists of points $s_1, \ldots, s_r$ where $1 \le r < K$, and suppose for contradiction that s_{r+1} is such that $\mathcal{Z}(s_{r+1}) = \mathcal{Z}(s_i)$ for some $1 \le i \le r$. Since s_i and s_{r+1} are both valid points, we have by triangle inequality

$$\|\widehat{A}_{s_{r+1}} - \widehat{A}_{s_i}\| \le \frac{1}{4\beta K}(p-q)^2 n.$$

On the other hand, because S does not yet have K nodes, there must be some $\mathcal{Z}_j$ which does not have a surrogate in S. The cluster that corresponds to $\mathcal{Z}_j$ must, by our neighborhood size analysis, contain a node u such that $N(u) \ge \frac{1}{\mu}\frac{n}{K}$ and that

$$\|\widehat{A}_u - \mathcal{Z}_j\|_2^2 \le \frac{1}{4}\mu K^2(p\vee q) \le \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n,$$

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

Since $\mathcal{Z}_j \ne \mathcal{Z}(s_i)$ for any $1 \le i \le r$, we have $\|\mathcal{Z}_j - \mathcal{Z}(s_i)\|_2^2 \ge 2\frac{1}{\beta K}(p-q)^2 n$ for all $1 \le i \le r$. By claim 1, we have that all s_i's are valid, and thus

$$\|\widehat{A}_{s_i} - \mathcal{Z}(s_i)\| \le \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n.$$

So we have that, by triangle inequality, that $\|\widehat{A}_u - \widehat{A}_{s_i}\|_2^2 \ge \frac{1}{\beta K}(p-q)^2 n$ for all $s_i \in S$. This is a contradiction because u is farther away from every point in S than s_{r+1}, and thus our assumption that $\mathcal{Z}(s_{r+1}) = \mathcal{Z}(s_i)$ must be invalid. □

B.3 Analysis of the Initialization Scheme

Proposition B.3. *Suppose that $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ 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 $1 - Ln^{-3+\delta_p}$, for a small $\delta_p > 0$, we have*

1. *For all colors l satisfying $P_l \vee Q_l > c/n$ and $\Delta_l \ge \sqrt{C_{test}}\sqrt{\frac{P_l\vee Q_l}{n}}$, the following holds:*

$$C_1\frac{|P_l - Q_l|}{\sqrt{P_l\vee Q_l}} \le \frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l\vee\widehat{Q}_l}} \le C_2\frac{|P_l - Q_l|}{\sqrt{P_l\vee Q_l}}, \tag{B.8}$$

   for some absolute constants C_1 and C_2.

2. *For all colors satisfying $P_l \vee Q_l > c/n$ and $\Delta_l < \sqrt{C_{test}}\sqrt{\frac{P_l\vee Q_l}{n}}$, the following holds:*

$$\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l\vee\widehat{Q}_l}} \le C\sqrt{\frac{1}{n}} \tag{B.9}$$

   for a suitable absolute constant C.

Proof. Recall from Proposition B.1 that given a clustering with error-rate γ, and under the assumptions $P_l \vee Q_l > \frac{c}{n}$ and $\Delta_l^2 \geq \frac{P_l \vee Q_l}{n}$ we have that the estimated probabilities $\widehat{P_l}$ and $\widehat{Q_l}$ satisfy

$$|\widehat{P_l} - \widehat{P}| \leq \eta \Delta_l \quad \text{and}$$
$$|\widehat{Q_l} - \widehat{Q}| \leq \eta \Delta_l,$$

with probability $1 - n^{-(3+\delta_p)}$, where η is as in Proposition B.1. We first pick a value of γ such that $\eta < \frac{1}{4}$. Notice that we can do so because as $\eta \to 0$ as $\gamma \to 0$. We will now ensure that the error-rate γ as obtained from Proposition B.2 matches our current choice of γ. Recall that Proposition B.2 states that under the assumptions $P_l \vee Q_l > \frac{c}{n}$ and

$$C_1 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \leq 1$$

for some constant C_1, we have

$$l(\sigma, \sigma_0) \leq C_2 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2}$$

for some constant C_2. More details concerning the constants C_1 and C_2 may be found in the statement of Proposition B.2. For the purpose of this proof, it is enough to note that if C_{test} is chosen large enough, then the following can be made to hold:

$$C_1 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \leq \frac{C_1}{C_{test}} < 1, \quad \text{and}$$
$$l(\sigma, \sigma_0) \leq C_2 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \leq \frac{C_2}{C_{test}} < \gamma$$

for all colors l satisfying $\Delta_l \geq \sqrt{C_{test}} \sqrt{\frac{P_l \vee Q_l}{n}}$. We pick C_{test} to be the larger of 1 and this value, so that the results from Proposition B.1 may also be applied for all colors satisfying the bound $\Delta_l \geq \sqrt{C_{test}} \sqrt{\frac{P_l \vee Q_l}{n}}$. In the rest of the proof, we shall use the bounds from Proposition B.1 and the fact that $|\eta| < 1/4$. 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:

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

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

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

This completes the proof for all colors satisfying $\Delta_l^2 \geq C_{test}\frac{P_l \vee Q_l}{n}$.

Now we move on to the second claim where we suppose $\Delta_l^2 \leq C_{test}\frac{P_l \vee Q_l}{n}$. Note that this does not necessarily imply $\Delta_l^2 \leq \frac{P_l \vee Q_l}{n}$, since $C_{test} \geq 1$. However, we may still take the maximum of the bounds provided in Proposition B.1, to conclude that with probability $1 - Ln^{-(3+\delta_p)}$,

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

where $|\eta| \leq 1/4$. The condition $\Delta_l^2 \leq C_{test}\frac{P_l \vee Q_l}{n}$ implies $\sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}} \geq \Delta_l$, which may be substituted in the bound above to obtain

$$|\widehat{P}_l - P_l| \leq \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}},$$

where we used the fact that $C_{test} \geq 1$. A similar bound also holds for $|\widehat{Q}_l - Q_l|$ using the same reasoning. Putting these together, we have that

$$\begin{aligned}
|\widehat{P}_l - \widehat{Q}_l| &= |\widehat{P}_l - P_l + P_l - Q_l + Q_l - \widehat{Q}_l| \\
&\leq \Delta_l + |\widehat{P}_l - P_l| + |\widehat{Q}_l - Q_l| \\
&\leq \Delta_l + \frac{\sqrt{C_{test}}}{2}\sqrt{\frac{P_l \vee Q_l}{n}} \\
&\leq \frac{3}{2}\sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}.
\end{aligned}$$

To bound the denominator term $\sqrt{\widehat{P}_l \vee \widehat{Q}_l}$, we have that $\widehat{P}_l \geq P_l - \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}}$ and $\widehat{Q}_l \geq Q_l - \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}}$. Suppose without loss of generality that $P_l \geq Q_l$ so that $P_l = (P_l \vee Q_l)$. Then we have

$$\widehat{P}_l \vee \widehat{Q}_l \geq (P_l \vee Q_l) - \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}} \geq CP_l \vee Q_l$$

for some constant C, where we used the assumption that $P_l \vee Q_l \geq \frac{c}{n}$. Thus, we have that

$$\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \leq C\sqrt{\frac{1}{n}}$$

for an appropriate constant C that depends on C_{test} and c. □

B.4 Choosing label l^*

Proposition B.4. *Let l^* be the color chosen by the initialization algorithm. Let $a_n = \frac{nI_L}{L\rho_L^4}$ and assume that $a_n \to \infty$. For large enough n, with probability at least $1 - 2Ln^{-(3+\delta_p)}$, we have that $\frac{n(P_{l^*}-Q_{l^*})^2}{(P_{l^*}\vee Q_{l^*})\rho_L^4} \geq C \cdot a_n$ for some constant C.*

Proof. Let C_{test} be the constant in Proposition B.3. As shown in the proof of Lemma B.6, we have that I_L is of the same order as $\sum_{l=1}^{L} \frac{\Delta_l^2}{P_l \vee Q_l}$, and therefore, there must exist some color l_n (we leave the dependency on n implicit and denote it just by l) such that $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq C\frac{nI_L}{L} = Ca_n\rho_L^4$ for some constant C. 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 $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq Ca_n\rho_L^4$ 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,

$$\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \overset{(a)}{\leq} \frac{|\widehat{P}_{l^*} - \widehat{Q}_{l^*}|}{\sqrt{\widehat{P}_{l^*} \vee \widehat{Q}_{l^*}}} \leq C'\sqrt{\frac{1}{n}},$$

where (a) follows from the definition of l^*. But, because l satisfies $\Delta_l \geq C_{test}\sqrt{\frac{P_l \vee Q_l}{n}}$, we also have

$$\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \geq \frac{1}{\sqrt{5}}\frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}, \geq C''\sqrt{a_n\rho_L^4\frac{1}{n}}$$

Since $a_n \to \infty$ and $\rho_L \geq 1$, we contradict the $C'\frac{1}{\sqrt{n}}$ upper bound derived earlier.

Step 2: Again, let l be a color such that $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq Ca_n\rho_L^4$. By Proposition B.3 and the definition of l^*, we obtain

$$\begin{aligned}
\frac{|P_{l^*} - Q_{l^*}|}{\sqrt{P_{l^*} \vee Q_{l^*}}} &\geq \frac{\sqrt{3}}{3}\frac{|\widehat{P}_{l^*} - \widehat{Q}_{l^*}|}{\sqrt{\widehat{P}_{l^*} \vee \widehat{Q}_{l^*}}} \\
&\geq \frac{\sqrt{3}}{3}\frac{|\widehat{P}_l - \widehat{Q}_l|}{\sqrt{\widehat{P}_l \vee \widehat{Q}_l}} \\
&\geq \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}\frac{1}{\sqrt{15}} \\
&\geq C'\sqrt{a_n\rho_L^4\frac{1}{n}},
\end{aligned}$$

for an appropriate constant C'. The conclusion thus follows. □

B.5 Analysis of probability of error for a single node

Proposition B.5. *Let node u be arbitrarily fixed, let $\widetilde{\sigma}_u$ be the output of Algorithm 3.3, and let $\pi_u \in S_K$ such that*

$$l(\sigma_0, \widetilde{\sigma}_u) = d(\sigma_0, \pi_u(\widetilde{\sigma}_u)),$$

where both l and d are taken with respect to the set $\{1, 2, \ldots, n\}\backslash\{u\}$. We assume that the conditions of Proposition 5.1 hold. Conditioning on the event that the error rate γ of $\widetilde{\sigma}_u$ satisfies $\rho_L^4\gamma \to 0$, and

also on the event that the result of Proposition B.1 holds with a sequence η that satisfies $\eta\rho_L^2 \to 0$, 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:

$$\pi_u^{-1}(\sigma_0(u)) = \arg\max_k \sum_{v:\widetilde{\sigma}_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$$

Proof. Throughout the proof, we assume n is large enough so that $\frac{1}{2}\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 \leq \frac{1}{2}$. Suppose without loss of generality that $\sigma_0(u) = 1$. We misclassify u in community k if

$$\sum_{v:\widetilde{\sigma}_u(v)=k} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l) \geq \sum_{v:\widetilde{\sigma}_u(v)=1} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l). \tag{B.10}$$

This can be restated as

$$\sum_{v:\widetilde{\sigma}_u(v)=k} \overline{A}_{uv} - \sum_{v:\widetilde{\sigma}_u(v)=1} \overline{A}_{uv} \geq 0, \tag{B.11}$$

where $\overline{A}_{uv} \equiv \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$. Note that the edges from u are independent of the clustering $\widetilde{\sigma}_u$, since this clustering was obtained by running the algorithm with vertex u excluded.

Define $m_1 = |\{v : \widetilde{\sigma}_u(v) = 1\}|$ and $m_k = |\{v : \widetilde{\sigma}_u(v) = k\}|$ as the size of clusters m_1, m_k under σ_u. Define $m_1' = \{v : \widetilde{\sigma}_u(v) = 1, \sigma_0(v) = 1\}$ as the points correctly clustered by σ_u, and $m_k' = \{v : \widetilde{\sigma}_u(v) \neq 1, \sigma_0(v) = k\}$ as a loose definition of points correctly classified by $\widetilde{\sigma}_u$ in community k. With these definitions, the probability of the bad event Equation B.11 is the probability of the event

$$\left(\sum_{i=1}^{m_k'} \widetilde{Y}_i + \sum_{i=1}^{m_k - m_k'} \widetilde{X}_i + \right) - \left(\sum_{i=1}^{m_1'} \widetilde{X}_i + \sum_{i=1}^{m_1 - m_1'} \widetilde{Y}_i\right) \geq 0,$$

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. (For simplicity, we abuse notation here by using the notation $\widetilde{Y}_i$ and $\widetilde{X}_i$ in both bracketed terms. These random variables are not the same, of course, but are independent and identical copies.) This is the same as the probability of the event

$$\exp\left(t\left(\sum_{i=1}^{m_k'} \widetilde{Y}_i + \sum_{i=1}^{m_k - m_k'} \widetilde{X}_i - \sum_{i=1}^{m_1'} \widetilde{X}_i - \sum_{i=1}^{m_1 - m_1'} \widetilde{Y}_i\right)\right) \geq 1$$

We bound the probability of this event as follows:

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

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

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

$$= \left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right)^{m_k - m_k'} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l}\right)^{m_1 - m_1'} \tag{B.12}$$

$$\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{m_k} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{m_1}. \tag{B.13}$$

We will bound term B.12 and B.13 separately. Loosely speaking, we will show that term B.12 is bounded in magnitude by $\exp(o(I_L)\frac{n}{K})$ and that term B.13 is bounded by $\exp(-\frac{n}{\beta K}(1 + o(1))I_L)$.

