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

May 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 [5, 9, 13, 16]. 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 [6, 11, 14, 17–19].

When studying community recovery, a standard model assumption is that, conditioned on the community labels of the nodes of the graph, edges are generated independently according to distributions that depend only upon the community labels of the endpoints of an edge. This is the setting of the Stochastic Block Model (SBM). We consider in this paper the homogeneous version of SBM, where all within-community edges are generated according to the Bernoulli(p) distribution and all between-communities edges are generated according to the Bernoulli(q) distribution.

Community recovery on the SBM is the problem of estimating the latent cluster memberships of the nodes from an instance of a network generated by the SBM. Astonishing progress has been made on this problem over the past decade, starting from the seminal conjecture of [CITE]. There now exists algorithms that achieve various notions of optimality, including minimum misclustering error rate, sharp detection threshold, and sharp recovery threshold.

A shortcoming of stochastic block model is that the edges are binary, which means that all edges are given equal weights in the determination of the community structure. On real world networks however, pairwise connections often have varying strength and characteristics; for example, some edges on social networks represent connections between close friends and other edges represent connections between distant acquaintances – these edges should not be given equal consideration for community recovery.

Many real world networks possess a natural weight structure that reflect the strength and characteristics of its edges. On social networks or cellular networks for example, the frequency of interactions between the individuals or users could quantify the strength of a connection. On gene co-expression networks, edges have weights that range from -1 to 1 that indicate the correlation between the expression levels of a gene pair. On co-citation network, edges have weights that represent the frequency with which two articles are co-cited. On brain neural networks, edge weights may be taken as the level of neural activity between regions in the brain. 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 consider the *weighted* setting of the stochastic block model, where, after an edge is generated from one of two Bernoulli distributions, 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. Under this model, we study the problem of estimating the cluster membership from a weighted network, without knowledge of the weight densities $p(x), q(x)$.

A critical assumption that underlies many of the existing work on weighted networks is that there is a separation between the means of between-cluster edge weight distribution $p(x)$ and within-cluster edge weight distribution $q(x)$. This assumption is common because it allows the application of some of the existing algorithms for the unweighted SBM–such as spectral clustering–to a weighted network with little to no modification. We do not make this assumption in our paper: our setting allows $p(x)$ and $q(x)$ to have the same mean. We make no assumption on the means of $p(x)$ and $q(x)$ not because we think that such assumptions are unreasonable, but because the edge weight distributions $p(x), q(x)$ may significantly differ in other aspects such as variance or higher moments. It is necessary to consider information beyond mean separation in order to achieve optimal performance.

Other existing approaches to the weighted networks often assume that the $p(x), q(x)$ belong to a parametric family. In our paper however, we let the densities $p(x), q(x)$ be nonparametric and impose only mild regularity conditions. The fact that we do not assume any parametric form of $p(x), q(x)$ adds a new challenge for community recovery: 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 from which density an edge weight is drawn without knowing the latent cluster structure.

The goal of our paper is to characterize the optimal rate of misclustering error for the weighted stochastic block model. From one side, we prove an information theoretic lower bound on the performance of any community recovery algorithms on the unweighted SBM. [TODO: mention permutation equivariance] From the other side, we present a computationally tractable algorithm whose rate of convergence matches the lower bound. Our results show that the optimal rate on the weighted SBM is governed by the 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 $q(x)$. This generalizes, in a natural way, the results of Zhang and Zhou [21], which show the optimal rate of the unweighted SBM to be governed by the Renyi-divergence of order $1/2$ between two Bernoulli distributions that capture the divergence between the edge probabilities only.

Furthermore, we show that the optimal rate for the weighted SBM depends on the densities $p(x), q(x)$ *only through* their Renyi-divergence. This implies that the optimal rate is adaptive to the densities $p(x), q(x)$, that is, it is possible, without knowing $p(x)$ and $q(x)$, to achieve same optimal rate as if one *does* know $p(x), q(x)$. And, in the cases where the densities comes from a parametric family, it is possible, without making any parametric assumptions, to get the same optimal rate as if one imposes the true parametric form. This is in constrast to most nonparametric problems in statistics where the nonparametric method usually have slower rate of convergence than parametric methods in the settings where a parametric form is known. This observation reflects an important intuition behind our results, that on the weighted SBM, one does not need to estimate the densities well in order to recover the clusters well.

The algorithm that we devise is based on discretization. In the cases where the weights are bounded, we discretize the weights with a uniformly spaced binning to convert the weighted SBM into an instance of colored or labeled SBM, where each of the edges are marked with a color from a set of colors whose cardinality is finite but divergent; we then solve the colored SBM by extending the now familiar coarse-to-fine clustering algorithm that computes an initialization through spectral clustering and then performs refinement through node-wise likelihood maximization. In the cases where the weights are unbounded, we reduce the problem to the bound case: we first apply a transformation on the edge weights so that the transformed weights are bounded. Our lower bound analysis uses and extends the change-of-measure proof technique. Our lower bound result applies to all parameters in the parameter space and is thus more than minimax. It does restrict the space of estimators to algorithms whose output does not depend on how the nodes are labeled – a property we call permutation equivariance – but this constraint is a mild one and satisfied by all effective clustering algorithms.

[TODO: fix this to the new structure] 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 4 describes our proposed algorithm for finding communities on the weighted SBM. In Section 5 we provide the statements of our main results concerning the behavior of our algorithm in terms of misclassification error rates and exact recovery. Section 6 highlights the key technical components employed in the analysis of our algorithm. We close in Section 7 with further implications of our work and open questions related to our results.

1.1 Notation

For a positive integer n, we use $[n]$ to denote the set $\{1, ..., n\}$ and S_n to denote the set of permutations on $[n]$. We let $o(1)$ denote a sequence indexed by n that tends to 0 as $n \to \infty$ and $\Theta(1)$ denote a sequence indexed by n that is bounded away from 0 and ∞ as $n \to \infty$. For two real numbers a, b, we let $a \vee b$ denote $\max(a, b)$ and $a \wedge b$ denote $\min(a, b)$.

2 Model and problem formulation

We begin with a formal definition of the weighted stochastic block model and a description of the community recovery problem.

2.1 Weighted Stochastic Block Model

We let n denote the number of nodes in the network and $K \geq 2$ denote the number of communities. We suppose that the communities are approximately balanced in that there exists a *cluster-imbalance constant* β such that the cluster size n_k for each cluster $k = 1, \ldots, K$ satisfies $\frac{\beta n}{K} \geq n_k \geq \frac{n}{\beta K}$.

Definition 2.1. We let $\sigma_0 : [n] \to [K]$ denote the true cluster assignment, i.e., for each node u, $\sigma_0(u) \in \{1, 2, \dots, K\}$ represent the cluster label of u. Because a clustering does not depend on how the clusters are labeled, we say that two cluster assignments σ, σ' are **equivalent** if there exists a permutation $\tau \in S_K$ such that $\tau \circ \sigma = \sigma'$.

For the homogeneous unweighted stochastic block model, the parameters consists of (n, K, β, σ_0) in conjunction with between-cluster edge probability p and within-cluster edge probability q. The unweighted SBM is then the following probability distribution over adjacency matrices $A \in \{0,1\}^{n \times n}$:

Definition 2.2. (Unweighted Homogeneous Stochastic Block Model)
For all pairs $\{(u,v) : u, v \in [n], u < v\}$, the binary entries A_{uv}'s are generated independently as such:

$$A_{uv} \sim \begin{cases} Ber(p) & \text{if } \sigma_0(u) = \sigma_0(v) \\ Ber(q) & \text{if } \sigma_0(u) \neq \sigma_0(v). \end{cases}$$

In the more general case of *heterogenous* SBM, we have a matrix $P \in \mathbb{R}^{K \times K}$ of probabilities instead of two numbers p, q. The edge random variables are generated independently as $A_{uv} \sim Ber(P_{\sigma_0(u), \sigma_0(v)})$. We focus on the homogeneous case in this paper and discuss how our results may extend to the heterogenous setting.

To incorporate real-valued edge weights in the SBM, we model each edge weight as a random variable; an edge weight, conditioned on the event that the edge exists, has the density $p(x)$ if the corresponding edge is between-cluster and $q(x)$ if the corresponding edge is within-cluster. The parameters for the weighted SBM consists of (n, K, β, σ_0) in addition to P_0, Q_0, which are the edge *absence* probabilities, and $p(x), q(x)$, which are the edge weight densities. We let S denote the support of $p(x), q(x)$, which could be either $S = [0,1]$, $S = \mathbb{R}$, or $S = \mathbb{R}^+$. The weighted SBM is then the following probability distribution over adjacency matrices $A \in S^{n \times n}$.

Definition 2.3. (Weighted Homogeneous Stochastic Block Model)
For all pairs $\{(u,v) : u, v \in [n], u < v\}$, we first independently generate the edge presence indicator variables Z_{uv}'s:

$$Z_{uv} \sim \begin{cases} Ber(1 - P_0) & \text{if } \sigma_0(u) = \sigma_0(v) \\ Ber(1 - Q_0) & \text{if } \sigma_0(u) \neq \sigma_0(v). \end{cases}$$

and then we independently generate the real-valued edge weights:

$$A_{uv} \sim \begin{cases} 0 & \text{if } Z_{uv} = 0 \\ p(x) & \text{if } Z_{uv} = 1 \text{ and } \sigma_0(u) = \sigma_0(v) \\ q(x) & \text{if } Z_{uv} = 1 \text{ and } \sigma_0(u) \neq \sigma_0(v). \end{cases}$$

We can define the weighted SBM more succinctly by defining probability measures P, Q to be mixed distributions where the singular part of P is a point mass at 0 with probability P_0 and the continuous part P is $(1 - P_0)p(x)$ and likewise for Q. Under this representation, the weighted SBM is the following equivalent definition:

Definition 2.4. (Weighted Homogeneous SBM)
For all pairs $\{(u,v) : u, v \in [n], u < v\}$, the real-valued entries A_{uv}'s are generated independently as such:

$$A_{uv} \sim \begin{cases} P & \text{if } \sigma_0(u) = \sigma_0(v) \\ Q & \text{if } \sigma_0(u) \neq \sigma_0(v) \end{cases}$$

We observe that if P, Q are Bernoulli distributions without any continuous parts, then the weighted SBM reduces to the unweighted version. It is possible to generalize the weighted SBM to a *weighted and labeled* stochast block model by letting the singular parts of P, Q possess additional point masses. We focus on the weighted SBM in this paper but our theory extends to simple cases of weighted and labeled SBM as well.