Bound for Term B.12. Establishing this bound requires a number of lemmas establishing bounds on the intermediate terms that appear in the computation. In particular, we use the bounds from Lemma B.5, Lemma B.4, and Lemma B.6 in the following sequence of inequalities:

$$\begin{aligned}
\left|1 - \frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right| &= \left|\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} (P_l - Q_l)}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l}\right| \\
&\overset{(a)}{\leq} \frac{8}{\sum_l \sqrt{P_l Q_l}} \left|\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} (P_l - Q_l)\right| \\
&\overset{(b)}{\leq} 16 \left|\sum_l \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right)(P_l - Q_l)\right| \\
&\leq 16 \left|\sum_{l \in L_1} \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right)(P_l - Q_l)\right| + 16 \sum_{l \notin L_1} \left|\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right| |P_l - Q_l| \\
&\overset{(c)}{\leq} 16 \sum_{l \in L_1} \frac{\Delta_l^2}{Q_l}(1 + \eta') + \sum_{i \notin L_1} 32 \rho_L \frac{\Delta_l}{\sqrt{n(P_l \vee Q_l)}} \\
&\leq 16 \rho_L \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l}(1 + \eta') + \sum_{i \notin L_1} 32 \rho_L \frac{\Delta_l}{\sqrt{n(P_l \vee Q_l)}} \\
&\overset{(d)}{\leq} C I_L \rho_L (1 + \eta') + C' \rho_L \frac{L}{n} \\
&\overset{(e)}{\leq} C \rho_L I_L.
\end{aligned}$$

In (a), we used the bound Lemma B.5, which states that

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

In (b), we the fact that $\sum_l \sqrt{P_l Q_l} \to 1$, and thus for large enough n it will exceed $1/2$. In (c), we employ Lemma B.4, which provides the appropriate bounds for the term $\left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right)$. Here η' is

a $o(1)$ sequence. Inequality (d) follows from Lemma B.6, which shows that the first term in the previous step is $\Theta(I_L)$, and from the definition of the set L_1^c. Finally, inequality (e) follows from the assumption that $\frac{I_L n}{L\rho_L^4} \to \infty$ (note that $\rho_L \geq 1$) and by appropriately redefining η'. Identical analysis shows that

$$\left| 1 - \frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l} \right| \leq C\rho_L I_L.$$

Now, we note that $|x| \leq \exp(|1-x|)$. Therefore, term B.12 can be bounded as

$$\begin{aligned}\left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l} \right)^{m_k - m_k'} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l} \right)^{m_1 - m_1'} &\leq \exp(C\rho_L I_L (m_k - m_k' + m_1 - m_1')) \\ &\leq \exp(C I_L \rho_L \gamma n).\end{aligned}$$

Since $\gamma\rho_L = o(1)$, we conclude that term B.12 is bounded by $\exp\left(\frac{n}{K} o(I_L)\right)$, as desired.

Bound for Term B.13. Define $\widehat{I} = -\log\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)$. With this definition,

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

We claim that the following three statements are true.

1. $m_1, m_k \geq \frac{n}{\beta K}(1 - \beta K \gamma)$
2. $\widehat{I} \geq I_L(1 + o(1))$, where the notation $1 + o(1)$ stands for $1 + \eta'$ for an $o(1)$ sequence η'
3. $\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)^{\frac{m_1-m_k}{2}} = \exp\left(\frac{n}{K} o(I_L)\right)$

Let us first suppose that these statements are true and see that term B.13 can be bounded.

$$\begin{aligned}\exp(-\widehat{I})^{\frac{m_1+m_k}{2}} \left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l \right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l \right)^{\frac{m_1-m_k}{2}} &\leq \exp\left(-I_L(1+o(1)) \frac{n}{\beta K} \cdot (1 - \beta K \gamma) + \frac{n}{K} o(I_L) \right) \\ &\leq \exp\left(-(1+o(1)) \frac{n}{\beta K} I_L \right),\end{aligned}$$

where last inequality holds because $\gamma = o(1)$. We will prove each of the three claims in the remainder of the proof.

Claim 1: This is straightforward. The labeling $\widetilde{\sigma}_u$ has at most γn errors and therefore, $m_1 \geq m_1' \geq \frac{n}{\beta K} - \gamma n$. A similar argument also works for m_k.

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

$$\widehat{I} - I_L = -\log \frac{\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)}{\left(\sum_l \sqrt{P_l Q_l}\right)^2}. \tag{B.14}$$

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

$$T_{l,l'} = \frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}\frac{P_{l'}}{\widehat{P}_{l'}}\frac{\widehat{Q}_{l'}}{Q_{l'}}.$$

Continuing,

$$\begin{aligned}
\widehat{I} - I_L &= -\log\left(1 + \frac{\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)}{\left(\sum_l \sqrt{P_l Q_l}\right)^2}\right) \\
&\geq -\log\left(1 + 4\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)\right) \quad \text{(assuming that } \sum_l \sqrt{P_l Q_l} \geq 1/2\text{)} \\
&\geq -4\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right). \qquad \text{(B.15)}
\end{aligned}$$

We proceed by first bounding $|T_{l,l'} - 1|$:

$$\begin{aligned}
|T_{l,l'} - 1| &= \left|\frac{\widehat{P}_l}{P_l}\frac{Q_l}{\widehat{Q}_l}\frac{P_{l'}}{\widehat{P}_{l'}}\frac{\widehat{Q}_{l'}}{Q_{l'}} - 1\right| \\
&= \left|\left(1 - \frac{P_l - \widehat{P}_l}{P_l}\right)\left(1 - \frac{\widehat{Q}_l - Q_l}{\widehat{Q}_l}\right)\left(1 - \frac{\widehat{P}_{l'} - P_{l'}}{\widehat{P}_{l'}}\right)\left(1 - \frac{Q_{l'} - \widehat{Q}_{l'}}{Q_{l'}}\right) - 1\right| \\
&\overset{(a)}{\leq} 2\left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{\widehat{Q}_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{\widehat{P}_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}}\right) \\
&\overset{(b)}{\leq} 4\left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{P_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}}\right),
\end{aligned}$$

where (a) and (b) follow from Lemma B.3, which says that $\frac{|P_l - \widehat{P}_l|}{P_l}$ and $\frac{|P_l - \widehat{P}_l|}{\widehat{P}_l}$ both go to 0, and at the same rate since $\widehat{P}_l = \Theta(P_l)$. The same also holds true for Q_l and $\widehat{Q}_l$. 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 $l < l'$,

$$|T_{l,l'} - 1| \leq 8\left(\frac{|\widehat{P}_l - P_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l}\right).$$

By Proposition B.1, we have that, for $l \in L_1$,

$$\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l} \leq \eta\Delta_l \left(\frac{1}{P_l} + \frac{1}{Q_l}\right) \leq \frac{\eta\Delta_l}{P_l \vee Q_l} \cdot 2\rho_L \leq \eta' \frac{\Delta_l}{P_l \vee Q_l}.$$

Similarly, for $l \notin L_1$,

$$\frac{|P_l - \widehat{P}_l|}{P_l} \leq \eta\sqrt{\frac{P_l \vee Q_l}{nP_l^2}} = \eta\frac{P_l \vee Q_l}{P_l}\sqrt{\frac{1}{n(P_l \vee Q_l)}} \leq \eta\rho_L\sqrt{\frac{1}{n(P_l \vee Q_l)}} \leq \eta'\sqrt{\frac{1}{n(P_l \vee Q_l)}},$$

and likewise for the $\frac{|\widehat{Q}_l - Q_l|}{Q_l}$ term. We plug these bounds into the previous derivation and obtain

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

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

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

Continuing on from equation B.15, we have that

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

where (a) follows from the following facts: We have $\sum_{l'} \sqrt{P_{l'} Q_{l'}} \leq 1$. Furthermore, Lemma B.6 states that $\sum_{l\in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} = \Theta(I_L)$, and finally our assumptions imply that $\sum_{l\notin L_1} \frac{1}{n} \leq \frac{L}{n} = o(I_L)$. This proves claim 2.

Claim 3. We rewrite the term in claim 3 as follows:

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

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

Assume that $m_k \geq m_1$. The reverse case can be analyzed in the identical manner. We can rewrite the term as

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

Note that $\sum_l \sqrt{P_l Q_l} \to 1$, and thus Lemma B.5 implies that the denominators are $\Theta(1)$. To bound the numerator term, we apply Lemma B.4:

$$\begin{aligned}
\left| \sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} (Q_l - \widehat{Q}_l) \right| &= \left| \sum_l \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right) (Q_l - \widehat{Q}_l) \right| \\
&\leq \left| \sum_{l \in L_1} \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right) (Q_l - \widehat{Q}_l) \right| + \left| \sum_{l \notin L_1} \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right) (Q_l - \widehat{Q}_l) \right| \\
&\overset{(a)}{\leq} \left| \sum_{l \in L_1} \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right) \eta \Delta_l \right| + \left| \sum_{l \notin L_1} \left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right) \eta \sqrt{\frac{P_l \vee Q_l}{n}} \right| \\
&\overset{(b)}{\leq} \sum_{l \in L_1} \eta \frac{\Delta_l^2}{Q_l} + \sum_{l \notin L_1} \eta \rho_L \frac{1}{n} \\
&\leq \eta \rho_L \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} + \sum_{l \notin L_1} \eta \rho_L \frac{1}{n} \\
&\overset{(c)}{\leq} \eta' I_L + \eta' \frac{L}{n} \\
&\overset{(d)}{\leq} \eta' I_L.
\end{aligned}$$

In the above sequence of inequalities, step (a) follows from Proposition B.1. Step (b) follows from Lemma B.4. Step (c) follows from Lemma B.6 and the assumption that $\eta \rho_L \to 0$. Step (d) follows from our assumption of $\frac{L}{n} = o(I_L)$. Thus, we obtain

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

proving claim 3.

Combining bounds for terms B.12 and B.13 : Multiplying the bounds for terms B.12 and B.13 shows that the probability of misclassifying 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 $k \neq 1$ completes the proof. □

B.6 Lemmas for Proposition 5.1

Lemma B.2. *Let L, P_l, Q_l, ρ_L, and I_L satisfy the assumptions in Proposition 5.1. Define the new probabilities of edge labels as follows:*

$$P'_l := P_l(1-\delta) + \frac{\delta}{L}, \quad \text{and } Q'_l := Q_l(1-\delta) + \frac{\delta}{L}, \tag{B.16}$$

for all $1 \leq l \leq L$, where $\delta = \frac{cL}{n}$ for a constant $c > 0$. Let I'_L be the Renyi divergence between P'_l and Q'_l. Then for all large enough n, we have that $P'_l, Q'_l > \frac{c}{n}$ for all $1 \leq l \leq L$, and

$$I'_L = I_L(1+o(1)). \tag{B.17}$$

Proof. We know that

$$\begin{aligned} I_L &= -2\log \sum_{i=1}^{L} \sqrt{P_l Q_l} \\ &= -2\log\left(1 - \frac{1}{2}\sum_{i=1}^{L}(\sqrt{P}_l - \sqrt{Q}_l)^2\right) \\ &= \left(\sum_{i=1}^{L}(\sqrt{P}_l - \sqrt{Q}_l)^2\right)(1+o(1)). \end{aligned}$$

Similarly, we also have $I'_L = \left(\sum_{i=1}^{L}(\sqrt{P'_l} - \sqrt{Q'_l})^2\right)(1+o(1))$, and it is enough to show that

$$\left(\sum_{i=1}^{L}(\sqrt{P}_l - \sqrt{Q}_l)^2\right) = \left(\sum_{i=1}^{L}(\sqrt{P'_l} - \sqrt{Q'_l})^2\right)(1+o(1)).$$

Equivalently, we want to show that

$$\left|\sum_{l=1}^{L}\frac{\Delta_l^2}{(\sqrt{P}_l + \sqrt{Q}_l)^2} - \sum_{l=1}^{L}\frac{(\Delta'_l)^2}{(\sqrt{P'_l} + \sqrt{Q'_l})^2}\right| = o(I_L).$$

We consider two cases: $\rho_L = \omega(1)$ and $\rho_L = \Theta(1)$. If $\rho_L = \omega(1)$, choose $a = \frac{nI_L}{\rho_L L}$. If $\rho_L = \Theta(1)$, choose $a = o\left(\frac{nI_L}{L}\right)$ such that $a \to \infty$. Note that in both cases, we have $\frac{a}{\rho_L} \to \infty$ and $\frac{aL}{n} = o(I_L)$. We now break up the set of colors into two groups: G_1 contains colors which satisfy $P_l \vee Q_l \leq \frac{a}{n}$, and $G_2 = G_1^c$.