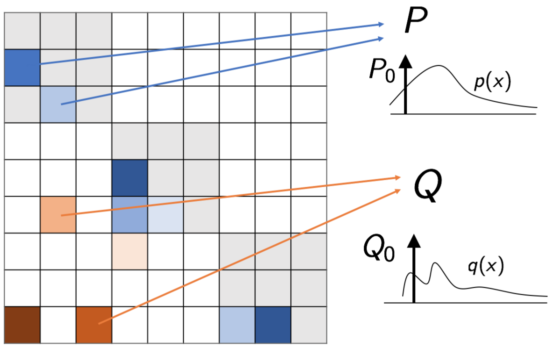

Figure 1: Weighted stochastic block model

2.2 Community estimation

Given an observation $A \in S^{n\times n}$ generated from the SBM, the problem of community estimation is to estimate the true cluster membership structure σ_0. We assume throughout the paper that the number of clusters K is known so that the only information available to a community recovery algorithm consists of A and K. [TODO:Discuss choosing K]

We evaluate the performance of a community recovery algorithm by looking at its misclustering error. Before formally defining misclustering error, let us first define some notation. For a clustering algorithm $\widehat{\sigma}$, let $\widehat{\sigma}(A) : [n] \to [K]$ be the clustering outputted by $\widehat{\sigma}$ when given A as an input.

Definition 2.5. We define the *misclustering error* to be

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

where $d_H(\cdot,\,\cdot)$ denotes the Hamming distance. We then define the risk of an estimator as

$$R(\widehat{\sigma}, \sigma_0) = \mathbb{E} l(\widehat{\sigma}(A), \sigma_0)$$

where the expectation is taken with respect to both the random network A and potential randomness in the algorithm $\widehat{\sigma}$. The loss function $l(\cdot, \cdot)$ invovles a minimization over the set of permutations τ on K cluster labels because clusterings are invariant with respect the labeling of the clusters.

The goal of this paper is to characterize the minimal achievable risk for community recovery on the weighted SBM in terms of the parameters $(n, K, \beta, \sigma_0, P_0, Q_0, p(x), q(x))$.

2.3 Permutation Equivariance

The cluster structure in a network does not depend on how the nodes are labeled. Therefore, it is natural to consider only estimation algorithms that output equivalent clusterings when given graph-isomorphic networks as inputs. We formally define this property as *permutation equivariance*. Permutation equivariance is a natural condition that most papers assume implicitly; we give a formal treatment here in order to state our lower bound.

Definition 2.6. For an $n \times n$ matrix A and a permutation $\pi \in S_n$, define πA as an $n \times n$ matrix such that $A_{uv} = [\pi A]_{\pi(u),\pi(v)}$. In other words, if for all nodes u, one relabels u as $\pi(u)$, then πA is the resulting adjacency matrix.

Let $\widehat{\sigma}$ be a deterministic clustering algorithm; i.e., $\widehat{\sigma}(A)$ is a clustering $[n] \to [K]$ for any $n \times n$ matrix A. $\widehat{\sigma}$ is *permutation equivariant* if, for any A and any $\pi \in S_n$,

$$\widehat{\sigma}(\pi A) \circ \pi \text{ is equivalent to } \widehat{\sigma}(A)$$

In other words, there exists $\tau \in S_K$ such that $\widehat{\sigma}(\pi A) \circ \pi = \tau \circ \widehat{\sigma}(A)$, i.e., $l(\widehat{\sigma}(A), \widehat{\sigma}(\pi A) \circ \pi) = 0$. We note that $\widehat{\sigma}(\pi A)$ by itself is not equivalent to $\widehat{\sigma}(A)$ because the nodes in πA are labeled with respect to the permutation π.

It is straightforward to extend definition 2.6 to randomized algorithms.

Definition 2.7. A randomized clustering algorithm $\widehat{\sigma}$ is permutation equivariant if, for all A and $\pi \in S_n$, and for all functions $\sigma : [n] \to [K]$,

$$P\Big(\widehat{\sigma}(A) \text{ equivalent to } \sigma\Big) = P\Big(\widehat{\sigma}(\pi A) \text{ equivalent to } \sigma\Big)$$

where the probability is taken with respect to randomization in the algorithm $\widehat{\sigma}$.

Permutation equivariance is a natural property satisfied by all the clustering algorithms that are proposed in literature except perhaps those that use extra side information in addition to the given network. In section 5.2, we study permutation equivariance in detail and give some properties of permutation equivariant estimators.

3 Overview of main results

Our results are asymptotic in n and we treat $P_0, Q_0, p(x), q(x), \sigma_0$ as varying with n; at various places, we do not explicitly show this dependence in our notation in order to simplify the presentation. We hold K, β to be fixed with respect to n.

Let P be the probability measure on S (recall that S can be $[0,1]$,, $\mathbb{R}$, or $\mathbb{R}^+$) induced by $(P_0, p(x))$, that is, the singular part of P is a point mass with probability P_0 at 0 and the continuous part of P has $(1-P_0)p(x)$ as its Radon-Nikodym derivative with respect to the Lebesgue measure. Let Q be defined likewise. The optimal rate of misclustering error naturally depends on the extent to which P and Q are different. The notions of divergence between P and Q that arises in the analysis of the weighted stochastic block model is the Renyi-divergence of order 1/2, which we denote by I:

$$I = -2\log \int \left(\frac{dP}{dQ}\right)^{1/2} dQ = -2\log\left(\sqrt{P_0 Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right) \tag{3.1}$$

Since P, Q depends on n, I depends on n as well. The fact that P, Q vary with n is analogous to analyses on the unweighted SBM [TODO:cite] where people make the common sparse graph assumption that $p, q \to 0$ as $n \to \infty$. Zhang and Zhou [TODO:cite] considers a more general setting where $|p - q|$ is assumed to converge to 0. In similar fashion, we focus on the case where P and Q tend toward each other in the sense that $I \to 0$; this case encompasses the sparse graph setting where the edge absence probabilities $P_0, Q_0 \to 1$ as $n \to \infty$. [TODO:why large I is uninteresting]

Under the setting where $I = o(1)$, we can characterize I in terms of the Hellinger distance: [TODO:reference lemma]

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

$$\cdot (1 + o(1)) \tag{3.2}$$

We make two observations from equation 3.2. First, we see that I decomposes into two terms: term 1 captures divergence between the edge presence probabilities and term 2 captures the divergence between the edge weight densities.

Second, under the weighted SBM, a network whose edges are dense, i.e. P_0, Q_0 do not tend to 1, is still interesting to analyze because it could be that the edge weight densities are very similar, i.e. $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$.

We first state the lower bound.

Theorem. *(Informal Statement)*
Let A be generated from a weighted SBM with parameters $(n, K, \beta, \sigma_0, P_0, Q_0, p(x), q(x))$. Under regularity condition on the densities $p(x), q(x)$, for any permutation equivariant estimator $\widehat{\sigma}$:

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

For the upper bound, we construct an algorithm $\widehat{\sigma}$ whose misclustering error is no more than $\exp\left(-(1+o(1))\frac{nI}{\beta K}\right)$ with high probability.

Theorem. *(Informal Statement)*
There exists a permutation equivariant algorithm $\widehat{\sigma}$ that satisfies, under regularity conditions on $p(x), q(x)$:

$$P\left(l(\widehat{\sigma}(A), \sigma_0) \geq \exp\left(-(1 + o(1))\frac{nI}{\beta K}\right)\right) \to 0$$

Because the upper bound is stated as convergence in probability, we cannot directly compare it to the lower bound on the risk. However, the same proof of the upper bound can be used to show the following:

$$\begin{array}{ll} \text{if } \dfrac{nI}{\beta K \log n} \leq 1 & \mathbb{E} l(\widehat{\sigma}(A), \sigma_0) \leq \exp\left(-(1 + o(1))\dfrac{nI}{\beta K}\right) \\ \text{if } \dfrac{nI}{\beta K \log n} > 1 & P\Big(l(\widehat{\sigma}(A), \sigma_0) = 0\Big) \geq 1 - \dfrac{1}{n} \end{array}$$

From this, we observe that in the regime where $\frac{nI}{\beta K \log n} \leq 1$, the optimal risk is tightly characterized as $\exp\left(-(1+o(1))\frac{nI}{\beta K}\right)$. In the regime where $\frac{nI}{\beta K \log n} > 1$, the optimal risk is lower bounded by $\exp\left(-(1+o(1))\frac{nI}{\beta K}\right)$; we have no corresponding upper bound but this second regime is less interesting than the first one because it is possible in this this to recover the true clustering σ_0 exactly with high probability.

3.1 Relation to Previous Work

[TODO: elaborate more]

Our result is a natural generalization of the work by Zhang and Zhou [21] in which they show that the minimax optimal risk for the unweighted stochastic block model is $\exp\left(-(1+o(1))\frac{nI_{Ber}}{\beta K}\right)$ where $I_{Ber} = \sqrt{pq} + \sqrt{(1-p)(1-q)}$. If we let P be Bernoulli(p) and Q be Bernoulli(q), then I exactly reduces to I_{Ber}. The algorithm that achieves optimality in Zhang and Zhou [21] is intractable but a computationally feasible version was later developed by Gao et al [7].

Our result is also analogous to that of Yun and Proutiere [20], who studied, under certain assumptions and with respect to a prior on the cluster assignment σ_0, the optimal risk for the heterogenous labeled stochastic block model with finite number of labels. They characterize the optimal rate under a seeming different notion of divergence but if we specialize their result to the homogenous setting, we find that their notion of divergence reduces to the Renyi-divergence of order $\frac{1}{2}$ between two discrete distributions over a fixed finite number of labels.

[TODO: re-structure next paragraph]

Weighted networks have received some attention in the physics community [4, 12], their statistical properties are not well understood. Stochastic block model in particular is rarely studied under the weighted network setting. One exception is the work by the work by Aicher et al [3], which also proposes a notion of weighted SBM by modeling the edge weights as generated by a distribution from a known exponential family. Hajek et al [10] considers a model similar to the one we propose except that it contains one community and that the distributions P, Q are known.

3.1.1 Other Notions of Recovery

A closely related problem is that of finding the exact recovery threshold. We say that the stochastic block model has an exact recovery threshold if there is some function of the parameters $\theta(p, q, n, K, \beta, \sigma_0)$ 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 [1] 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 [21] beyond the $\log n$ degree setting where $\frac{n I_{\text{Ber}}}{K \log n}$ was shown to be the threshold. Apart from exact recovery (also known as strong consistency), a notion of detection threshold has also been considered [15].

[TODO: relate our result to exact recovery threshold]

4 Estimation algorithm

The weighted stochastic block model presents an extra layer of difficulty because the densities $p(x)$ and $q(x)$ are unknown. One consequence of not knowing $p(x)$ and $q(x)$ is that the MLE does not exist, for the same reason that the MLE 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. A natural thought is to estimate the edge weight densities $p(x)$ and $q(x)$, but this is hindered by the fact that we do not know whether a edge weight observation originates from $p(x)$ or $q(x)$ without knowing the cluster structure.

[TODO: Why can't we just plug the weighted network into spectral clustering?]

[TODO: plug in two sample test?]

Our approach is appreciably different and consists of combining the idea of discretization from nonparametric density estimation with clustering techniques for the unweighted stochastic block model.

4.1 Algorithm overview

We now outline the main components of our algorithm. The key ideas are to convert the edge weights into a finite set of labels by discretization, and then cluster nodes on the labeled network. We first provide a broad overview of our algorithm and then describe each step 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 & discretization.** We take as input a weighted matrix A and apply an invertible transformation function $\Phi : S \rightarrow [0, 1]$ to obtain a matrix $\Phi(A)$ with weights between 0 and 1.

   Next, we divide the $[0, 1]$ interval into $l = 1, \dots, L$ equally-spaced subintervals, which we call bins. We replace the real-valued weight entries $\Phi(A)$ with a categorical label $l \in \{1, \dots, L\}$: $[\Phi(A)]_{uv}$ is assigned label l if the value $[\Phi(A)]_{uv}$ falls into bin l. We output a network with each edge assigned one of L possible colors. We continue to denote the adjacency matrix by A.

2. **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. Again, we continue to denote the modified adjacency matrix as A.

3. **Initialization Parts 1 & 2.** For each color l, we create a sub-network by including only edges of color l. For each sub-network, we perform spectral clustering. We output l^*, the color that induces the maximally separated spectral clustering.

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

4. **Refinement & consensus.** From each $\widetilde{\sigma}_u$, we generate a clustering $\widehat{\sigma}_u$ on $\{1, 2, \dots, n\}$ which retains the assignments specified by $\widetilde{\sigma}_u$ for $\{1, 2, \dots, n\} \setminus \{u\}$, and assigns $\widehat{\sigma}_u(u)$ by maximizing the likelihood taking into account only the neighborhood around u.

   We then align the cluster assignments made in the previous step.

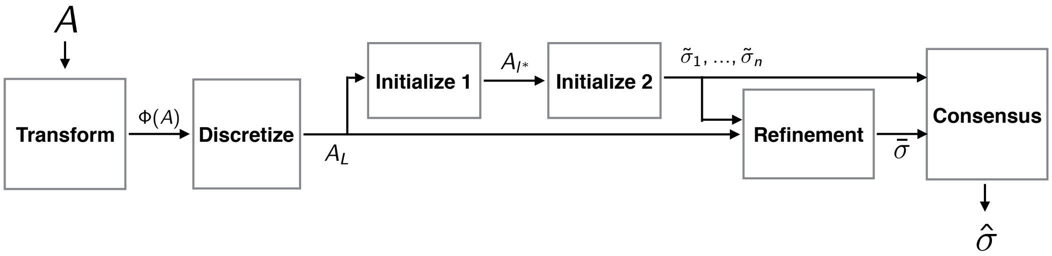

Figure 2: Add a box indicating the add noise step. Pipeline for the our proposed algorithm

4.2 Transformation and discretization

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

Algorithm 4.1 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 $A$
```

4.3 Add noise

This part of the algorithm is required for technical reasons. As detailed in the proof of Proposition 6.1 in Appendix A, deliberately forming a noisy version of the graph barely affects the separation between the distributions specifying within- and between-community edge labels, but has the desirable effect of ensuring that all edge labels occur with probability at least $\frac{c}{n}$. This property is crucial to our analysis in subsequent steps of the recovery algorithm.

In the description of the algorithm below, we treat the label 0 (i.e., an empty edge) as a separate label, so we have a network with a total of $L+1$ labels.

Algorithm 4.2 Add noise

```
Input: A labeled network with $L+1$ labels and a constant $c$
Output: A labeled network $A$ with $L+1$ labels

  for every edge $(u,v)$ do
      With probability $1 - \frac{c(L+1)}{n}$, do nothing. With probability $\frac{c(L+1)}{n}$ replace edge label with a
  label drawn uniformly at random from $\{0,1,2,\ldots,L\}$
  end for
  Output $A$
```

4.4 Initialization

The initialization procedure takes as input a network with edges labeled $\{1,\ldots,L\}$. The goal of the initialization procedure is to create a rough clustering $\widetilde{\sigma}$ that is suboptimal but still consistent. As outlined in Algorithm 4.3, the rough clustering is based on a single label l^*, selected based on the maximum value of the estimated Renyi divergence between within-community and between-community distributions for the unweighted SBMs based on individual labels.

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

Spectral clustering: Note that Algorithm 4.3 involves several applications of SPECTRAL CLUSTERING. We describe the spectral clustering algorithm used as a subroutine in Algorithm 4.4 below:

Importantly, note that we may always choose the parameter μ sufficiently large such that Algorithm 4.4 generates a set S with $|S| = K$.

Algorithm 4.3 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 $\{\widetilde{\sigma}_u\}_{u=1,\dots,n}$.
```

Algorithm 4.4 Spectral clustering

```
Input: An unweighted network $A$, trim threshold $\tau$, number of communities $K$, tuning parameter $\mu$
Output: A clustering $\sigma$

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, formed by truncating the SVD
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
```

4.5 Refinement and consensus

Our refinement and consensus steps closely follow the method described by Gao et al. [7]. 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 by locally maximizing an approximate log-likelihood expression for each of the nodes $u = 1,\dots,n$. The consensus step then resolves a cluster label consistency problem arising after the refinement stage.