For colors in G_1, we have $\Delta_l \leq \frac{a}{n}$. Thus,

$$\frac{\Delta_l^2}{(\sqrt{P}_l + \sqrt{Q}_l)^2} \leq \Delta_l \leq \frac{a}{n},$$

and

$$\frac{(\Delta'_l)^2}{(\sqrt{P'_l} + \sqrt{Q'_l})^2} \leq \Delta'_l = (1-\delta)\Delta_l \leq (1-\delta)\frac{a}{n}.$$

Thus, we have

$$\begin{aligned}\left|\sum_{l\in G_1}\frac{\Delta_l^2}{(\sqrt{P}_l+\sqrt{Q}_l)^2}-\sum_{l\in G_1}\frac{(\Delta_l')^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2}\right| &\le \sum_{l\in G_1}\left(\frac{a}{n}+(1-\delta)\frac{a}{n}\right)\\ &\le \frac{2aL}{n}\\ &= o(I_L).\end{aligned}$$

For colors in G_2, we may write a similar expression

$$\begin{aligned}&\left|\sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P}_l+\sqrt{Q}_l)^2}-\sum_{l\in G_2}\frac{(\Delta_l')^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2}\right|\\ &=\sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P}_l+\sqrt{Q}_l)^2}\left|1-(1-\delta)^2\frac{(\sqrt{P}_l+\sqrt{Q}_l)^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2}\right|.\end{aligned}$$

We analyze the term inside the absolute value as follows.

$$\begin{aligned}\frac{(\sqrt{P}_l+\sqrt{Q}_l)^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2} &= \left(\frac{\sqrt{P}_l+\sqrt{Q}_l}{\sqrt{P}_l\sqrt{\frac{P_l'}{P_l}}+\sqrt{Q}_l\sqrt{\frac{Q_l'}{Q_l}}}\right)^2\\ &= \left(\frac{\sqrt{P}_l+\sqrt{Q}_l}{\sqrt{P}_l\sqrt{1-\delta+\frac{c}{nP_l}}+\sqrt{Q}_l\sqrt{1-\delta+\frac{c}{nQ_l}}}\right)^2.\end{aligned}$$

Since $P_l\vee Q_l>\frac{a}{n}$,

$$\begin{aligned}&\frac{1}{n(P_l\vee Q_l)}<\frac{1}{a}=o(1),\quad \text{and}\\ &\frac{1}{nP_l}\le\frac{\rho_L}{n(P_l\vee Q_l)}<\frac{\rho_L}{a}=o(1).\end{aligned}$$

Furthermore, since $\delta=o(1)$,

$$\begin{aligned}\left(\frac{\sqrt{P}_l+\sqrt{Q}_l}{\sqrt{P}_l\sqrt{1-\delta+\frac{c}{nP_l}}+\sqrt{Q}_l\sqrt{1-\delta+\frac{c}{nQ_l}}}\right)^2 &= \left(\frac{\sqrt{P}_l+\sqrt{Q}_l}{\sqrt{P}_l(1+o(1))+\sqrt{Q}_l(1+o(1))}\right)^2\\ &= 1+o(1).\end{aligned}$$

Hence, we may conclude that

$$\begin{aligned}\sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P}_l+\sqrt{Q}_l)^2}\left|1-(1-\delta)^2\frac{(\sqrt{P}_l+\sqrt{Q}_l)^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2}\right| &= \sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P}_l+\sqrt{Q}_l)^2}o(1)\\ &= o(I_L).\end{aligned}$$

Combining the result for colors in G_1 and G_2, we conclude that $I_L'=(1+o(1))I_L$. □

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

Lemma B.3. *Let l be any color and suppose that $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ where $\rho_L > 1$; suppose also that $P_l, Q_l \ge \frac{c}{n}$ for some absolute constant c. Let us condition on the event that the conclusion of proposition B.1 holds with a sequence η such that $\eta\rho_L^2 \to 0$. Then we have that*

$$\max_l \frac{|\widehat{P}_l - P_l|}{P_l} \to 0 \quad \text{and} \quad \max_l \frac{|\widehat{Q}_l - Q_l|}{Q_l} \to 0.$$

In particular, for small enough η, we have that $\frac{1}{2}P_l \le \widehat{P}_l \le 2P_l$ 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 B.1, we have that either $|\widehat{P}_l - P_l| \le \eta\Delta_l$ or $|\widehat{P}_l - P_l| \le \eta\sqrt{\frac{P_l \vee Q_l}{n}}$ holds. Let us suppose $|\widehat{P}_l - P_l| \le \eta\Delta_l$ first. Then,

$$\frac{|\widehat{P}_l - P_l|}{P_l} \le \eta\frac{\Delta_l}{P_l} \le \eta(1+\rho_L) \to 0.$$

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

$$\frac{|\widehat{P}_l - P_l|}{P_l} \le \eta\sqrt{\frac{\rho_L}{P_l n}} \le \eta\sqrt{\rho}_L\sqrt{\frac{1}{c}} \to 0,$$

where we use that $\frac{P_l \vee Q_l}{P_l}$ is at most ρ_L. □

Lemma B.4. *Suppose that $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ for $\rho_L > 1$ and that $P_l, Q_l \ge \frac{c}{n}$ for some absolute constant c. Let us condition on the event that the conclusion of Proposition B.1 holds with a sequence η such that $\eta\rho_L^2 \to 0$. The following are true:*

1. *For all l satisfying $n\frac{\Delta_l^2}{P_l \vee Q_l} \ge 1$, we have*

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

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

2. *For all l satisfying $n\frac{\Delta_l^2}{P_l \vee Q_l} < 1$, we have that*

$$\left|\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1\right| \le 2\rho_L\frac{1}{\sqrt{n(P_l \vee Q_l)}}.$$

The same bounds also hold for $\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} - 1$ by symmetry.

Proof. First, suppose that l satisfies $n\frac{\Delta_l^2}{P_l \vee Q_l} \ge 1$. We note that by Lemma B.3, we have that $\frac{\widehat{Q}_l - Q_l}{Q_l} = \eta'$ 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 $\max_l |\eta\rho_L| = o(1)$.

$$\frac{\widehat{P}_l}{\widehat{Q}_l} - 1 = \frac{\widehat{P}_l - P_l + P_l}{\widehat{Q}_l - Q_l + Q_l} - 1$$

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

In (a), we used the fact that $|\widehat{P}_l - P_l| \le \eta\Delta_l$ and $|\widehat{Q}_l - Q_l| \le \eta\Delta_l$ by the conclusion of Proposition B.1. Using the inequality

$$|\sqrt{x} - 1| \le |x - 1|$$

now completes the proof of the first case.

The proof of the second case is almost identical. Let us assume that l is such that $n\frac{\Delta_l^2}{P_l \vee Q_l} < 1$. In this case, we have that

$$
\begin{aligned}
|\widehat{P}_l - P_l| &= \eta\sqrt{\frac{P_l \vee Q_l}{n}} \quad \text{and} \\
|\widehat{Q}_l - Q_l| &= \eta\sqrt{\frac{P_l \vee Q_l}{n}},
\end{aligned}
$$

where $\max_l |\eta\rho_L^2| \to 0$. By Lemma B.3, we have $\frac{\widehat{Q}_l - Q}{Q_l} = \eta'$ 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. We may check that the following equalities hold:

$$
\begin{aligned}
\frac{\widehat{P}_l}{\widehat{Q}_l} - 1 &= \left(\frac{P_l}{Q_l} + \frac{\widehat{P}_l - P_l}{Q_l}\right)\left(1 - \frac{\widehat{Q}_l - Q_l}{Q_l}(1+\eta')\right) - 1 \\
&= \frac{P_l}{Q_l} + \frac{\widehat{P}_l - P_l}{Q_l} - \frac{P_l}{Q_l}\frac{\widehat{Q}_l - Q_l}{Q_l}(1+\eta') - \frac{\widehat{P}_l - P_l}{Q_l}\frac{\widehat{Q}_l - Q_l}{Q_l}(1+\eta') - 1 \\
&= \frac{P_l}{Q_l} + \eta\sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} + \rho_L\eta\sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} + \eta\sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} - 1 \\
&= \frac{P_l}{Q_l} + \rho_L\eta\sqrt{\frac{(P_l \vee Q_l)^2}{nQ_l^2(P_l \vee Q_l)}} - 1 \\
&= \frac{P_l - Q_l}{Q_l} + \eta\rho_L^2\sqrt{\frac{1}{n(P_l \vee Q_l)}} \\
&= \frac{P_l - Q_l}{Q_l} + \eta'\rho_L\sqrt{\frac{1}{n(P_l \vee Q_l)}}.
\end{aligned}
$$

Using the inequality $|\sqrt{1+x} - 1| \le x$ for $x \ge 0$, we conclude that

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

$$\leq \left| \frac{P_l - Q_l}{Q_l} + \eta' \rho_L \frac{1}{\sqrt{n(P_l \vee Q_l)}} \right|$$
$$\overset{(a)}{\leq} 2\rho_L \frac{1}{\sqrt{n(P_l \vee Q_l)}},$$

where (a) follows because we assumed $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$, which means

$$\left| \frac{P_l - Q_l}{Q_l} \right| \leq \sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} = \sqrt{\frac{(P_l \vee Q_l)^2}{nQ_l^2(P_l \vee Q_l)}} \leq \rho_L \frac{1}{\sqrt{n(P_l \vee Q_l)}}.$$

□

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

Lemma B.5. *Suppose that*

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

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. We have the sequence of inequalities

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

where in the second to last equality, we used the fact that $\frac{\widehat{Q}_l - Q_l}{Q_l} = \eta' \to 0$. A small enough value of η leads to the proof of the first inequality.

Also, for small enough η, we have that $\widehat{P}_l \geq \frac{1}{2} P_l$ and $\widehat{Q}_l \geq \frac{1}{2} Q_l$, giving us the second inequality. □

Lemma B.6. *Define* $L_1 = \{l : n \frac{\Delta_l^2}{P_l \vee Q_l} \geq 1\}$. *Then*

$$C_1 \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} \leq I_L \leq C_2 \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l},, \tag{B.18}$$

for some constants C_1 *and* C_2.

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

$$I_L = -2\log\sum_l \sqrt{P_l Q_l} = -2\log\left(1 - \frac{\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2}{2}\right).$$

Using the fact that $I_L \to 0$ as $n \to \infty$, and thus $\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2 \to 0$, we see that the following bounds hold for all large enough n:

$$\frac{1}{2}\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2 \le I_L \le 2\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2.$$

Expressing $\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2$ as $\sum_l \frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2}$, we conclude that there exist constants $\widetilde{C}_1$ and $\widetilde{C}_2$ such that

$$\widetilde{C}_1\sum_l \frac{\Delta_l^2}{P_l \vee Q_l} \le I_L \le \widetilde{C}_2\sum_l \frac{\Delta_l^2}{P_l \vee Q_l}.$$

Observe that

$$\sum_l \frac{\Delta_l^2}{P_l \vee Q_l} = \sum_{l\in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} + \sum_{l\in L_1^c} \frac{\Delta_l^2}{P_l \vee Q_l}. \tag{B.19}$$

The contribution of the colors in L_1^c is bounded as

$$\sum_{l\in L_1^c} \frac{\Delta_l^2}{P_l \vee Q_l} \le \frac{L}{n}, \tag{B.20}$$

and using our assumptions $\frac{L}{n}$ is $o(I_L)$. This implies that the contribution from the colors in L_1 must be $\Theta(I_L)$; i.e.. we can find constants C_1 and C_2 such that

$$C_1\sum_{l\in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} \le I_L \le C_2\sum_{l\in L_1} \frac{\Delta_l^2}{P_l \vee Q_l}.$$

□

We state Lemma 4 from [10] which analyzes the consensus step of the algorithm.

Lemma B.7. *Let σ, σ' be two clusters such that, for some constant $C \ge 1$, 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 B.8. *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*

$$d(\pi(\sigma), \sigma') < T/2 \qquad \text{and} \qquad d(\xi(\sigma), \sigma') < T/2$$

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,

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

□

C Proof of Proposition 5.2

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

Proposition 5.2. *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

$$\widetilde{P}_l = (1 - P_0) \int_{a_l}^{b_l} p(x)dx := (1 - P_0)P_l \quad \text{and}$$
$$\widetilde{Q}_l = (1 - Q_0) \int_{a_l}^{b_l} q(x)dx := (1 - Q_0)Q_l.$$

Suppose $L \leq \frac{2}{H}$. Define

$$I = -2 \log\left(\sqrt{P_0 Q_0} + \int \sqrt{(1 - P_0)(1 - Q_0)p(x)q(x)}dx\right) \quad \text{and}$$
$$I_L = -2 \log\left(\sqrt{P_0 Q_0} + \sum_{l=1}^{L} \sqrt{\widetilde{P}_l \widetilde{Q}_l}\right).$$

Then, we have that

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

and that $\frac{1}{2\rho c_0} \leq \frac{\widetilde{P}_l}{\widetilde{Q}_l} \leq 2\rho c_0$ for all l.

Proof. We first show that the likelihood ratio $\frac{\widetilde{P}_l}{\widetilde{Q}_l} = \frac{1-P_0}{1-Q_0}\frac{P_l}{Q_l}$ satisfies the bounds claimed. Let us consider an l such that $\text{bin}_l \cap R^c = \emptyset$. Then, for all $x \in \text{bin}_l$, we have that $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$ by assumption C2. It follows immediately that $\frac{P_l}{Q_l} = \frac{\int_{\text{bin}l} p(x)dx}{\int_{\text{bin}_l} q(x)dx} \leq \rho$. The lower bound follows in the same manner. Using the fact that $\frac{1-P_0}{1-Q_0}$ is bounded between $1/c_0$ and c_0, we conclude that $\frac{P_l}{Q_l}$ is bounded between $\frac{1}{\rho c_0}$ and ρc_0.

Now suppose $\text{bin}_l \cap R^c \neq \emptyset$. By the fact that R is an interval and that $\mu\{R^c\} = o(H)$, and since $L \leq 2/H$ we conclude that $\mu\{R^c\} < 1/2L$ for all large enough n. This means only bins $[0, \frac{1}{L}]$ and $[1-\frac{1}{L}, 1]$ can potentially satisfy $\text{bin}_l \cap R^c \neq 0$. Notice that assumption C5 indicates that both $p(x)$ and $q(x)$ are increasing in the interval $[0, 1/L]$ and decreasing in the interval $[1-1/L, 1]$. Define $P_l' = \int_{\text{bin}_l \cap R} p(x)dx$ and $Q_l' = \int_{\text{bin}_l \cap R} q(x)dx$. Define $P_l'' = \int_{\text{bin}_l \cap R^c} p(x)dx$ and $Q_l'' = \int_{\text{bin}_l \cap R^c} q(x)dx$. Thus, we have $P_l = P_l' + P_l''$ and $Q_l = Q_l' + Q_l''$ for all l. Note that $\frac{1}{\rho} \leq \frac{P_l'}{Q_l'} \leq \rho$ by same argument as before. Furthermore, using the monotonic properties of $p(x)$ and $q(x)$ in the relevant intervals, we have

$$P_l' \geq \min_{x \in \text{bin}_l \cap R} p(x)\frac{1}{2L} \geq \max_{x \in \text{bin}_l \cap R^c} p(x)\frac{1}{2L} \geq P_l'',$$

where the first inequality follows because $\mu(R^c) \leq \frac{1}{2L}$ and the second inequality follows from assumption C5. Similarly, we can derive that $Q_l' \geq Q_l''$. Thus,

$$\frac{P_l}{Q_l} \leq \frac{2P_l'}{Q_l'} \leq 2\rho, \quad \text{and}$$
$$\frac{P_l}{Q_l} \geq \frac{P_l'}{2Q_l'} \geq \frac{1}{2\rho}.$$

Using the bound on $\frac{1-P_0}{1-Q_0}$ completes the proof.

We now proceed with the bounding $|I - I_L|$. Using the simple relation between Renyi divergence by Hellinger distance detailed in Lemma I.1, 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

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

The key step in completing our proof is Proposition D.1, proved in Appendix D.1.

Proposition D.1. *Under assumptions C1-C5, we have that*

$$\left|\int(\sqrt{p(x)}-\sqrt{q(x)})^2 dx - \sum_{l=1}^{L}(\sqrt{P_l}-\sqrt{Q_l})^2\right| = o\left(\int(\sqrt{p(x)}-\sqrt{q(x)})^2 dx\right).$$

The claimed result follows from Proposition D.1 by noticing that

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

The proof Proposition D.1 contains a number of subparts, which we briefly describe below. Since $p(x)$ and $q(x)$ are easier to handle on the interval R, we initially only concern ourselves with comparing

$$H_R := \int_R (\sqrt{p(x)}-\sqrt{q(x)})^2dx \quad \text{and} \quad H_L^R := \sum_{l=1}^{L}(\sqrt{P_l'}-\sqrt{Q_l'})^2.$$

We first notice that $\{\text{bin}_l\} \cap R$ constitute an approximately uniform binning of R; i.e., we can find constants c_{bin} and C_{bin} such that $\frac{c_{\text{bin}}}{L} \le |\text{bin}_l \cap R| \le \frac{C_{\text{bin}}}{L}$. This is reasoned as follows. Since R is an interval, we know that $\text{bin}_l \cap R$ is an interval as well. Secondly, we have the inequality $|\text{bin}_l \cap R^c| \le \mu\{R^c\} \le \frac{1}{2L}$. So we have the inequality

$$\frac{1}{2L} \le |\text{bin}_l \cap R| \le \frac{1}{L}.$$

Then, in a series of lemmas, we show that such an approximately uniform binning of R leads to several useful bounds on H^R and H_L^R. In particular, we show in Lemma D.1 that as long as L grows, we have $d_L \to 1/4$, where

$$d_L := \sum_l Q_l'\left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right) \quad \text{and} \quad \int_R q(x)\left(\frac{1}{2}\frac{\gamma(x)}{q(x)}\right)^2 dx \overset{(a)}{=} \frac{1}{4}.$$

Here $\gamma_l' = Q_l' - P_l'$ and equality (a) follows from the definition of α in Assumption C.3. In Lemma D.2, we show that

$$H^R = \frac{\alpha^2}{4}(1+\eta),$$

where $\eta = \Theta(\alpha)$. Similarly, in Lemma D.3 we show that

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

where $\eta_L = \Theta(\alpha)$. We combine the results of Lemmas D.1, D.2 and D.3 in Lemma D.4 and show that

$$|H^R - H_L^R| = o(H^R).$$

Finally, we use Lemma D.4 to complete the proof of Proposition D.1. □

D Appendix for Proposition 5.2

D.1 Proof of Proposition D.1

Proposition D.1. *Suppose assumptions C1-C5 hold. Then we have that*

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

Proof. Let $H = \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$. Let a_L be an $o(1)$ sequence such that $\mu(R^c) \le a_L H$. We divide the set of bins into three sets L_1, L_2, L_3:

$$\begin{aligned}
L_1 &= \{l \,:\, \mathrm{bin}_l \cap R^c = \emptyset\} \\
L_2 &= \{l \,:\, \mathrm{bin}_l \cap R^c \neq \emptyset,\, P_l \vee Q_l \ge 2Ca_L H\} \\
L_3 &= \{l \,:\, \mathrm{bin}_l \cap R^c \neq \emptyset,\, P_l \vee Q_l \le 2Ca_L H\}.
\end{aligned}$$

For each bin l, define $P_l' = \int_{\mathrm{bin}_l \cap R} p(x)dx$ and $P_l'' = \int_{\mathrm{bin}_l \cap R^c} p(x)dx$. Likewise for Q_l' and Q_l''. We now proceed in two steps:

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

$$\left|(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2\right| \le a_L H.$$

Since $\mu(R^c) \le a_L H$, we have that $P_l'' = \int_{\mathrm{bin}_l \cap R^c} p(x)dx \le Ca_L H$ and likewise for Q_l''.

$$\begin{aligned}
(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2 &= P_l + Q_l - P_l' - Q_l' - 2\sqrt{P_l Q_l} + 2\sqrt{P_l' Q_l'} \\
&\overset{(a)}{\le} P_l'' + Q_l'' - 2\sqrt{P_l'' Q_l''} \\
&\le P_l'' + Q_l'' \\
&\le 2Ca_L H,
\end{aligned}$$

Here, (a) follows from the following reasoning. First, by AM-GM inequality, we have that $2\sqrt{P_l' Q_l' P_l'' Q_l''} \le P_l' Q_l'' + P_l'' Q_l'$. Thus:

$$P_l' Q_l' + P_l'' Q_l'' + 2\sqrt{P_l' Q_l' P_l'' Q_l''} \le (P_l' + P_l'')(Q_l' + Q_l'').$$

Taking square roots and replacing $P_l' + P_l'' = P_l$ and likewise for Q_l, we conclude

$$\sqrt{P_l' Q_l'} + \sqrt{P_l'' Q_l''} \le \sqrt{P_l Q_l},$$

which yields inequality (a).

On the other hand, we have that

$$\begin{aligned}
\sqrt{P_l Q_l} - \sqrt{P_l' Q_l'} &= \frac{P_l Q_l - P_l' Q_l'}{\sqrt{P_l Q_l} + \sqrt{P_l' Q_l'}} \\
&= \frac{P_l' Q_l'' + P_l'' Q_l' + P_l'' Q_l''}{\sqrt{P_l Q_l} + \sqrt{P_l' Q_l'}} \\
&\le \frac{P_l' Q_l'' + P_l'' Q_l' + P_l'' Q_l''}{2\sqrt{P_l' Q_l'}} \\
&\le Q_l'' \frac{P_l'}{2\sqrt{P_l' Q_l'}} + P_l'' \frac{Q_l'}{2\sqrt{P_l' Q_l'}} + Q_l'' \frac{P_l''}{2\sqrt{P_l' Q_l'}}.
\end{aligned}$$

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

$$\left|\frac{P_l'}{Q_l'}\right| = \left|\int_{\mathrm{bin}_l \cap R} \frac{p(x)}{Q_l'} dx\right| \le \int_{\mathrm{bin}_l \cap R} \left|\frac{p(x)}{q(x)}\right| \frac{q(x)}{Q_l'} dx \le \rho.$$

Thus, $\sqrt{\frac{P_l'}{Q_l'}} \vee \sqrt{\frac{Q_l'}{P_l'}} \le \sqrt{\rho}$. This bounds the terms $Q_l'' \frac{P_l'}{2\sqrt{P_l' Q_l'}} + P_l'' \frac{Q_l'}{2\sqrt{P_l' Q_l'}} \le \sqrt{\rho}(Q_l'' + P_l'')$.

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 $P_l \geq 2Ca_L H$ or that $Q_l \geq 2Ca_L H$. Let us suppose the former; the latter case can be handled in an identical manner. Since $P_l'' \leq Ca_L H$ and $P_l \geq 2Ca_L H$, 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

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

Thus,

$$\begin{aligned}(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2 &= P_l + Q_l - P_l' - Q_l' - 2\sqrt{P_l Q_l} + 2\sqrt{P_l' Q_l'} \\ &\geq P_l'' + Q_l'' - 4\sqrt{\rho}(Q_l'' + P_l'') \\ &\geq -(4\sqrt{\rho} - 1)(P_l'' + Q_l'') \\ &\geq -(4\sqrt{\rho} - 1) \cdot 2Ca_L H\end{aligned}$$

Combining these two bounds, we have

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

for an appropriate constant $C_{C,\rho}$. This completes step 1.

Step 2: In step 2, we verify that $\{\text{bin}_l\}_{l \in L_1} \cup \{\text{bin}_l \cap R\}_{l \in L_2} \cup \{\text{bin}_l \cap R\}_{l \in L_3}$ constitute a valid approximately uniform binning of R. First, because R is an interval, it is easy to see that $\text{bin}_l \cap R$ is an interval as well. Secondly, $|\text{bin}_l \cap R^c| \leq \mu\{R^c\} \leq a_L H$. 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 $\frac{C_{\text{bin}}}{L} \leq |\text{bin}_l \cap R| \leq \frac{1}{L}$.

Step 3: We now turn to main step of the proof. We can bound the difference between H and H_L as

$$\begin{aligned}&\left|\sum_{l=1}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| \\ &= \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_3}(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| \\ &\overset{(a)}{\leq} \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l} - \sqrt{Q_l})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + 8Ca_L H \\ &\overset{(b)}{\leq} \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l'} - \sqrt{Q_l'})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + C_{C,\rho} a_L H \\ &\overset{(c)}{\leq} \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l'} - \sqrt{Q_l'})^2 + \sum_{l \in L_3}(\sqrt{P_l'} - \sqrt{Q_l'})^2 - \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + C_{C,\rho} a_L H \\ &\overset{(d)}{\leq} \left|\sum_{l \in L_1}(\sqrt{P_l} - \sqrt{Q_l})^2 + \sum_{l \in L_2}(\sqrt{P_l'} - \sqrt{Q_l'})^2 + \sum_{l \in L_3}(\sqrt{P_l'} - \sqrt{Q_l'})^2 - \int_R(\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right| + C_{C,\rho} a_L H\end{aligned}$$

$\overset{(e)}{\leq} C_{C,\rho} a_L H.$

where (a) follows because $P_l \vee Q_l \leq 2Ca_L H$ for all $l \in L_3$ and because $|L_3| \leq 2$, (b) follows from step 1 and the fact that $|L_2| \leq 2$, (c) 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. Inequality (d) follows because $\int_{R^c} (\sqrt{p(x)} - \sqrt{q(x)})^2 \leq C\mu\{R^c\} = Ca_L H$, and (e) follows by Lemma D.4. Since $a_L \to 0$, the conclusion follows.