Algorithm 4.5 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
for each node $u$ do
    Estimate $\{\widehat{P}_l, \widehat{Q}_l\}_{l=0,\dots,L}$ from $\widetilde{\sigma}_u$
    Let $\widehat{\sigma}_u : [n] \to [K]$ where $\widehat{\sigma}_u(v) = \widetilde{\sigma}_u(v)$ for all $v \neq u$ and
        $\widehat{\sigma}_u(u) = \arg\max_k \sum_{v \,:\, \widetilde{\sigma}_u(v)=k,\, v\neq u} \sum_l \log \frac{\widehat{P}_l}{\widehat{Q}_l} \mathbf{1}(A_{uv} = l)$
end for
Let $\widehat{\sigma}(1) = \widehat{\sigma}_1(1)$                                   ▷ Consensus Stage
for each node $u \neq 1$ do
        $\widehat{\sigma}(u) = \arg\max_k |\{v \,:\, \widehat{\sigma}_1(v) = k\} \cap \{v \,:\, \widehat{\sigma}_u(v) = \widehat{\sigma}_u(u)\}|$
end for
Output $\widehat{\sigma}$
```

5 Optimal Misclustering Error

In this section, we give the formal statements of our main results.

Literatures on the unweighted SBM typically study the sparse graph case where the within-cluster probability p and the between-cluster probability q are dependent on n and both converging to 0. Zhang and Zhou [21] considers a more general setting where $|p - q|$ is assumed to converge to 0. In similar fashion, we also assume that the probability measures P, Q vary with n tend toward each other in the sense that $I \to 0$.

In this section, we explicitly use subscript to denote all quantities that vary with n. For example, we write $p_n(x), q_n(x)$ in place of $p(x), q(x)$ and we write

$$I_n = -2 \log \left(\sqrt{P_{0,n} Q_{0,n}} + \int_S \sqrt{(1 - P_{0,n})(1 - Q_{0,n}) p_n(x) q_n(x)} dx \right)$$

Assumption A0: There exist absolute constants $c_0 > 0$ such that $\frac{1}{c_0} \leq \frac{1 - P_{0,n}}{1 - Q_{0,n}} \leq c_0$. If $P_{0,n} \vee Q_{0,n} > 0$, we also assume $\frac{1}{c_0} \leq \frac{P_{0,n}}{Q_{0,n}} \leq c_0$.

Assumption A0 states that the density of within-community edges has the same order as the density of between-community edges. This assumption is standard in the existing literature on unweighted stochastic block models.

Let $\mathcal{P}$ be the set of all densities on S and let $(\mathcal{P} \times \mathcal{P})^\infty$ denote the set of infinite sequences of density pairs, i.e., $(\mathcal{P} \times \mathcal{P})^\infty = \left\{ (p_n, q_n)_{n=1,\dots,\infty} : p_n, q_n \geq 0, \int_S p_n d\lambda = \int_S q_n d\lambda = 1 \right\}$, where λ denotes the Lebesgue measure.

5.1 Upper Bound

Definition 5.1. Let S be either $[0, 1]$ or $\mathbb{R}$ or $\mathbb{R}^+$. We say that $\Phi : S \to [0, 1]$ is a *transformation function* if it is a differentiable bijection (hence a cumulative distribution function), and $\phi = \Phi'$

satisfies (a) $\int_S \phi(x)^s dx < \infty$ for any $0 < s < 1$, and (b) $\left|\frac{\phi'(x)}{\phi(x)}\right|$ is bounded.

For $S = [0,1]$, we always take Φ as the identity. For $S = \mathbb{R}$ or $\mathbb{R}^+$, Φ must be chosen so that ϕ is not too heavy tailed in order to satisfy (a) and not too light tailed in order to satisfy (b). For example, we may take ϕ as:

$$\phi(x) = \frac{e^{1-\sqrt{x+1}}}{4}, \qquad \text{when } S = \mathbb{R}^+, \tag{5.1}$$

$$\phi(x) = \frac{e^{1-\sqrt{|x|+1}}}{8}, \qquad \text{when } S = \mathbb{R}. \tag{5.2}$$

The transformation function Φ induces a probability measure on S, and we let $\Phi\{\cdot\}$ denote the Φ-measure of a set.

Definition 5.2. We call a nonnegative function $f(x)$ is (c_{s1}, c_{s2}, C_s)-*bowl-shaped* if $f(x)$ is nonincreasing for all $x \leq c_{s1}$, nondecreasing for all $x \geq c_{s2}$, and bounded by C_s for all $x \in [c_{s1}, c_{s2}]$.

[TODO: describe the two cases on H_n]

5.1.1 The case $H_n = \Theta(1)$

We state the required assumptions and then present our main result.

Define $\mathcal{G}_\Phi^{\text{upper}} \subset (\mathcal{P} \times \mathcal{P})^\infty$ as sequences of densities $p_n(x), q_n(x)$ that satisfy $H \equiv \int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 dx = \Theta(1)$ as well as the following regularity conditions:

A1 $p_n(x), q_n(x) > 0$ on the interior of S, $\sup_x \{p_n(x) \vee q_n(x)\} < \infty$, and $\inf_{x \in S} \left\{\frac{p_n(x) \vee q_n(x)}{\phi(x)}\right\} < \infty$.

A2 For an absolute constant $r \geq 8$, $\int \left|\log \frac{p_n(x)}{q_n(x)}\right|^r \phi(x) dx < \infty$.

A3 There exists a function $h_n(x)$ such that

(a) $h_n(x) \geq \max\left\{\left|\frac{q_n'(x)}{q_n(x)}\right|, \left|\frac{p_n'(x)}{p_n(x)}\right|\right\}$,

(b) $h_n(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped for some absolute constants c_{s1}, c_{s2}, and C_s, and

(c) For some constant t such that $\frac{4}{r} < 2t < 1$, we have

$$\sup_n \int |h_n(x)|^{2t} \phi(x) dx < \infty.$$

A4 $(\log p_n)'(x), (\log q_n)'(x) \geq (\log \phi)'(x)$ for all $x < c_{s1}$, and $(\log p_n)'(x), (\log q_n)'(x) \leq (\log \phi)'(x)$ for all $x > c_{s2}$.

We also note that we allow $c_{s1} = 0$ in the case where $S = \mathbb{R}^+$ and we allow $c_{s1} = 0$ and $c_{s1} = 1$ in the case where $S = [0,1]$.

These conditions depend on the choice of Φ, but it is enough to choose ϕ as a heavy-tailed density where all moments exist in order for the class $\mathcal{G}_\Phi^{\text{upper}}$ to be very broad. In particular, we show in section 5.1.5 that choosing Φ as 5.3 or 5.4 suffices for $\mathcal{G}_\Phi^{\text{upper}}$ to encompass Gaussian, Laplace, and other broad classes of densities.

We then have our upper bound.

Theorem 5.1. *Suppose $I_n \to 0$ and $nI_n \to \infty$. Let $\widehat{\sigma}$ be the algorithm described in section 4 with transformation function Φ and discretization level L_n chosen such that $L_n \to \infty$ and $\frac{nI_n}{L_n \exp(L_n^{1/r})} \to \infty$. Suppose that $P_{0,n}$ and $Q_{0,n}$ satisfy Assumption A0 and $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{upper}$, that is, they satisfy Assumptions A1–A4 with respect to Φ. Then,*

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

For the proof of Theorem 5.1, see Appendix F.2.

5.1.2 The case $H_n = o(1)$

We begin by stating the required assumptions.

Define $\mathcal{G}_\Phi^{\text{upper}\prime} \subset (\mathcal{P} \times \mathcal{P})^\infty$ as sequences of densities $p_n(x), q_n(x)$ that satisfy $H_n \equiv \int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 dx = o(1)$ as well as the following regularity conditions:

A1' $p_n(x), q_n(x) > 0$ on the interior of S, $\sup_n \sup_x \{p_n(x) \vee q_n(x)\} < \infty$, and $\sup_n \inf_{x\in S} \left\{ \frac{p_n(x) \vee q_n(x)}{\phi(x)} \right\} < \infty$.

A2' There is an absolute constant ρ such that there exist subintervals R_n of S where

(a) $\frac{1}{\rho} \leq \frac{p_n(x)}{q_n(x)} \leq \rho$ for $x \in R_n$ and

(b) $\Phi\{R_n^c\} = o(H_n)$.

A3' Denoting $\alpha_n^2 = \int_{R_n} q_n(x) \left(\frac{p_n(x) - q_n(x)}{q_n(x)} \right)^2 dx$ and $\gamma(x) = \frac{q_n(x) - p_n(x)}{\alpha_n}$, there exists an absolute constant $r > 4$ such that

$$\sup_n \int_{R_n} q_n(x) \left| \frac{\gamma_n(x)}{q_n(x)} \right|^r dx < \infty,$$

A4' There exists a function $h(x)$ such that

(a) $h_n(x) \geq \max\left\{ \left| \frac{\gamma_n'(x)}{q_n(x)} \right|, \left| \frac{q_n'(x)}{q_n(x)} \right|, \left| \frac{\gamma_n(x)}{q_n(x)} \right| \right\}$,

(b) $h_n(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped for absolute constants c_{s1}, c_{s2}, and C_s, and

(c) for an absolute constant t where $\frac{4}{r} < 2t < 1$, we have

$$\sup_n \int_{R_n} |h_n(x)|^{2t} \phi(x) dx < \infty.$$

A5' $(\log p_n)'(x), (\log q_n)'(x) \geq (\log \phi)'(x)$ for all $x \leq c_{s1}$, and $(\log p_n)'(x), (\log q_n)'(x) \leq (\log \phi)'(x)$ for all $x \geq c_{s2}$.

Again, we note that we allow $c_{s1} = 0$ in the case where $S = \mathbb{R}^+$ and we allow $c_{s1} = 0$ and $c_{s1} = 1$ in the case where $S = [0, 1]$.

We have the following result:

Theorem 5.2. *Suppose $I_n \to 0$ and $nI_n \to \infty$. Let $\widehat{\sigma}$ be the algorithm described in Section 4 with transformation Φ and discretization level L_n chosen such that $L_n \to \infty$, $L_n = o(\frac{1}{H_n})$, and $L_n = o(nI_n)$. Suppose $P_{0,n}, Q_{0,n}$ satisfy Assumption A0 and $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{upper\prime}$, that is, they satisfy Assumptions A1'–A5' with respect to Φ. Then,*

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

The proof of Theorem 5.2 is outlined in Appendix F.1.

5.1.3 Additional discussion of assumptions

It is crucial to note that our algorithm does not require any prior knowledge about the form of $p_n(x)$ and $q_n(x)$: the same algorithm and guarantees apply so long as $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{\text{upper}}$, i.e. they satisfy conditions A1-A4 or if $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{\text{upper}\prime}$, i.e. they satisfy conditions A1'-A5'. As we show in section 5.1.5, the conditions are mild and satisfied by the Gaussian, Laplace scale and location family, Gamma shape and scale family, as well as other broad families of distributions.

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

Interpretation of Conditions A1–A4, A1'–A5':