□

D.2 Lemmas for Proposition 5.2

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 $\frac{C_{\text{bin}}}{L} \leq b_l - a_l \leq \frac{C_{\text{bin}}}{L}$ for some constants C_{bin} and C_{bin}. We also use the notation bin_l to denote the bin $[a_l, b_l]$. Let $B_l = b_l - a_l$.

Lemma D.1. *Let $d_L = \sum_l Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2$. Suppose assumptions C1-C5 hold. Then we have that*

$$\lim_{L \to \infty} d_L = 1/4.$$

Proof. 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

$$\sup_{x \in \text{bin}_l} |h(x)| \leq L^\tau$$

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 $L \to \infty$. By assumption C4, for all large enough L, $\{x : |h(x)| \geq L^\tau\}$ is a union of at most K_h intervals. Thus, we have that

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

Here (a) follows by the finiteness of the integral $\int_R |h(x)|^t dx$ given by assumption C4, and (c) follows from because $t \leq 1$ by assumption C4, and $\tau t < 1$ by our selection, and so the first term dominates. We can now bound the number of bad bins:

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

where we redefine the constant $C_{M',K}$ suitably. For a bad bin l, we can bound $Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2$ as follows:

$$\begin{aligned}
Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2 &= Q'_l \left(\frac{1}{Q'_l} \int_{\text{bin}_l} \gamma(x) dx\right)^2 \\
&= Q'_l \left(\int_{\text{bin}_l} \frac{\gamma(x)}{q(x)} \frac{q(x)}{Q'_l} dx\right)^2 \\
&\overset{(a)}{\leq} Q'_l \int_{\text{bin}_l} \frac{q(x)}{Q'_l} \left(\frac{\gamma(x)}{q(x)}\right)^2 dx
\end{aligned}$$

$$
\begin{aligned}
&\leq \int_{\text{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&\overset{(b)}{\leq} \left(\int_{\text{bin}_l} q(x) \left|\frac{\gamma(x)}{q(x)}\right|^r dx\right)^{2/r} \left(\int_{\text{bin}_l} q(x)dx\right)^{(r-2)/r} \\
&\overset{(c)}{\leq} M^{2/r}(CC_{\text{bin}})^{(r-2)/r} L^{-(r-2)/r} = C_{M,C} L^{-(r-2)/r}.
\end{aligned}
$$

Here (a) follows from Jensen's inequality, (b) follows from Holder's inequality, and (c) follows from the finiteness of the integral as per assumption C3. Now, we have

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

For each good bin l, define $x_l = \arg\max_{x \in \text{bin}_l} |q(x)|$. The maximum is attainable since q is continuous and $q(x_l) < \infty$ since q is bounded. Now, for a good bin, we have that

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

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 $B_l \left|\frac{\xi_l}{q(x_l)}\right|$:

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

The second inequality follows because $q(c_x) \leq q(x_l)$. The third inequality follows because l is a good bin and thus $\left|\frac{q'(c_x)}{q(c_x)}\right| \leq L^\tau$. The last inequality follows because $B_l \leq C_{\text{bin}} 1/L$. We perform similar analysis on γ:

$$
\gamma'_l = \int_{\text{bin}_l} \gamma(x)dx
$$

$$= \int_{a_l}^{b_l} \gamma(x_l) + \gamma'(c_x)(x - x_l)dx$$
$$= B_l\gamma(x_l) + B_l^2\xi_l',$$

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 $B_l \left|\frac{\xi_l'}{q(x_l)}\right| \leq \frac{1}{2}C_{\text{bin}}L^{\tau-1}$. For any bin l, we also have that

$$Q_l' = \int_{\text{bin}_l} q(x)dx \leq CB_l,$$

where C is the bound on $p(x) \vee q(x)$. Now we look at a single $Q_l'\left(\frac{\gamma_l'}{Q_l'}\right)^2$ term for a single good bin l:

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

To arrive at the last equality, we assume that $L \geq C_{\text{bin}}^{1/(1-\tau)}$. Under this assumption, $\left|B_l\frac{\xi_l'}{q(x_l)}\right| \leq \frac{1}{2}$ and thus it is valid to take the Taylor approximation. Here, η_l is some scalar that satisfies $|\eta_l| \leq 16$. Expanding the right hand side,

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

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

$$\left|B_l\frac{\xi_l}{q(x_l)}\right| + \left|\eta_l(B_l\frac{\xi_l}{q(x_l)})^2\right| \leq C_{\text{bin}}L^{\tau-1}, \text{ and}$$
$$\left|1 - B_l\frac{\xi_l}{q(x_l)} + \eta_l(B_l\frac{\xi_l}{q(x_l)})^2\right| \leq 2.$$

Now, we can bound

$$\begin{aligned}
&\left|Q_l'\left(\frac{\gamma_l'}{Q_l'}\right)^2 - B_lq(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \\
&\leq B_lq(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 C_{\text{bin}}L^{\tau-1} + 2B_lq(x_l)\frac{\gamma(x_l)}{q(x_l)}C_{\text{bin}}L^{\tau-1} + B_lq(x_l)C_{\text{bin}}^2L^{2(\tau-1)}.
\end{aligned}$$

The third term is bounded by $C_1L^{2\tau-3}$ for a suitable constant C_1. 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 $2B_l C C_{\text{bin}} L^{\tau-1} \leq C_2 L^{\tau-2}$ for some constant C_2

In any case, we have that

$$\left|Q_l'\left(\frac{\gamma_l'}{Q_l'}\right)^2 - Bq(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \leq C_3 B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 L^{\tau-1} + C_4 L^{\tau-2}.$$

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

$$\begin{aligned}
|d_L - d_R| &= \left|\sum_l Q_l'\left(\frac{\gamma_l'}{Q_l'}\right)^2 - \sum_{l\ good} B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \\
&\leq \sum_{l\ good}\left|Q_l'\left(\frac{\gamma_l'}{Q_l'}\right)^2 - B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| + C_{M,M',K,C} L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\tau-1} + L \cdot C_4 L^{\tau-2} + C_5 L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\tau-1} + C_4 L^{\tau-1} + C_5 L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\frac{2-rt}{r(1+t)}} + C_6 L^{\frac{2-rt}{r(1+t)}} \quad \text{setting } \tau = \frac{2+r}{r(1+t)}.
\end{aligned}$$

The τ is chosen to balance $L^{\tau-1}$ and $L^{2/r-\tau t}$. Notice that $0 < \tau < 1$ by assumption C4, which says $rt > 2$. Furthermore, since $2 > rt$, we have that

$$|d_L - d_R| = o(d_R) + o(1).$$

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

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

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

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

The points c_l, c_l' are in bin_l and they are dependent on x; we leave that dependency implicit to make the notations simpler. For bin_l, we have

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

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 $|x-x_l| \le B_l$ and that

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

and likewise, $\left|\frac{q'(c_l)}{q(x_l)}\right| \le L^\tau$. Thus, we have that $|T_1|, |T_2| \le C_{\text{bin}}L^{\tau-1}$. Now, suppose $C_{\text{bin}}L^{\tau-1} \le \frac{1}{2}$, which is satisfied if $L \ge (2C_{\text{bin}})^{\frac{1}{1-\tau}}$. We obtain

$$\begin{aligned}\int_{\text{bin}_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx &= \int_{\text{bin}_l} q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}+T_2\right)^2\left(\frac{1}{1+T_1}\right)dx \\ &= \int_{\text{bin}_l}\left(q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 + q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)T_2 + q(x_l)T_2^2\right)(1-T_1+\eta T_1^2)dx,\end{aligned}$$

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

$$\begin{aligned}&\left|\int_{\text{bin}_l} q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 dx - B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \\ &\le B_l q(x_l)\left(\frac{\gamma(x_l)}{q(x_l)}\right)^2 C_{\text{bin}}L^{\tau-1} + B_l q(x_l)\frac{\gamma(x_l)}{q(x_l)}C_{\text{bin}}L^{\tau-1} + B_l q(x_l)C_{\text{bin}}^2 L^{2(\tau-1)}.\end{aligned}$$

Analyzing exactly as in the case of $|d_L - d_R|$, we conclude that

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

Since $d = 1/4$, we have that $d_R \to 1/4$, which in turn implies $d_L \to 1/4$. This completes the proof. □

Lemma D.2. *Let $H^R = \int_R (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$, $\delta(x) = q(x) - p(x)$, $\alpha^2 = \int_R q(x)\left(\frac{\delta(x)}{q(x)}\right)^2 dx$, and $\gamma(x) = \frac{\delta(x)}{\alpha}$. Suppose that assumptions C1-C5 hold. Then, we have that,*

$$H^R = \frac{\alpha^2}{4}(1+\eta),$$

where $|\eta| \le C(\alpha + \alpha^2)$ for some constant C. In particular, we have that if $H^R \to 0$, then $\alpha \to 0$ and $\eta \to 0$.

Proof. We write H^R as

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

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 $x \in [0, 1]$ and rewrite

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

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

$$\int_R \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx = (1 + \eta)\frac{\alpha^2}{4},$$

where

$$\eta = \frac{\int_R q(x) \left(\frac{1}{2}\frac{\delta(x)}{q(x)}\right)^2 (\xi_2(x)^2 + 2\xi_2(x)) dx}{\alpha^2/4} = \int_R q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 (\xi_2(x)^2 + 2\xi_2(x)) dx.$$

By Lemma I.3, we have that $\xi_2(x) \leq 2\left|\frac{\delta(x)}{q(x)}\right|$. This gives

$$\begin{aligned} |\eta| &\leq \int_R q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 \left(4\left|\frac{\delta(x)}{q(x)}\right|^2 + 4\left|\frac{\delta(x)}{q(x)}\right|\right) dx \\ &= 4\alpha^2 \int_R q(x) \left|\frac{\gamma(x)}{q(x)}\right|^4 dx + 4\alpha \int_R q(x) \left|\frac{\gamma(x)}{q(x)}\right|^3 dx \\ &\leq C(\alpha^2 + \alpha), \end{aligned}$$

using the finiteness of integrals in assumption C3. □

Lemma D.3. *Let* $H_L^R = \sum_{l=1}^L \left(\sqrt{P_l'} - \sqrt{Q_l'}\right)^2$. *Let* $\delta(x) = q(x) - p(x)$, $\alpha^2 = \int_R q(x) \left(\frac{\delta(x)}{q(x)}\right)^2 dx$, *and* $\gamma(x) = \frac{\delta(x)}{\alpha} dx$. *Suppose that assumptions C1-C5 hold. Then, we have that,*

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

where $d_L = \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^2 dx$, $\gamma_l' = \frac{Q_l' - P_l'}{\alpha}$ *and* $\sup_L |\eta_L| \leq C(\alpha + \alpha^2)$ *for some constant* C.

Proof. Let us define $\delta_l = Q'_l - P'_l$. We write H_L^R as

$$\begin{aligned} H_L^R &= \sum_{l=1}^{L} (\sqrt{P'_l} - \sqrt{Q'_l})^2 \\ &= \sum_{l=1}^{L} Q'_l \left(1 - \sqrt{\frac{P'_l}{Q'_l}}\right)^2 \\ &= \sum_{l=1}^{L} Q'_l \left(1 - \sqrt{1 - \frac{\delta_l}{Q'_l}}\right)^2 \\ &= \sum_{l=1}^{L} Q'_l \left(1 - \left(1 - \frac{1}{2}\frac{\delta_l}{Q'_l} - \xi_l\right)\right)^2, \end{aligned}$$

where by convention, we define $\frac{\delta_l}{Q'_l} = 0$ when $Q'_l, P'_l = 0$ and where $\xi_l = 1 - \frac{1}{2}\frac{\delta_l}{Q'_l} - \sqrt{1 - \frac{\delta_l}{Q'_l}}$. 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. This may also be written as

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

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 I.3, we have the inequality $|\xi_{2l}| \leq 2\left|\frac{\delta_l}{Q'_l}\right|$. Therefore,

$$\begin{aligned} |\eta_L| &= \left| \frac{\sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\delta_l}{Q'_l}\right)^2 (2\xi_{2l} - \xi_{2l}^2)}{\sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\delta_l}{Q'_l}\right)^2} \right| \\ &\leq \frac{\sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2 (2|\xi_{2l}| + \xi_{2l}^2)}{\sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2} \\ &\leq 4\frac{\alpha \sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^3 + \alpha^2 \sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^4}{\sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2}. \end{aligned}$$

The denominator tends to $1/4$ by Lemma D.1 and can be bounded by $1/(2C')$ for large enough L . To bound the numerator, we note that for a single l:

$$\int_{a_l}^{b_l} \frac{q(x)}{Q'_l} \left|\frac{\gamma(x)}{q(x)}\right|^3 dx \geq \left| \int_{\text{bin}_l} \frac{q(x)}{Q'_l} \frac{\gamma(x)}{q(x)} dx \right|^3 = \left|\frac{\gamma'_l}{Q'_l}\right|^3.$$

Therefore,

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

Thus,

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

□

Lemma D.4. *Suppose assumptions A1-4 hold. Let $n \to \infty$, then for any sequence $L_n \to \infty, \alpha_n \to 0$ we have $H_L^R = H^R(1+o(1))$; i.e.,*

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

Proof. By Lemma D.2 and Lemma D.3, we have that, for all α,

$$|H_L^R - H^R| = \left|d_L\alpha^2(1+\eta_L) - \frac{\alpha^2}{4}(1+\eta)\right|,$$

which implies

$$\left|\frac{H_L^R}{H^R}-1\right| = \left|4d_L\frac{(1+\eta_L)}{1+\eta}-1\right|,$$

where $|\eta|, |\eta_L| \le C_1(\alpha+\alpha^2)$ 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, by Lemma D.1,

$$\lim_{L_n\to\infty}|4d_L - 1| = 0,$$

uniformly for all α. From these two observations, it is clear that $\left|\frac{H_L^R}{H^R}-1\right| \to 0$, which completes the proof. □

E Proof of Proposition 5.3

Proposition 5.3. *Suppose $p(x), q(x)$ are supported on $[0,1]$.*

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

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

C3' Let $h(x) \ge \sup_n \max\left\{\left|\frac{p'(x)}{p(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|\right\}$. Suppose $\int |h(x)|^t dx \le M'$ for some constant M' and $1 \ge t \ge 2/r$. Suppose also that the level set $\{x : |h(x)| \ge \kappa\}$ is a union of at most K_h 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* $\widetilde{P}_l = (1-P_0)\int_{a_l}^{b_l} p(x)dx$ *and* $\widetilde{Q}_l = (1-Q_0)\int_{a_l}^{b_l} q(x)dx$. *Define* $I = -2\log\left(\sqrt{P_0Q_0} + \int\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right)$ *and* $I_L = -2\log\left(\sqrt{P_0Q_0} + \sum_{l=1}^{L}\sqrt{\widetilde{P}_l\widetilde{Q}_l}\right)$. *Then, we have that*

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

and that $\frac{1}{4\rho_L c_0} \leq \frac{\widetilde{P}_l}{\widetilde{Q}_l} \leq 4\rho_L c_0$ *for all* l.