A1 Condition A1 is simple; the last part states that ϕ must have a tail at least as heavy as that of $p_n(x)$ and $q_n(x)$.

A2 Condition A2 requires the likelihood ratio to be integrable. It is analogous to a bounded likelihood ratio condition but much weaker.

A3 Condition A3 controls the smoothness of the derivatives of $\log p_n(x)$ and $\log q_n(x)$. We add a mild bowl-shape constraint for technical reasons related to the analysis of binning.

A4 Condition A4 is a mild shape constraint on $p_n(x)$ and $q_n(x)$. When $S = \mathbb{R}$, this condition essentially requires p_n, q_n to be monotonically increasing in x for $x \to -\infty$ and decreasing in x for $x \to \infty$.

An analogous interpretation may be used to describe to the conditions A1'–A5'. We must pay special attention to A2' and A3' however because of the $H_n = o(1)$ assumption. In condition A2', we require that the likelihood ratio $\frac{p_n(x)}{q_n(x)}$ be bounded away from 0 and ∞ except on a region $R_n^c \subseteq \mathbb{R}_n$. Since $H \to 0$, we the densities $p_n(x)$ and $q_n(x)$ are becoming increasingly similar and R_n^c is shrinking. We require that the measure of R_n^c, with respect to Φ, shrinks faster than H. This condition intuitively states that $|\frac{p_n(x)}{q_n(x)}|$ and its reciprocal tend to infinity slowly with respect to x.

In condition A3', note that $H \to 0$ implies $\alpha \to 0$, as well, so $\gamma_n(x) = \frac{p_n(x) - q_n(x)}{\alpha_n}$ is a function of constant order. The integrability condition on $\gamma_n(x)$ states that $|p_n(x) - q_n(x)|$ must converge to 0 almost uniformly for all x in the region R_n. (Having an L_∞-bound on γ_n would imply uniform convergence.)

[TODO:how does choice of Φ affect anything?]

5.1.4 Examples for $S = [0, 1]$

When $S = [0, 1]$, we always take Φ as the identity. The transformation is not needed when $S = [0, 1]$, but we still define Φ so that we can present the results in a unified format.

In the case where $H_n = \Theta(1)$, the simplest example of $p_n(x), q_n(x)$ that satisfy conditions A1-A4 is if $p_n(x), q_n(x)$ are bounded away from 0 and ∞ on $[0, 1]$ uniformly in n and if $|p_n'(x)|, |q_n'(x)|$ are also bounded uniformly in n. A1-A4 however is much more general in that it allows p_n, q_n to be 0 and p_n', q_n' to be infinity at the boundary points $0, 1$. In the case where $H_n = o(1)$, p_n, q_n satisfy A1'-A5' if they satisfy the boundedness conditions described above and if the function $x \mapsto \frac{p_n(x) - q_n(x)}{\|p_n - q_n\|_2}$ is also, uniformly in n, bounded away from 0 and infinity in its value and bounded away from infinity in its first derivative.

5.1.5 Examples for $S = \mathbb{R}$ or $\mathbb{R}^+$

We start with a proposition that allows us to generate large classes of examples. The proposition considers the case where the p_n, q_n both belong to a family of distributions that is smoothly parametrized by a parameter vector θ.

Let $f_\theta : S \to \mathbb{R}$ be a class of function indexed by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^{d_\Theta}$. Suppose Θ is compact and the class $\{f_\theta\}_\theta$ satisfy the following conditions with respect to Φ:

B1 $\inf_{\theta\in\Theta} \inf_x \{\log \phi(x) - f_\theta(x)\} > -\infty$.

B2 The Fisher information matrix $G_\theta := \int_S (\nabla_\theta f_\theta(x))(\nabla_\theta f_\theta(x))^\intercal \exp(f_\theta(x))dx$ is full-ranked:

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

for absolute constants $c_{\min}$ and $c_{\max}$.

B3 There exist $g_1(x) \geq \sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(x)\|$ and $g_{2,\theta}(x) \geq \max\{\|\nabla_\theta f'_\theta(x)\|, |f'_\theta(x)|\}$ such that g_1 and $g_{2,\theta}$ are $(c_{s1}, c_{s2}, \widetilde{C}_s)$-bowl-shaped, and

$$\int g_1(x)^r \phi(x)dx < \infty, \quad \text{and} \quad \sup_{\theta\in\Theta} \int g_{2,\theta}(x)^{4t} \phi(x)dx < \infty,$$

where t and r are constants satisfying $\frac{8}{r} \leq 4t < 1$.

B4 $\inf_{\theta\in\Theta} f'_\theta(x) \geq (\log \phi)'(x)$ for all $x \leq c_{s1}$, and $\sup_{\theta\in\Theta} f'_\theta(x) \leq (\log \phi)'(x)$ for all $x \geq c_{s2}$.

We then have the following result, proved in Appendix G:

Proposition 5.1. *Let $\{f_\theta\}_{\theta\in\Theta}$ be a class of functions that satisfy assumptions B1–B5 with respect to a transformation function Φ. Let $\theta_{1,n}, \theta_{0,n}$ be two sequences of parameters in Θ. Let $p_n(x) = \exp(f_{\theta_{1,n}}(x))$ and $q_n(x) = \exp(f_{\theta_{0,n}}(x))$. Then, we have:*

(a) If $\|\theta_{1,n} - \theta_{0,n}\|_2 = \Theta(1)$, then the sequence $(p_n, q_n) \in \mathcal{G}_\Phi^{upper}$.

(b) If $\|\theta_{1,n} - \theta_{0,n}\|_2 = o(1)$, then the sequence $(p_n, q_n) \in \mathcal{G}_\Phi^{upper\prime}$.

Proposition 5.1 is useful for generating various examples of densities belonging to $\mathcal{G}_\Phi^{\text{upper}}$ or $\mathcal{G}_\Phi^{\text{upper}\prime}$. For all the examples we show, it suffices to choose the transformation function as:

$$\phi(x) = \frac{e^{1-\sqrt{x+1}}}{4}, \qquad \text{when } S = \mathbb{R}^+, \tag{5.3}$$

$$\phi(x) = \frac{e^{1-\sqrt{|x|+1}}}{8}, \qquad \text{when } S = \mathbb{R}. \tag{5.4}$$

These functions ϕ are similar to a generalized normal density, modified so that $\left|\frac{\phi'(x)}{\phi(x)}\right|$ is bounded. It is easy to verify that $\Phi(x) = \int_0^x \phi(t)dt$ (respectively, $\Phi(x) = \int_{-\infty}^x \phi(t)dt$) is a valid transformation function.

Example 5.1. (Location-scale family over $\mathbb{R}$)
Let $\exp(f(x))$ be a positive base density over $\mathbb{R}$. Define $\theta = (\mu, \sigma)$ and $\Theta = [-C_\mu, C_\mu] \times \left[\frac{1}{c_\sigma}, c_\sigma\right]$ for some absolute constants C_μ and c_σ, and let

$$f_\theta = f\left(\frac{x-\mu}{\sigma}\right) - \log \sigma$$

If $f(x)$ satisfies the conditions

(a) $|f^{(k)}(x)|$ is bounded for some $k \geq 2$, and

(b) there exist absolute constants c and M such that $f'(x) > M$ for $x < -c$ and $f'(x) < -M$ for $x > c$,

then $\{f_\theta\}_{\theta \in \Theta}$ satisfy Assumptions B1–B4 when ϕ is chosen according to equation (5.4). Details are provided in Appendix G.2.

For a sequence of $\{(\mu_{1,n}, \sigma_{1,n}), (\mu_{0,n}, \sigma_{0,n})\}_n \subset (\Theta \times \Theta)^\infty$, let p_n, q_n be defined as:

$$p_n(x) = \frac{1}{\sigma_{1,n}} \exp\left(f\left(\frac{x - \mu_{1,n}}{\sigma_{1,n}}\right)\right) \quad q_n(x) = \frac{1}{\sigma_{0,n}} \exp\left(f\left(\frac{x - \mu_{0,n}}{\sigma_{0,n}}\right)\right)$$

where f satisfy assumptions (a) and (b) from above.

As a direct consequence of proposition 5.1, we have that, if $|\mu_{1,n} - \mu_{0,n}| + |\sigma_{1,n} - \sigma_{0,n}| = \Theta(1)$, then $(p_n, q_n)_n \in \mathcal{G}_\Phi^{\text{upper}}$, that is, $(p_n, q_n)_n$ satisfies assumptions A1'–A4' and if $|\mu_{1,n} - \mu_{0,n}| + |\sigma_{1,n} - \sigma_{0,n}| = o(1)$, then $(p_n, q_n)_n \in \mathcal{G}_\Phi^{\text{upper}'}$, that is, $(p_n, q_n)_n$ satisfies assumption A1–A5.

The above assumptions on $f(x)$ are not strong, and are satisfied for **Gaussian location-scale families**, where the base density is the standard Gaussian density with

$$f(x) = -x^2 - \frac{1}{2} \log 2\pi,$$

and **Laplace location-scale families**, where the base density is the standard Laplace density with

$$f(x) = -|x| - \log 2.$$

We emphasize that we make no assumption on any separation between $\mu_{1,n}, \mu_{0,n}$. Our results apply even when $\mu_{1,n} = \mu_{0,n}$ for all n which implies that $p_n(x)$ and $q_n(x)$ have the same mean.

Example 5.2. (Scale family over $\mathbb{R}^+$)

If the base density $\exp(f(x))$ is supported and positive on $\mathbb{R}^+$, we can define a scale family parametrized by σ as $\exp(f(\frac{x}{\sigma}))$. Let $\theta = (\sigma)$ and $\Theta = \left[\frac{1}{c_\sigma}, c_\sigma\right]$ for some absolute constant c_σ.

And let

$$f_\theta = f\left(\frac{x}{\sigma}\right) - \log \sigma$$

Again, if $f(x)$ satisfy conditions

(a) $|f^{(k)}(x)|$ is bounded for some $k \geq 2$, and

(b) there exist absolute constants $c > 0$ and M such that $f'(x) < -M$ for $x > c$,

then $\{f_\theta\}_{\theta \in \Theta}$ satisfy Assumptions B1–B4 when ϕ is chosen according to equation (5.3).

Again, for a sequence of $\{\sigma_{1,n}, \sigma_{0,n}\}_n \subset (\Theta \times \Theta)^\infty$, let p_n, q_n be defined as:

$$p_n(x) = \frac{1}{\sigma_{1,n}} \exp\left(f\left(\frac{x}{\sigma_{1,n}}\right)\right), \quad \text{and} \quad q_n(x) = \frac{1}{\sigma_{0,n}} \exp\left(f\left(\frac{x}{\sigma_{0,n}}\right)\right).$$

As a direct consequence of proposition 5.1, we have that, if $|\sigma_{1,n} - \sigma_{0,n}| = \Theta(1)$, then $(p_n, q_n)_n \in \mathcal{G}_\Phi^{\text{upper}}$, that is, $(p_n, q_n)_n$ satisfies assumptions A1'–A4' and if $|\sigma_{1,n} - \sigma_{0,n}| = o(1)$, then $(p_n, q_n)_n \in \mathcal{G}_\Phi^{\text{upper}'}$, that is, $(p_n, q_n)_n$ satisfies assumption A1–A5.

An example of a f that satisfies condition (a) and (b) is

$$f(x) = -x$$

which forms the base density of the **exponential distributions**.