Proof. The proof of bounds on $\frac{\widetilde{P}_l}{\widetilde{Q}_l}$ is exactly as in Proposition 5.3, and we omit it here. The rest of the proof also follows that of Proposition 5.2, except the final step where we need to show that

$$\left|\int(\sqrt{p(x)} - \sqrt{q(x)})^2dx - \sum_l(\sqrt{P_l} - \sqrt{Q_l})^2\right| = o(\int(\sqrt{p(x)} - \sqrt{q(x)})^2dx) = o(1).$$

We establish this fact in Proposition E.1 in Appendix E.1. □

E.1 Appendix for Proposition 5.3

Proposition E.1. *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*

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

Proof. The proof will be similar to that of Proposition D.1. First, we observe that $\int(\sqrt{p(x)} - \sqrt{q(x)})^2dx = \int p(x)dx + \int q(x)dx - 2\int\sqrt{p(x)q(x)}dx = 2 - 2\int\sqrt{p(x)q(x)}$. And, that $\sum_l(\sqrt{P_l} - \sqrt{Q_l})^2 = \sum_l P_l + \sum_l Q_l - 2\sum_l\sqrt{P_lQ_l}$. Thus, we need only show that

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

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

$$\sup_{x \in \text{bin}_l}|h(x)| \leq L^\tau,$$

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

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

The last inequality follows because $t \leq 1$ and thus $\tau t < 1$ and so the first term dominates. The second inequality follows from the assumption C3'. We can now bound the number of bad bins:

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

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

Now we consider a good bin l. Let x_l be $\arg\max_{x \in \text{bin}_l} p(x)$. 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 $B_l \left| \frac{\xi_l}{p(x_l)} \right|$:

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

Likewise, define $x_l' = \arg\max_{x \in \text{bin}_l} q(x)$. 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

$$B_l \left| \frac{\xi_l'}{q(x_l')} \right| \leq \frac{1}{2} C_{\text{bin}} L^{\tau - 1}.$$

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 nuisance term

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

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

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

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

$$\left|(x-x_l)\frac{p'(c_x)}{p(x_l)}\right| \le B_l\left|\frac{p'(c_x)}{p(c_x)}\right| \le C_{\text{bin}}L^{\tau-1},$$

and

$$\left|(x-x_l)\frac{q'(c'_x)}{q(x_l)}\right| \le B_l\left|\frac{q'(c'_x)}{q(c'_x)}\right| \le C_{\text{bin}}L^{\tau-1}.$$

Therefore, we can bound the nuisance term:

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

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

$$\left|\int_{a_l}^{b_l}\sqrt{p(x)q(x)}dx - B_l\sqrt{p(x_l)q(x'_l)}\right| \le B_lCC_{\text{bin}}L^{\tau-1} \tag{E.2}$$

By combining inequalities (E.1) and (E.2), we have that

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

We can then complete the proof:

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

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

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

□

F 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.

F.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:

$$p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))} \quad \text{and} \quad q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}.$$

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

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

Under A1-A5, Proposition F.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 5.2 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 5.2 that $I_L = I(1 + o(1))$, we know that $L = o(nI_L)$ as well and thus, we can use Proposition 5.1 (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 $I_L = I(1 + o(1))$.

F.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 F.2 then shows that assumptions A1'-A4' implies C1'-C4'. Therefore, Proposition 5.3 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 5.1 (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 $I_L = I(1 + o(1))$.

F.3 Transformation Analysis

Proposition F.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 $h(x) \geq \sup_n \max\left\{\left|\frac{\gamma'(x)}{q(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{\phi'(x)}{\phi(x)}\right|, \left|\frac{\gamma(x)}{q(x)}\right|\right\}$. 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 $p_\Phi(z), q_\Phi(z)$ be the Φ-transformed densities over $[0,1]$. Then, we have that the following conditions are satisfied for $p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$:

C1 We have $p_\Phi(z), q_\Phi(z) \leq C$ on $[0,1]$ and are absolutely continuous.

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

C3 Define $\alpha^2 = \int_R \frac{(p_\Phi(z)-q_\Phi(z))^2}{q_\Phi(z)} dz$ and $\gamma_\Phi(z) = \frac{q_\Phi(z)-p_\Phi(z)}{\alpha}$. Then $\int_R q_\Phi(z) \left|\frac{\gamma_\Phi(z)}{q_\Phi(z)}\right|^r dz \leq M$ for constants $M, r \geq 4$.

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

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

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. We prove that C5 holds first. Note that

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

Therefore, $p'_\Phi(z) \geq 0$ if and only if $p'(x) \geq p(x)\frac{\phi'(x)}{\phi(x)}$. Likewise for $q'_\Phi(z)$. Claim C5 follows. For C4, note that

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

By a change of variables, we have that

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

$$\overset{(a)}{\leq} \int_R \left|\frac{q'(x)}{q(x)}\frac{1}{\phi(x)}\right|^t \phi(x)dx + \int_R \left|\frac{\phi'(x)}{\phi(x)}\frac{1}{\phi(x)}\right|^t \phi(x)dx,$$

where (a) follows because $t \leq 1$. This is finite since

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

and each of the two terms in the product are finite by A4. Finally, we also have

$$\begin{aligned}
\int_R \left|\frac{\gamma_\Phi'(z)}{q_\Phi(z)}\right|^t dz &= \int_R \left|\frac{1}{\alpha}\frac{p'(x) - p(x)\frac{\phi'(x)}{\phi(x)} - q'(x) + q(x)\frac{\phi'(x)}{\phi(x)}}{q(x)\phi(x)}\right|^t \phi(x)dx \\
&= \int_R \left|\frac{1}{\alpha}\frac{p'(x)-q'(x)}{q(x)} - \frac{1}{\alpha}\frac{p(x)-q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right|^t \left|\frac{1}{\phi(x)}\right|^t \phi(x)dx \\
&\overset{(a)}{\leq} \left\{\int_R \left|\frac{1}{\alpha}\frac{p'(x)-q'(x)}{q(x)} - \frac{1}{\alpha}\frac{p(x)-q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right|^{2t/(1-t)} \phi(x)dx\right\}^{(1-t)/2} \left\{\int_R \phi(x)^{\frac{1-t}{1+t}}dx\right\}^{(1+t)/2},
\end{aligned}$$

where we use Holder's inequality in step (a). Using A4 again, we see that it is enough to show the finiteness of

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

and

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

The former is finite by A4, and the latter is finite by Cauchy-Schwartz and A4. □

Proposition F.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* $h(x) \geq \sup_n \max\left\{\left|\frac{p'(x)}{p(x)}\right|, \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{\phi'(x)}{\phi(x)}\right|, \left|\frac{\gamma(x)}{q(x)}\right|\right\}$. *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_h 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 $p_\Phi(z), q_\Phi(z)$ be the Φ-transformed densities over $[0,1]$. Then, we have that the following conditions are satisfied for $p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ *and* $q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$, *for a sequence L such that $L \to \infty$:*

C1' *We have $p_\Phi(z), q_\Phi(z) \leq C$ 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* $\mu\{R^c\} \leq \frac{1}{2L}$, *where $r > 2$ is a constant.*

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

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

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

G Proof of Proposition 4.1

First suppose $\|\theta_1 - \theta_0\| \to 0$. In Lemma G.2, we show that in this case $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$. Condition A1 follows directly from condition B1 and condition A5 follows directly from condition B5. In Propositions G.1, G.2, G.3 below, we establish conditions A2, A3, and A4, respectively.

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

G.1 Supporting propositions

Proposition G.1. *Suppose assumptions B1-B5 hold. 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

$$\begin{aligned}\log \frac{p(x)}{q(x)} &= f_{\theta_1}(x) - f_{\theta_0}(x) \\ &= (\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x).\end{aligned}$$

This implies

$$\begin{aligned}\left|\log \frac{p(x)}{q(x)}\right| &\leq \|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\| \\ &\leq \|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\|,\end{aligned}$$

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. By B3, this set R is an interval 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

$$\Phi(R^c) = \Phi\left\{x : \sup_\theta \|\nabla_\theta f_\theta(x)\| > \frac{\log \rho}{\|\theta_1 - \theta_0\|}\right\} \leq C\frac{\|\theta_1 - \theta_0\|^r}{(\log \rho)^r} \overset{(a)}{=} o(H),$$

where (a) follows $r \geq 4$, $\rho > 1$, and from Lemma G.2. □

Proposition G.2. *Suppose assumptions B1-B5 are satisfied. Let $R \subset \mathbb{R}$ be the interval in Proposition G.1. Then, we have that,*

$$\int_R \frac{1}{\alpha^r} \left| \frac{p(x)}{q(x)} - 1 \right|^r q(x)dx \leq \infty.$$

Proof. By Lemma G.1, we have that $\alpha \asymp \|\theta_1 - \theta_0\|$. Thus, we need only show that

$$\int_R \frac{1}{\|\theta_0 - \theta_1\|^r} \left| \frac{p(x)}{q(x)} - 1 \right|^r q(x)dx \leq \infty$$

$$\begin{aligned}
\frac{1}{\|\theta_1 - \theta_0\|} \left| \frac{p(x)}{q(x)} - 1 \right| &= \frac{1}{\|\theta_1 - \theta_0\|} |\exp(f_{\theta_1}(x) - f_{\theta_0}(x)) - 1| \\
&= \frac{1}{\|\theta_1 - \theta_0\|} \left| (\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\bar{\theta}}(x) \right| \exp(f_{\bar{\theta}}(x) - f_{\theta_0}(x)) \\
&\leq \|\nabla_\theta f_{\bar{\theta}}(x)\| \exp(f_{\bar{\theta}}(x) - f_{\theta_0}(x)) \\
&= \|\nabla_\theta f_{\bar{\theta}}(x)\| \exp\left((\bar{\theta} - \theta_1)^\mathsf{T} \nabla_\theta f_{\widetilde{\theta}}(x) \right) \\
&\leq \|\nabla_\theta f_{\bar{\theta}}(x)\| \exp\left(\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\widetilde{\theta}}(x)\| \right),
\end{aligned}$$

Where $\bar{\theta}, \widetilde{\theta}$ are some convex combinations of θ_0, θ_1. Therefore,

$$\begin{aligned}
\int_R \left(\frac{1}{\|\theta_1 - \theta_0\|} \left| \frac{p(x)}{q(x)} - 1 \right| \right)^r q(x)dx &\leq \int_R \|\nabla_\theta f_{\bar{\theta}}(x)\|^r \exp\left(r\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\widetilde{\theta}}(x)\| \right) q(x)dx \\
&\overset{(a)}{\leq} \int_R \|\nabla_\theta f_{\bar{\theta}}(x)\|^r e^{r \log \rho} q(x)dx \\
&= \int_R \|\nabla_\theta f_{\bar{\theta}}(x)\|^r \rho^r q(x)dx \\
&\leq \rho^r \int_{-\infty}^{\infty} \|\nabla_\theta f_{\bar{\theta}}(x)\|^r q(x)dx \\
&\overset{(b)}{<} \infty,
\end{aligned}$$

where (a) follows from the definition of R and (b) follows from assumptions B1 and B4. □

Proposition G.3. *Suppose assumptions B1-B4 are satisfied. Let $R \subset \mathbb{R}$ be the interval in Proposition G.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. Something's wrong here. Exponent needs to be $4t/(1-t)$, which leads to changes in assumptions B1-B5. Also, assumption A4 was modified. That also needs to be incorporated into the proof. Note that we only need to prove assumption A4 for functions $\left|\frac{\gamma'(x)}{q(x)}\right|$ and $\left|\frac{\gamma(x)}{q(x)}\right|$, since assumption B4 takes care of the rest. We first consider $\left|\frac{\gamma'(x)}{q(x)}\right|$. Using the fact that $\alpha \asymp \|\theta_1 - \theta_0\|$, we need only prove that

$$\int_R \left| \frac{1}{\|\theta_1 - \theta_0\|} \frac{p'(x) - q'(x)}{q(x)} \right|^{2t/(1-t)} \phi(x)dx < \infty.$$

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:

$$\begin{aligned}&\int_R \left|\frac{1}{\|\theta_1-\theta_0\|}\frac{p'(x)-q'(x)}{q(x)}\right|^{2t/(1-t)}\phi(x)dx \\ &= \int_R \left|\nabla_\theta f'_{\overline{\theta}}(x)\frac{p(x)}{q(x)} + \frac{1}{\|\theta_1-\theta_0\|}f'_{\theta_0}(x)\left(\frac{p(x)}{q(x)}-1\right)\right|^{2t/(1-t)}\phi(x)dx.\end{aligned}$$

To show that this integral is finite, we need only show, regardless of the value of $2t/(1-t)$, that the two components have finite integrals:

$$\int_R \left|\nabla_\theta f'_{\overline{\theta}}(x)\frac{p(x)}{q(x)}\right|^{2t/(1-t)}\phi(x)dx < \infty \quad \text{and} \quad \int_R \left|\frac{1}{\|\theta_1-\theta_0\|}f'_{\theta_0}(x)\left(\frac{p(x)}{q(x)}-1\right)\right|^{2t/(1-t)}\phi(x)dx < \infty.$$

We bound the first integral as follows:

$$\int_R \left|\nabla_\theta f'_{\overline{\theta}}(x)\frac{p(x)}{q(x)}\right|^{2t/(1-t)}\phi(x)dx \leq \int_R \left|\nabla_\theta f'_{\overline{\theta}}(x)\right|^{2t/(1-t)}\rho\phi(x)dx \overset{(a)}{<} \infty,$$

where (a) follows from assumption B4. For the second term, we simplify as follows:

$$\begin{aligned}&\int_R \left|\frac{1}{\|\theta_1-\theta_0\|}f'_{\theta_0}(x)\left(\frac{p(x)}{q(x)}-1\right)\right|^{2t/(1-t)}\phi(x)dx \\ &\leq \int_R |f'_{\theta_0}(x)|^{2t/(1-t)}\left|\frac{1}{\|\theta_1-\theta_0\|}\left(\frac{p(x)}{q(x)}-1\right)\right|^{2t/(1-t)}\phi(x)dx \\ &\leq \left\{\int_R |f'_{\theta_0}(x)|^{4t/(1-t)}\phi(x)dx\right\}^{1/2}\left\{\int_R \left|\frac{1}{\|\theta_1-\theta_0\|}\left(\frac{p(x)}{q(x)}-1\right)\right|^{4t/(1-t)}\phi(x)dx\right\}^{1/2}\end{aligned}$$

The first quantity is finite by assumption. A bound for the second quantity follows from proposition G.2. □

G.2 Supporting lemmas

Lemma G.1. *Suppose assumptions B1-B5 hold. Let R be the interval in Proposition G.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. We write α^2 as

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

$$= \int_R \left|\exp\Big(f_{\theta_1}(x) - f_{\theta_0}(x)\Big) - 1\right|^2 q(x)dx$$
$$= \int_R \Big((\theta_1-\theta_0)^\mathsf{T}\nabla_\theta f_{\bar\theta}(x)\exp\Big(f_{\bar\theta}(x) - f_{\theta_0}(x)\Big)\Big)^2 q(x)dx.$$

First we show an upper bound:

$$\alpha^2 \le \int_R \|\theta_1-\theta_0\|^2\|\nabla_\theta f_{\bar\theta}(x)\|^2 \exp\Big(f_{\bar\theta}(x) - f_{\theta_0}(x)\Big)\exp(f_{\bar\theta}(x))dx$$
$$\le \int_R \|\theta_1-\theta_0\|^2\|\nabla_\theta f_{\bar\theta}(x)\|^2 \exp\Big(\|\theta_1-\theta_0\|\|\nabla_\theta f_{\tilde\theta}(x)\|\Big)\exp(f_{\bar\theta}(x))dx.$$

Since we are in R, we have that $\|\theta_1-\theta_0\|\sup_\theta\|\nabla_\theta f_\theta(x)\| \le \log\rho$. We can thus continue the bound:

$$\alpha^2 \le \|\theta_1-\theta_0\|^2\int_R \|\nabla_\theta f_{\bar\theta}(x)\|^2 e^{\log\rho}\exp(f_{\bar\theta}(x))dx$$
$$\le \|\theta_1-\theta_0\|^2\rho\int_{-\infty}^{\infty}\|\nabla_\theta f_{\bar\theta}(x)\|^2\exp(f_{\bar\theta}(x))dx$$
$$\overset{(a)}{\lesssim} \|\theta_1-\theta_0\|^2.$$

where (a) follows from assumptions B1 and B4. We now establish a lower bound:

$$\alpha^2 \ge \int_R \big((\theta_1-\theta_0)^\mathsf{T}\nabla_\theta f_{\bar\theta}(x)\big)^2\exp\Big(-|f_{\bar\theta}(x) - f_{\theta_0}(x)|\Big)\exp(f_{\bar\theta}(x))dx$$
$$\ge \int_R \big((\theta_1-\theta_0)^\mathsf{T}\nabla_\theta f_{\bar\theta}(x)\big)^2\exp\Big(-\|\theta_1-\theta_0\|\|\nabla_\theta f_{\bar\theta}(x)\|\Big)\exp(f_{\bar\theta}(x))dx$$
$$\overset{(a)}{\ge} e^{-C_R}(\theta_1-\theta_0)^\mathsf{T}\left(\int_R(\nabla_\theta f_{\bar\theta}(x))(\nabla_\theta f_{\bar\theta}(x))^\mathsf{T}\exp(f_{\bar\theta}(x))dx\right)(\theta_1-\theta_0),$$

where (a) follows from assumption B3. Define

$$\widetilde{G}_{\bar\theta} = \int_R(\nabla_\theta f_{\bar\theta}(x))(\nabla_\theta f_{\bar\theta}(x))^\mathsf{T}\exp(f_{\bar\theta}(x))dx.$$

For increasing ρ or as $\|\theta_1-\theta_0\|\to 0$, we have that R increases unboundedly, therefore, $\lambda_{min}(\widetilde{G}_{\bar\theta}) \to \lambda_{min}(G_{\bar\theta}) > 0$. Hence, $\alpha^2 \gtrsim \|\theta_1-\theta_0\|^2$. □

Lemma G.2. *When $\|\theta_1-\theta_0\| = \Theta(1)$, we have*

$$\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx = c\|\theta_0-\theta_1\|_2^2,$$

where $c_{\min} \le c \le \frac{1}{4}c_{\max}d_\Theta$.

Proof. Expanding the left hand side, wehave

$$\int(\sqrt{p(x)}-\sqrt{q(x)})^2dx = \int q(x)\left(\sqrt{\frac{p(x)}{q(x)}}-1\right)^2dx$$
$$= \int q(x)\Big(\exp\Big(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2\Big)-1\Big)^2dx.$$

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 $h(\theta_1) - h(\theta_0)$. We bound this as follows:

$$\begin{aligned}
|h(\theta_1) - h(\theta_0)| &= |(\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta h(\overline{\theta})| \\
&= \left| \frac{1}{2} (\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x) \exp(f_{\overline{\theta}}(x)/2 - f_{\theta_0}(x)/2) \right| \\
&\leq \frac{1}{2} \|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\| \exp(f_{\overline{\theta}}(x)/2 - f_{\theta_0}(x)/2),
\end{aligned}$$

where $\overline{\theta} \in \Theta$ is some convex combination of θ_1, θ_0. Thus, we have that

$$\begin{aligned}
\int q(x) \Big(\exp\Big(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2 \Big) - 1 \Big)^2 dx &\leq \int q(x) \frac{1}{4} \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp(f_{\overline{\theta}}(x) - f_{\theta_0}(x)) dx \\
&= \frac{1}{4} \|\theta_1 - \theta_0\|^2 \int \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp(f_{\overline{\theta}}(x)) dx \\
&\leq \frac{1}{4} \|\theta_1 - \theta_0\|^2 \operatorname{tr}(G_{\overline{\theta}}) \\
&\leq \frac{1}{4} \|\theta_1 - \theta_0\|^2 c_{\max} d_\Theta,
\end{aligned}$$

where $\Theta \subset \mathbb{R}^{d_\Theta}$. To get an upper bound,

$$\begin{aligned}
\int q(x) \Big(\exp\Big(f_{\theta_1}(x)/2 - f_{\theta_0}(x)/2 \Big) - 1 \Big)^2 dx &= \int \Big((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x) \Big)^2 \exp(f_{\overline{\theta}}(x)) dx \\
&= (\theta_1 - \theta_0)^\mathsf{T} G_{\overline{\theta}} (\theta_1 - \theta_0) \\
&\geq c_{\min} \|\theta_1 - \theta_0\|^2.
\end{aligned}$$

□

H Proof of Theorem 4.3

We first state a lemma that gives an alternative characterization of the Renyi divergence.

Lemma H.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*

$$I = -2 \log \int \left(\frac{dP}{dQ} \right)^{1/2} dQ \qquad \text{and} \quad D = \inf_{Y \in \mathcal{P}} \max \left\{ \int \log \frac{dY}{dP} dY, \int \log \frac{dY}{dQ} dY \right\},$$

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:

$$\inf_{Y \in \mathcal{P}} \int \log \frac{dY}{dP} dY \quad \text{such that} \quad \int \log \frac{dP}{dQ} dY = 0.$$

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. The quantity 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 $-2\log Z = I$. □

Throughout this proof, we let C denote a $\Theta(1)$ sequence whose value may change from instance to instance. Without loss of generality, let us suppose that cluster 1 and 2 are the two smallest clusters, each of size $\frac{n}{\beta K}$. Let us condition on the event that node 1 was placed in cluster 1, i.e, $\sigma_0(1) = 1$. Let $C_i = \{x : \sigma_0(x) = i\}$ be the i-th community. 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 H.1; i.e., $Y_0 \propto (P_0 Q_0)^{1/2}$ 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 $\mathcal{E} = \mathcal{E}(\widehat{\sigma}(G))$ be the set of vertices which are incorrectly clustered by $\widehat{\sigma}$, and let v_1 denote node 1. We have the equality

$$P_\Psi(\mathcal{Q} \leq f(n)) = P_\Psi(\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}) + P_\Psi(\mathcal{Q} \leq f(n), v_1 \notin \mathcal{E})$$

We first bound the first term as follows:

$$\begin{aligned}
P_\Psi(\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}) &= \int_{\mathcal{Q}\leq f(n), v_1\in\mathcal{E}} dP_\Psi \\
&= \int_{\mathcal{Q}\leq f(n), v_1\in\mathcal{E}} \exp(\mathcal{Q}) dP_\Phi \\
&\leq \exp(f(n)) P_\Phi(\mathcal{Q} \leq f(n), v_1 \in \mathcal{E}) \\
&\leq \exp(f(n)) P_\Phi(v_1 \in \mathcal{E}) \\
&\leq \exp(f(n)) \mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0).
\end{aligned}$$

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 $\frac{1}{2}$ 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

$$P_\Psi\left(\mathcal{Q} \leq \mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}\right) \geq 4/5,$$

where $V_\Psi(Q)$ is the variance of Q under measure Ψ. Hence, we get that $\log \frac{1}{4\mathbb{E}_\Phi l(\widehat{\sigma},\sigma_0)} \leq \mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}$, stated equivalently as

$$\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma_0) \geq \frac{1}{4}\exp\Big(- (\mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})})\Big).$$

We now just need to compute $\mathbb{E}_\Psi \mathcal{Q}$ and $V_\Psi(\mathcal{Q})$. Calculating the expected value first, we obtain

$$\begin{aligned}
\mathbb{E}_\Psi \mathcal{Q} &= \mathbb{E}_\Psi \sum_{v\neq 1,\, v\in C_1} \log \frac{Y(A_{1v})}{P(A_{1v})} + \sum_{v\neq 1,\, v\in C_2} \log \frac{Y(A_{1v})}{Q(A_{1v})} \\
&= (\frac{n}{\beta K} - 1)\int \log \frac{dY}{dP} dY + (\frac{n}{\beta K})\int \log \frac{dY}{dQ} dY \\
&= (\frac{n}{\beta K} - 1/2)2D \\
&= (\frac{n}{\beta K} - 1/2)I \\
&= \frac{nI}{\beta K}(1 + o(1)).
\end{aligned}$$