We end this section with the Gamma distributions on $\mathbb{R}^+$, which falls outside the above example but satisfy assumption B1-B4 nevertheless.

Example 5.3. (Gamma distribution)

Let $\theta = (\alpha, \beta)$ and $\Theta = [\frac{1}{C}, C]^2$ for some absolute constant C, and let

$$f_\theta(x) = (\alpha - 1)\log x - \beta x + \alpha \log \beta - \log \Gamma(\alpha).$$

where $\Gamma(\cdot)$ is the Gamma function. If we choose ϕ as equation (5.3), then $\{f_\theta\}_{\theta \in \Theta}$ satisfies Assumptions B1–B4. [TODO:proof reference]

For a sequence of $\{(\alpha_{1,n}, \beta_{1,n}), (\alpha_{0,n}, \beta_{0,n})\}_n \subset (\Theta \times \Theta)^\infty$, let p_n, q_n be defined as:

$$p_n(x) = \frac{\beta_{1,n}^{\alpha_{1,n}}}{\Gamma(\alpha_{1,n})} x^{\alpha_{1,n}-1} e^{-\beta_{1,n} x}, \quad \text{and} \quad q_n(x) = \frac{\beta_{0,n}^{\alpha_{0,n}}}{\Gamma(\alpha_{0,n})} x^{\alpha_{0,n}-1} e^{-\beta_{0,n} x},$$

defined over $\mathbb{R}^+$. Then, as a direct consequence of proposition 5.1, we have that, if $|\alpha_{1,n} - \alpha_{0,n}| + |\beta_{1,n} - \beta_{0,n}| = \Theta(1)$, then $(p_n, q_n)_n \in \mathcal{G}_\Phi^{\text{upper}}$, that is, $(p_n, q_n)_n$ satisfies assumptions A1'–A4' and if $|\alpha_{1,n} - \alpha_{0,n}| + |\beta_{1,n} - \beta_{0,n}| = o(1)$, then $(p_n, q_n)_n \in \mathcal{G}_\Phi^{\text{upper}\prime}$, that is, $(p_n, q_n)_n$ satisfies assumption A1–A5.

We note that our results apply even when $\frac{\alpha_{1,n}}{\beta_{1,n}} = \frac{\alpha_{0,n}}{\beta_{0,n}}$ for all n which implies that $p_n(x)$ and $q_n(x)$ have the same mean.

5.2 Lower bound

As with the upper bound analysis, the condition required for the our lower bound depends on whether $H \equiv \int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 dx$ is $o(1)$ or $\Theta(1)$. We capture these conditions by defininng sets of sequences $\mathcal{G}_\Phi^{\text{lower}}$ and $\mathcal{G}_\Phi^{\text{lower}\prime}$.

Define $\mathcal{G}^{\text{lower}} \subset (\mathcal{P} \times \mathcal{P})^\infty$ as the sequences of densities that satisfy $\int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 dx = \Theta(1)$ and

$$\sup_n \int p_n(x) \left| \log \frac{p_n(x)}{q_n(x)} \right|^2 dx < \infty$$
$$\sup_n \int q_n(x) \left| \log \frac{p_n(x)}{q_n(x)} \right|^2 dx < \infty$$

A comparison of these conditions with assumptions A1'-A4' shows that $\mathcal{G}^{\text{lower}} \supseteq \mathcal{G}_\Phi^{\text{upper}}$ for any Φ. To see this, observe that, by assumption A1', $\left| \log \frac{p_n(x)}{q_n(x)} \right|^2$ must be integrable with respect to $\phi(x)$.

When $H_n = o(1)$, we impose much stronger condition on p_n, q_n–we require the likelihood ratio to be bounded. Define $\mathcal{G}^{\text{lower}\prime} \subset (\mathcal{P} \times \mathcal{P})^\infty$ as the sequences of densities that satisfy $\int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 dx = o(1)$ and

$$\sup_n \sup_x \left| \log \frac{p_n(x)}{q_n(x)} \right| < \infty$$

In contrast to the $H_n = \Theta(1)$ case, the bounded likelihood ratio condition that defines $\mathcal{G}^{\text{lower}\prime}$ is generally more restrictive than assumptions A1-A5. However, $\mathcal{G}^{\text{lower}\prime}$ still has significant overlap with $\mathcal{G}_\Phi^{\text{upper}\prime}$ for Φ defined as 5.3 or 5.4.

Theorem 5.3. *Suppose we have K clusters, of which at least one has size $\frac{n}{\beta K}$ and at least one has size $\frac{n}{\beta K}+1$, for some constant $\beta \geq 1$. Let σ_0 denote the true clustering. Suppose $I_n \to 0$ and $P_{0,n}$ and $Q_{0,n}$ satisfy Assumption A0, and let $(p_n(x), q_n(x))$ be a sequence of densities either in $\mathcal{G}^{lower}$ or in $\mathcal{G}^{lower\prime}$. Then any permutation-equivariant algorithm $\widehat{\sigma}$ satisfies the following:*

(i) If $nI_n \to \infty$, then $\mathbb{E}l(\widehat{\sigma}(A), \sigma_0) \geq \exp\left(-(1+o(1))\frac{nI_n}{\beta K}\right)$.

(ii) If $nI_n \to c < \infty$, for some constant c, then $\mathbb{E}l(\widehat{\sigma}(A), \sigma_0) \geq c' > 0$, for some constant c'.

Theorem 5.3 applies to any parameters $(p_n(x), q_n(x), P_0, Q_0, K, \beta, \sigma_0)$ that satisfy the assumptions, rather than being a minimax lower bound involving a supremum over a parameter space.

Remark 5.1. It is interesting to observe that Theorem 5.3, in conjunction with Theorem 5.2, shows that one does not have to pay a price for making nonparametric assumptions: Our nonparametric method achieves the optimal rate of recovery even if the densities $p_n(x)$ and $q_n(x)$ take on a parametric form. This seemingly counterintuitive phenomenon arises because the cost of discretization is reflected in the $o(1)$ term in the exponent and is thus of lower order.

The proof of theorem is given in section H of the appendix and is inspired by the change of measure technique employed by Yun and Proutiere [TODO:cite].

We give a brief sketch of the proof here. Without loss of generality, let cluster 1 and 2 be the clusters in σ_0 that have sizes $\frac{n}{\beta K}$ and $\frac{n}{\beta K}+1$ respectively. We consider another cluster assignment σ_0^2 that differ from σ_0 on only one node – a node in cluster 2 is re-assigned to cluster 1 for σ_0^2.

Since σ_0, σ_0^2 are identical in all other aspects, including the sizes of all the clusters, we can leverage the symmetry of permutation equivariance

[TODO:summarize proof]

6 Proof sketch: Recovery algorithm

A large portion of the Appendix is devoted to proving that our recovery algorithm succeeds and achieves the optimal error rates. We provide an outline of the proofs here.

We divide our argument into propositions that focus on successive stages of our algorithm. A birds-eye view of our method reveals that it consists of two major components: (1) convert a weighted network into a labeled network, and then (2) run a community recovery algorithm on the labeled ne

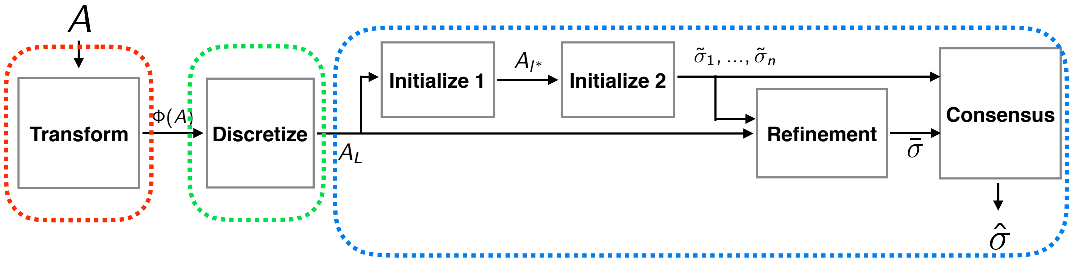

Figure 3: The add noise component also goes into the blue section. Analysis of the right-most blue region is in subsection 6.1, of the middle green region in subsection 6.2, and of the left-most red region in subsection 6.3

6.1 Analysis of community recovery on a labeled network

The workhorse behind our algorithm is a subroutine (right-most blue region in Figure 3) for recovering communities on a network where the edges have discrete labels $l = 1, \dots, L_n$. The following proposition characterizes the rate of convergence of the output of the subroutine, where within-community edges are assigned edge labels with probabilities $\{P_{l,n}\}$, and between-community edges are assigned edge labels according to $\{Q_{l,n}\}$.

Proposition 6.1. *Suppose the edge label probabilities satisfy $\frac{1}{\rho_n} \leq \frac{P_{l,n}}{Q_{l,n}} \leq \rho_n$ for a sequence $\rho_n = \Omega(1)$. Define $I_{L_n} = -2 \log \sum_{l=0}^{L_n} \sqrt{P_{l,n} Q_{l,n}}$ and suppose $I_{L_n} \to 0$. Suppose $L_n = \Omega(1)$ and $\frac{n I_{L_n}}{L_n \rho_n^4} \to \infty$. Let $\widehat{\sigma}$ be the algorithm [TODO:what algorithm]. Then*

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

Yun and Proutiere [20] proposed an algorithm for the labeled SBM that achieves the same rate of convergence. Proposition 6.1 is more general than their result, however, in that, [TODO:we relax their assumption], we allow both the number of labels L_n and the bound ρ_n on the ratio $\frac{P_{l,n}}{Q_{l,n}}$ to diverge to infinity. This extension is critical in analyzing the weighted SBM, since to achieve consistency for continuous distributions, the discretization level L_n must increase with n.

6.2 Discretization of the Renyi divergence

The rate of convergence in Proposition 6.1 resembles the expressions in Theorems 5.2 and 5.1, except the Renyi divergence I is replaced by the discretized Renyi divergence I_L. Thus, we may derive Theorems 5.2 and 5.1 from Proposition 6.1 by showing that $|I_n - I_{L_n}|$ is sufficiently small. It is easy to show that $I_{L_n} \leq I_n$, since discretization always leads to a loss of information. If $p_n(x)$ ad $q_n(x)$ are sufficiently regular in that they may be well-approximated via discretization, one might expect that I_{L_n} to be not much smaller than I_n. Proposition 6.3 and 6.2 formalizes this notion.

For both of the propositions below, we let $\text{bin}_l = [a_l, b_l]$, for $l = 1, \dots, L_n$ be a uniformly-spaced binning of $[0,1]$ and let $P_{l,n} = (1 - P_{0,n}) \int_{a_l}^{b_l} p_n(z) dz$ and $Q_{l,n} = (1 - Q_{0,n}) \int_{a_l}^{b_l} q_n(z) dz$.
The following proposition is useful for proving Theorem 5.1:

Proposition 6.2. *Let $p_n(z), q_n(z)$ be two densities supported on $[0,1]$. Suppose $H_n = \Theta(1)$. Also suppose*

C1 $p_n(z), q_n(z) > 0$ *on* $(0,1)$, *and* $\sup_z \{p_n(z) \vee q_n(z)\} < \infty$.

C2 $\int \left| \log \frac{p_n(z)}{q_n(z)} \right| dz < \infty$.

C3 There exists $h_n(z)$ such that

- *(a)* $h_n(z) \geq \max \left\{ \left| \frac{p_n'(z)}{p_n(z)} \right|, \left| \frac{q_n'(z)}{q_n(z)} \right| \right\}$,
- *(b) $h_n(z)$ is (c_{s1}', c_{s2}', C_s')-bowl-shaped, and*
- *(c)* $\int |h_n(z)|^t dz < \infty$ *for some constant t such that $\frac{2}{r} \leq t \leq 1$.*

C4 $p_n'(z), q_n'(z) \geq 0$ *for all $z < c_{s1}'$, and $p_n'(z), q_n'(z) \leq 0$ for all $z > c_{s2}'$.*

Suppose $L_n \to \infty$ and $P_{0,n}, Q_{0,n}$ satisfy assumption A0. Then

$$\left| \frac{I_n - I_{L_n}}{I_n} \right| = o(1)$$

and $\frac{1}{4c_0} \exp(-L_n^{1/r}) \leq \frac{P_{l,n}}{Q_{l,n}} \leq 4c_0 \exp(L_n^{1/r})$, *for all l, where c_0 is the constant in assumption A0.*

The following proposition is useful for proving Theorem 5.2:

Proposition 6.3. *Let $p_n(z)$ and $q_n(z)$ be two densities supported on $[0,1]$. Suppose $H_n = o(1)$. Let L_n be a sequence such that $L_n \to \infty$. Suppose the following assumptions are satisfied:*

C1' $p_n(z), q_n(z) > 0$ on $(0,1)$, and $\sup_z\{p_n(z) \vee q_n(z)\} < \infty$.

C2' There exists a subinterval $R \subseteq [0,1]$ such that

- *(a) $\frac{1}{\rho} \leq \left|\frac{p_n(z)}{q_n(z)}\right| \leq \rho$ for all $z \in R$, where ρ is an absolute constant, and*
- *(b) $\mu\{R_n^c\} = o(H_n)$, where μ is the Lebesgue measure and $R_n^c \equiv [0,1]\ R_n$.*

C3' Let $\alpha_n^2 = \int_{R_n} \frac{(p_n(z)-q_n(z))^2}{q_n(z)} dz$ and $\gamma_n(z) = \frac{q_n(z)-p_n(z)}{\alpha}$, and suppose $\int_{R_n} q_n(z) \left|\frac{\gamma_n(z)}{q_n(z)}\right|^r dz < \infty$ for an absolute constant $r \geq 4$.

C4' There exists $h_n(z)$ such that

- *(a) $h_n(z) \geq \max\left\{\left|\frac{\gamma_n'(z)}{q_n(z)}\right|, \left|\frac{q_n'(z)}{q_n(z)}\right|\right\}$, and*
- *(b) $h_n(z)$ is (c_{s1}', c_{s2}', C_s')-bowl-shaped for absolute constants c_{s1}', c_{s2}', and C_s', and*
- *(c) $\int_R |h_n(z)|^t dz < \infty$ for an absolute constant $\frac{2}{r} < t < 1$.*

C5' $p_n'(z), q_n'(z) \geq 0$ for all $z < c_{s1}'$, and $p_n'(z), q_n'(z) \leq 0$ for all $z > c_{s2}'$.

Suppose $L_n \to \infty$ and $P_{0,n}, Q_{0,n}$ satisfy assumption A0. Then

$$\left|\frac{I_n - I_{L_n}}{I_n}\right| = o(1)$$

and $\frac{1}{4\rho c_0} \leq \frac{P_{l,n}}{Q_{l,n}} \leq 4\rho c_0$, for all l, where c_0 is the constant in A0.

6.3 Analysis of the transformation function

Propositions 6.3 and 6.2 consider 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 $p_n(x)$ and $q_n(x)$ denote densities over S, and let $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ denote the transformed densities over $[0,1]$. It is easy to see that $p_{\Phi,n}(z) = \frac{p_n(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q_{\Phi,n}(z) = \frac{q_n(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$. Therefore, via the change of variables $z = \Phi^{-1}(x)$, we have the following relations:

$$\int_S \sqrt{p_n(x)q_n(x)}dx = \int_0^1 \sqrt{p_{\Phi,n}(z)q_{\Phi,n}(z)}dz,$$

$$\int_S \left(\sqrt{p_n(x)} - \sqrt{q_n(x)}\right)^2 dx = \int_0^1 \left(\sqrt{p_\Phi(z)} - \sqrt{q_\Phi(z)}\right)^2 dz.$$

Therefore, the divergences I and H between $p_n(x)$ and $q_n(x)$ are equal to the divergences between $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$.

To prove Theorems 5.2 and 5.1, it thus remains to show that if the densities $p(x)$ and $q_n(x)$ satisfy Assumptions A1–A4 (or A1'–A5'), the transformed densities $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ satisfy Assumptions C1–C4 (or C1'–C5') in Proposition 6.3 (or Proposition 6.2). This is done in Propositions F.1 and F.2.

7 Conclusion

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

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

1. An important problem is to extend our analysis to the case of a *heterogenous* stochastic block model, where edge weight distributions depend on the exact community assignments of both endpoints. In such a setting, Abbe and Sandon [2] and Yun and Proutiere [20] 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 [8, 22] 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, and G. Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

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

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

[4] Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

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

[6] 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.

[7] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

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

[9] 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.

[10] Bruce Hajek, Yihong Wu, and Jiaming Xu. Information limits for recovering a hidden community. *IEEE Transactions on Information Theory*, 2017.

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

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

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

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

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

[16] 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.

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

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

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

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

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

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

A Proof of Proposition 6.1

We structure the proof according to the flow of our algorithm. Since this proposition addresses the case of discrete labels, we do not need to consider the "transformation and discretization" step.

We begin by analyzing Algorithm 4.2, where we deliberately add noise to the graph by changing edge colors at random. Although adding random noise destroys information and increases the difficulty of community recovery, Lemma B.2 in Appendix B.5 shows that the process does not significantly affect the Renyi divergence I_L. Furthermore, the new probabilities of edge labels are at least $\frac{c}{n}$, which is important for our later analysis. 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 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:

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

We then select $l^* \in \arg\max_{1 \le l \le L} \widehat{I}_L$. Note that if $|\widehat{P}_l - P_l|$ and $|\widehat{Q}_l - Q_l|$ are small, $\widehat{I}_l$ provides a measure of how "good" a color is for clustering: larger values of $\widehat{I}_l$ correspond to greater separation between P_l and Q_l. Naturally, the accuracy of the estimated edge probabilities $\widehat{P}_l$ and $\widehat{Q}_l$ depends on the accuracy of the spectral clustering step. Proposition B.1 below makes this statement rigorous. Before stating the proposition, we define the set of "good" labels as follows:

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

We bound the difference between the estimated and true probabilities for good and bad colors in Proposition B.1, the formal statement and proof of which are contained in Appendix B.1.

Proposition B.1. *Suppose σ is a clustering with error rate at most γ; i.e., $l(\sigma, \sigma_0) \le \gamma$, for sufficiently small $\gamma \ge \frac{1}{n}$. With probability at least $1 - Ln^{-(3+\delta_p)}$, for a small $\delta_p > 0$, the following hold:*

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

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

We now work toward obtaining a suitable initial clustering with small error rate γ. In Proposition B.2, we show that if the edge probabilities for a particular label are well-separated, the spectral clustering output of Algorithm 4.4 is reasonably accurate. We provide a rough statement of the proposition here, and refer to Appendix B.2 for the precise statement and proof.

Proposition B.2. *If P_l and Q_l satisfy $C_1 \frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2} \le 1$ for an absolute constant C_1, the output σ^l of Algorithm 4.4 satisfies the inequality*

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

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

Thus, if we want to cluster to an arbitrary degree of accuracy, we need $\frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2} \to 0$ for at least one well-separated label l. We show that the label l^* selected in Algorithm 4.3 satisfies $\frac{(P_{l^*} \vee Q_{l^*})}{n(P_{l^*} - Q_{l^*})^2} \to 0$ in the following proposition. A more detailed restatement and proof is contained in Appendix B.3.

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

Now let E_1 denote the high-probability event that the label l^* is chosen according to Proposition B.4. We perform spectral clustering n times, omitting one vertex and clustering on the remaining graph each time, and denote the resulting community assignments by $\{\widetilde{\sigma}_u\}_{1 \le u \le n}$. Note that Proposition B.2, together with a simple union bound, implies that with probability at least $1 - n^{-3}$, all clusterings have error rate bounded by $\gamma := C \frac{P_{l^*} \vee Q_{l^*}}{n(P_{l^*} - Q_{l^*})^2}$, for some constant C. Denote this event by E_2. On $E_1 \cap E_2$, we then have $\gamma \rho_L^4 \to 0$. Thus, we may apply Proposition B.1 on each clustering $\widetilde{\sigma}_u$ to show that the conclusion holds simultaneously for all $\widetilde{\sigma}_u$'s, with probability at least $1 - Ln^{-(2+\delta_p)}$. We denote this last event by E_3. Furthermore, $\eta = \Theta\left(\sqrt{\gamma \log \frac{1}{\gamma}}\right)$, so $|\eta \rho_L| \to 0$.