The bound for the variance term is involved. We start with the decomposition:

$$\begin{aligned}
V_\Psi(\mathcal{Q}) &= \sum_{v\neq 1, v\in C_1} V_\Psi\left(\log \frac{Y(A_{1v})}{P(A_{1v})}\right) + \sum_{v\neq 1, v\in C_2} V_\Psi\left(\log \frac{Y(A_{1v})}{Q(A_{1v})}\right) \\
&\leq (\frac{n}{\beta K} - 1)\mathbb{E}_\Psi\left(\log \frac{Y(A_{1v})}{P(A_{1v})}\right)^2 + \frac{n}{\beta K}\mathbb{E}_\Psi\left(\log \frac{Y(A_{1v})}{Q(A_{1v})}\right)^2.
\end{aligned}$$

We will show that $\mathbb{E}_\Psi\left(\log \frac{Y(A_{1v})}{P(A_{1v})}\right)^2$ can be bounded by CI so that $\sqrt{5V_\Psi(\mathcal{Q})} \leq C\sqrt{\frac{nI}{\beta K}}$. We have that

$$\begin{aligned}
\mathbb{E}_\Psi\left(\log \frac{Y(A_{1v})}{P(A_{1v})}\right)^2 &= \int \left(\log \frac{dY}{dP}\right)^2 dY \\
&= Y_0 \log^2 \frac{Y_0}{P_0} + (1 - Y_0)\int y(x) \log^2 \frac{(1 - Y_0)y(x)}{(1 - P_0)p(x)} dx. \qquad \text{(H.1)}
\end{aligned}$$

We bound the first term $Y_0 \log^2 \frac{Y_0}{P_0}$. We may bound $\left|\log \frac{Y_0}{P_0}\right|$ as

$$\begin{aligned}
\left|\log \frac{Y_0}{P_0}\right| &= \left|\frac{1}{2}\log \frac{Q_0}{P_0} - \log Z\right| \\
&\leq \frac{1}{2}\left|\log\left(1 - \frac{P_0 - Q_0}{P_0}\right)\right| + I/2 \\
&\overset{(a)}{\leq} \frac{1}{2}\left|\frac{P_0 - Q_0}{P_0}\right| + \left|\frac{P_0 - Q_0}{P_0}\right|^2 C + I/2 \\
&\overset{(b)}{\leq} C\left|\frac{P_0 - Q_0}{P_0}\right| + CI
\end{aligned}$$

Inequality (a) follows from Lemma I.2 and from the fact that $\frac{Q_0}{P_0}$ is bounded. Inequality (b) follows from the fact that $\left|\frac{P_0 - Q_0}{P_0}\right| = \left|1 - \frac{Q_0}{P_0}\right| \leq 1 + \left|\frac{Q_0}{P_0}\right|$ is bounded by a constant.

Therefore, we have that:

$$\begin{aligned}
Y_0 \log^2 \frac{Y_0}{P_0} &\leq Y_0 \left(C\frac{|P_0 - Q_0|}{P_0} + CI\right)^2 \\
&\overset{(a)}{\leq} Y_0 \frac{|P_0 - Q_0|^2}{P_0^2} C + Y_0 I^2 C \\
&\overset{(b)}{\leq} \frac{|P_0 - Q_0|^2}{P_0} C + I^2 C
\end{aligned}$$

where (a) follows because $(x+y)^2 \leq 2x^2 + 2y^2$ and (b) follows because $Y_0 = \frac{\sqrt{P_0 Q_0}}{Z} = (1 + o(1))CP_0$.

Now, $I = o(1)$ so $I^2 \leq I$. Also, $I \geq (1 + o(1))(\sqrt{P_0} - \sqrt{Q_0})^2 = (1 + o(1))\frac{(P_0 - Q_0)^2}{(\sqrt{P_0} + \sqrt{Q_0})^2} = C(1 + o(1))\frac{(P_0 - Q_0)^2}{P_0}$. Therefore, we have that:

$$Y_0 \log^2 \frac{Y_0}{P_0} \leq CI \tag{H.2}$$

Now we turn our attention to the second term in equation H.1: $(1 - Y_0)\int y(x) \log^2 \frac{(1-Y_0)y(x)}{(1-P_0)p(x)} dx$. First, we observe that

$$\begin{aligned}
\left|\log \frac{(1 - Y_0)y(x)}{(1 - P_0)p(x)}\right| &\leq \frac{1}{2}\left|\log \frac{(1 - Q_0)q(x)}{(1 - P_0)p(x)} - \log Z\right| \\
&\leq \frac{1}{2}\left|\log \frac{1 - Q_0}{1 - P_0}\right| + \frac{1}{2}\left|\log \frac{q(x)}{p(x)}\right| + I/2
\end{aligned}$$

We bound $\log \frac{1-Q_0}{1-P_0}$ and $\log \frac{p(x)}{q(x)}$ separately.

$$\begin{aligned}
\left|\log \frac{1 - Q_0}{1 - P_0}\right| &= \left|\log\left(1 - \frac{Q_0 - P_0}{1 - P_0}\right)\right| \\
&\overset{(a)}{\leq} \left|\frac{Q_0 - P_0}{1 - P_0}\right| + C\left(\frac{Q_0 - P_0}{1 - P_0}\right)^2 \\
&\overset{(b)}{\leq} C\left|\frac{Q_0 - P_0}{1 - P_0}\right|
\end{aligned}$$

where (a) follows from Lemma I.2 and the fact that $\frac{1-Q_0}{1-P_0}$ is bounded. (b) follows from the fact that $\left|\frac{Q_0-P_0}{1-P_0}\right| = \left|1-\frac{1-Q_0}{1-P_0}\right| \le 1+\left|\frac{1-Q_0}{1-P_0}\right| \le C$.

Now,

$$\begin{aligned}\left|\log\frac{q(x)}{p(x)}\right| &= \left|\log\left(1-\frac{p(x)-q(x)}{p(x)}\right)\right| \\ &\overset{(a)}{\le} \left|\frac{p(x)-q(x)}{p(x)}\right| + \left(\frac{p(x)-q(x)}{p(x)}\right)^2 C \\ &\overset{(b)}{\le} C\left|\frac{p(x)-q(x)}{p(x)}\right|\end{aligned}$$

where (a) follows from Lemma I.2 and the fact that $\frac{q(x)}{p(x)}$ is bounded and (b) follows from the fact that $\left|\frac{p(x)-q(x)}{p(x)}\right| = \left|1-\frac{q(x)}{p(x)}\right| \le 1+\left|\frac{q(x)}{p(x)}\right| \le C$.

Therefore,

$$\begin{aligned}(1-Y_0)\int y(x)\left(\log\frac{1-Y_0}{1-P_0}\frac{y(x)}{p(x)}\right)^2 dx &\le (1-Y_0)\int y(x)\left\{C\frac{|Q_0-P_0|}{1-P_0}+C\frac{|p(x)-q(x)|}{p(x)}+I/2\right\}^2 dx \\ &\le (1-Y_0)\int y(x)\left\{C\left(\frac{Q_0-P_0}{1-P_0}\right)^2+C\left(\frac{p(x)-q(x)}{p(x)}\right)^2+CI^2\right\}dx\end{aligned}$$

In the last inequality, we used the fact that $(x+y+z)^2 \le 9x^2+9y^2+9z^2$.

Again, we bound this by bounding the following three terms: **term a:** $(1-Y_0)\int y(x)C\left(\frac{Q_0-P_0}{1-P_0}\right)^2 dx$, **term b:** $(1-Y_0)\int y(x)C\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx$, and **term c:** $(1-Y_0)\int y(x)CI^2dx$.

term a:

$$\begin{aligned}(1-Y_0)\int y(x)C\left(\frac{Q_0-P_0}{1-P_0}\right)^2 dx &= C\left(\frac{Q_0-P_0}{1-P_0}\right)^2\int\frac{\sqrt{(1-P_0)p(x)(1-Q_0)q(x)}}{Z}dx \\ &\overset{(a)}{\le} C(1+o(1))\left(\frac{Q_0-P_0}{1-P_0}\right)^2\int\sqrt{\frac{(1-Q_0)q(x)}{(1-P_0)p(x)}}(1-P_0)p(x)dx \\ &\overset{(b)}{\le} C(1+o(1))\left(\frac{Q_0-P_0}{1-P_0}\right)^2(1-P_0) \\ &\le C(1+o(1))\frac{(Q_0-P_0)^2}{1-P_0}\end{aligned}$$

In (a), we use the fact that $\frac{1}{Z} = (1+o(1))$ since $Z \to 1$. In (b), we use the fact that $\frac{1-Q_0}{1-P_0}$ and $\frac{q(x)}{p(x)}$ are bounded by some absolute constant.

Now, note that $I \ge (1+o(1))(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 = (1+o(1))\frac{(P_0-Q_0)^2}{(\sqrt{1-P_0}+\sqrt{1-Q_0})^2} = C(1+o(1))\frac{(P_0-Q_0)^2}{1-P_0}$. Therefore, we have that

$$(1-Y_0)\int y(x)C\left(\frac{Q_0-P_0}{1-P_0}\right)^2 dx \le C(1+o(1))I \le CI$$

Moving on to **term b**:

$$C(1-Y_0)\int y(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx = \frac{C}{Z}\int\sqrt{\frac{1-Q_0}{1-P_0}\frac{q(x)}{p(x)}}(1-P_0)p(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx$$

$$\overset{(a)}{\leq} \frac{C}{Z}(1-P_0)\int p(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx$$

where in (a), we used the fact that $\frac{1}{Z} = (1+o(1))$ and that $\frac{1-Q_0}{1-P_0}$ and $\frac{p(x)}{q(x)}$ are both bounded by an absolute constant by assumption. Now, note that $H = \int(\sqrt{p(x)}-\sqrt{q(x)})^2dx = \int \frac{(p(x)-q(x))^2}{(\sqrt{p(x)}+\sqrt{q(x)})^2}dx = C\int \frac{(p(x)-q(x))^2}{p(x)}dx$.

Therefore, we have that

$$C(1-Y_0)\int y(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx \leq C(1-P_0)H$$
$$\leq C\sqrt{(1-P_0)(1-Q_0)}H \leq CI$$

Finally, we have, for **term c**, that

$$(1-Y_0)\int y(x)CI^2dx = (1-Y_0)CI^2 \leq CI$$

Therefore, we have that

$$(1-Y_0)\int y(x)\left(\log\frac{1-Y_0}{1-P_0}\frac{y(x)}{p(x)}\right)^2 dx \leq CI$$

Combining the above bound with H.2, we can now go back to Equation H.1 and get that

$$\mathbb{E}_\Psi\left(\log\frac{Y(A_{1v})}{P(A_{1v})}\right)^2 \leq CI$$

Hence, $\sqrt{5V_\Psi(\mathcal{Q})} \leq C\sqrt{\frac{nI}{\beta K}}$.

Now, suppose $\frac{nI}{K} \to \infty$. Then, we have that $\sqrt{\frac{nI}{K}} = o\left(\frac{nI}{K}\right)$ and thus, $\sqrt{5V_\Psi(\mathcal{Q})}$ is $o(nI/K)$. Therefore, we have that $\mathbb{E}_\Phi l(\widehat{\sigma},\sigma_0) \geq \frac{1}{4}\exp\left(-(1+o(1))\frac{nI}{\beta K}\right)$.

Suppose $nI/K \to c < \infty$, then $\mathbb{E}_\Psi\mathcal{Q} = c(1+o(1))$ and $\sqrt{5V_\Psi(\mathcal{Q})} \leq C(1+o(1))$. Therefore, $\mathbb{E}_\Phi l(\widehat{\sigma},\sigma_0) > c' > 0$ for some constant c'.

I Additional useful lemmas

Lemma I.1. *Let* $I = -2\log\left(\sqrt{P_0Q_0}+\int\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right)$ *and let* $H = (\sqrt{P_0}-\sqrt{Q_0})^2+\int\left(\sqrt{(1-P_0)p(x)}-\sqrt{(1-Q_0)q(x)}\right)^2dx$. *If* $I < 1-\epsilon$, *then we have that*

$$I = H(1+\eta)$$

where $|\eta| \leq \frac{H}{4\epsilon}$. *Therefore, we have that* $I \to 0$ *iff* $H \to 0$ *and that if* $I \to 0$, $I = H(1+o(1))$.

Proof.

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

$$= -2\log\left(1 - \frac{1}{2}H\right)$$
$$= 2\frac{1}{2}H(1+\eta)$$

where $|\eta| \leq \frac{H}{2\epsilon}$. □

Lemma I.2. *Suppose $x \geq 0$ and $1 \geq \epsilon > 0$, then we have that, for all $0 \leq x < 1 - \epsilon$,*

$$\log(1-x) = -(1+\eta)x$$

where $|\eta| \leq \frac{x}{2\epsilon}$

Proof. This follows by taking the Taylor expansion of $\log(1-x)$ around $x = 0$. □

Lemma I.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,

$|f(z)| \leq \frac{1}{2} < z$.

□