We now construct the clustering $\widehat{\sigma}_u$ by assigning vertex u to an appropriate community in $\widetilde{\sigma}_u$, using the relation from Algorithm 4.5. In Proposition B.5, we show that with high probability, the assignment $\widehat{\sigma}_u(u)$ is "correct." We provide a rough statement of the proposition here, and defer the exact statement and proof to Appendix B.4:

Proposition B.5. *Let* $\pi_u \in S_K$ *be such that* $l(\sigma_0, \widetilde{\sigma}_u) = d(\sigma_0, \pi_u(\widetilde{\sigma}_u))$. *Conditioned on* $E_1 \cap E_2 \cap E_3$, *with probability at least* $1 - (K-1)\exp\left(-(1 - o(1))\frac{n}{\beta K} I_L\right)$, *we have* $\pi_u^{-1}(\sigma_0(u)) = \widehat{\sigma}_u(u)$.

We briefly discuss the uniqueness of π_u. By construction, the error rate of $\widehat{\sigma}_u$ is at most $\gamma + \frac{1}{n}$, so $l(\sigma_0, \widehat{\sigma}_u) < \frac{1}{8\beta K}$ for sufficiently small γ. Now note that for any u, the minimum cluster size of the clustering $\widehat{\sigma}_u$ is at least $\frac{n}{\beta K} - (n\gamma + 1) \ge \frac{n(1 - \beta K \gamma - \beta K/n)}{\beta K} \ge \frac{n}{2\beta K}$, for small γ. A simple argument (cf. Lemma B.8) shows that π_u is the unique permutation to obtain such a small error rate.

Define π_1 and π_u to be the permutations minimizing $d(\sigma_0, \pi_1(\widehat{\sigma}_1))$ and $d(\sigma_0, \pi_u(\widehat{\sigma}_u))$, respectively, and let ξ_u denote the permutation minimizing $d(\widehat{\sigma}_1, \xi_u(\widehat{\sigma}_u))$. We know that $d(\sigma_0, \pi_1(\widehat{\sigma}_1)) < \frac{1}{8\beta K}$ and $d(\sigma_0, \pi_u(\widehat{\sigma}_u)) < \frac{1}{8\beta K}$. Thus, the triangle inequality implies

$$d(\widehat{\sigma}_1, \pi_1^{-1}(\pi_u(\widehat{\sigma}_u))) = d(\pi_1(\widehat{\sigma}_1), \pi_u(\widehat{\sigma}_u)) \le d(\sigma_0, \pi_1(\widehat{\sigma}_1)) + d(\sigma_0, \pi_u(\widehat{\sigma}_u)) < \frac{1}{4\beta K}.$$

Since the minimum cluster size of both $\widehat{\sigma}_1$ and $\widehat{\sigma}_u$ is $\frac{n}{2\beta K}$, Lemma B.8 implies that $\xi_u = \pi_1^{-1} \circ \pi_u$; and by Lemma B.7, we also have $\widehat{\sigma}(u) = \xi_u(\widehat{\sigma}_u(u))$.

Restating Proposition B.5, we have

$$P\left(\widehat{\sigma}_u(u) \neq \pi_u^{-1}(\sigma_0(u)) \middle| E_1 \cap E_2 \cap E_3\right) \le \exp\left(-(1 + o(1))\frac{nI_L}{\beta K}\right).$$

Furthermore, the left-hand expression is equivalent to $\xi_u^{-1}(\widehat{\sigma}(u)) \neq \pi_u^{-1}(\sigma_0(u))$, or $\widehat{\sigma}(u) \neq \xi_u \circ \pi_u^{-1}(\sigma_0(u)) = \pi_1^{-1}(\sigma_0(u))$.

Altogether, we conclude that

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

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

Finally, suppose

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

Defining $\eta'' = \eta' + \beta\sqrt{\frac{K}{nI_L}} = o(1)$, we have

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

On the other hand, if

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

we have

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

This completes the proof of Proposition 6.1.

B Supporting results for Proposition 6.1

We now provide proofs for the supporting results stated in Appendix A.

B.1 Analysis of estimation error of $\widehat{P_l}$ and $\widehat{Q_l}$

Proposition B.1. *Let A be the adjacency matrix of a labeled network with true clustering assignment σ_0. Suppose σ is a random initial clustering satisfying $l(\sigma,\sigma_0) \leq \gamma$. Let $\widehat{P_l} = \frac{\sum_{u\neq v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(A_{uv}=l)}{|\{u\neq v\colon\,\sigma(u)=\sigma(v)\}|}$ and $\widehat{Q_l} = \frac{\sum_{u\neq v\,:\,\sigma(u)\neq\sigma(v)} \mathbf{1}(A_{uv}=l)}{|\{u\neq v\colon\,\sigma(u)=\sigma(v)\}|}$ be the MLE of P_l and Q_l based on σ. Let δ_p be a positive, fixed, and arbitrarily small real number, and let $c > 0$ be an absolute constant. Then with probability at least $1 - Ln^{-(3+\delta_p)}$, the following hold for all sufficiently small 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*

$$|\widehat{P_l} - P_l| \leq \eta\Delta_l, \quad \textit{and} \quad |\widehat{Q_l} - Q_l| \leq \eta\Delta_l.$$

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*

$$|\widehat{P_l} - P_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}, \quad \textit{and} \quad |\widehat{Q_l} - Q_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}.$$

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

Proof. Our proof proceeds by showing that for any fixed assignment σ with error rate bounded by γ, the event described in Proposition B.1 holds with high probability. For a fixed assignment σ, we call a random graph a "bad graph for σ" if the event does not hold. For each σ, we upper-bound the probability that a randomly chosen graph lies in the set of bad graphs for σ; we then use a union bound over all choices of σ and all L colors to show that the probability of choosing a bad graph is bounded by $Ln^{-(3+\delta_p)}$.

We begin by bounding the bias of $\widehat{P}_l$. We have

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

where $\lambda := \frac{\sum_{u\neq v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(\sigma_0(u)\neq\sigma_0(v))}{|\{u\neq v:\ \sigma(u)=\sigma(v)\}|}$. Thus, $|\mathbb{E}(\widehat{P}_l) - P_l| \le \lambda|Q_l - P_l|$. Furthermore, if $\widehat{n}_k$ denotes the number of vertices in cluster k according to σ, we have

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

Define $\gamma_k = \frac{1}{n}\sum_{u\,:\,\sigma(u)=k} \mathbf{1}(\sigma_0(u)\neq k)$ to be the error rate within the estimated cluster k. Then $\sum_k \gamma_k \le \gamma$ and $\sum_{u\,:\,\sigma(u)=k}\sum_{v\,:\,\sigma(v)=k} \mathbf{1}(\sigma_0(v)\neq k) = \gamma_k n \widehat{n}_k$, implying that

$$\lambda \le \frac{\sum_k 2\gamma_k n\widehat{n}_k}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} = \frac{n}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)}\sum_k 2\gamma_k\widehat{n}_k \overset{(a)}{\le} \frac{K}{n-K}\sum_k 2\gamma_k\widehat{n}_k \overset{(b)}{\le} 4\gamma K,$$

where (a) uses the fact that

$$\sum_k \frac{\widehat{n}_k}{n}(\widehat{n}_k - 1) = n\sum_k \left(\frac{\widehat{n}_k}{n}\right)^2 - 1 \ge \frac{n}{K} - 1, \tag{B.2}$$

and (b) uses the assumption $K < \frac{n}{2}$. Altogether, we conclude that $|\mathbb{E}(\widehat{P}_l) - P_l| \le 4\gamma K\Delta_l$. A similar calculation may be performed for $\widehat{Q}_l$, so

$$\max\left\{|\mathbb{E}(\widehat{P}_l) - P_l|,\ |\mathbb{E}(\widehat{Q}_l) - Q_l|\right\} \le C_2\gamma\Delta_l,$$

for a constant $C_2 > 0$. To simplify presentation, we define $\eta_1 = C_2\gamma$, so

$$\max\left\{|\mathbb{E}(\widehat{P}_l) - P_l|,\ |\mathbb{E}(\widehat{Q}_l) - Q_l|\right\} \le \eta_1\Delta_l. \tag{B.3}$$

We now turn to bounding $|\widehat{P}_l - P_l|$ and $|\widehat{Q}_l - Q_l|$. Denoting $\widetilde{A}_{uv} = \mathbf{1}(A_{ij} = l)$ and using Bernstein's inequality, we have

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

By equation (B.1), we have

$$\sum_{u,v\,\sigma(u)=\sigma(v)} \mathbb{E}\widetilde{A}_{uv} = \sum_k \widehat{n}_k(\widehat{n}_k - 1)\mathbb{E}\widehat{P}_l \le (P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1),$$

implying that

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

Let

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

for a constant C_1 to be defined later. Let

$$A = 2(P_l \vee Q_l)\sum_k \widehat{n}_k(\widehat{n}_k - 1), \quad \text{and} \quad B = C_1\gamma n\log\frac{1}{\gamma} + (3+\delta_p)\log n.$$

We split into two cases:

1. Suppose $A \ge B$. Then $t^2 = 4AB$, and the probability term is at most

$$2\exp\left(-\frac{4AB}{A + \frac{4}{3}\sqrt{AB}}\right) \le 2\exp\left(-\frac{4AB}{A + \frac{4}{3}A}\right) \le 2\exp(-B).$$

2. Suppose $A \le B$. Then $t^2 = 4B^2$, and the probability term is at most

$$2\exp\left(-\frac{4B^2}{A + \frac{4}{3}B}\right) \le 2\exp\left(-\frac{4B^2}{B + \frac{4}{3}B}\right) \le 2\exp(-B).$$

In either case, with probability at least $1 - 2\exp\left(-\left(C_1\gamma n\log\frac{1}{\gamma} + (3+\delta_p)\log n\right)\right)$, we have

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

We now derive a more manageable upper bound for t. Using the notation from above, we have $t^2 = \max(4AB, 4B^2) \le 4(\sqrt{AB}+B)^2$. Since $\gamma \ge \frac{1}{n}$, we have $C_1\gamma n\log\frac{1}{\gamma} + (3+\delta_p)\log n \le \widetilde{C}_1\gamma n\log\frac{1}{\gamma}$, implying that

$$\begin{aligned}\frac{t}{\sum_{u\neq v}\mathbf{1}(\sigma(u)=\sigma(v))} &\le 2\frac{\sqrt{2(P_l \vee Q_l)\widetilde{C}_1\gamma n\log\frac{1}{\gamma}}}{\sqrt{\sum_k \widehat{n}_k(\widehat{n}_k - 1)}} + 2\frac{\widetilde{C}_1\gamma n\log\frac{1}{\gamma}}{\sum_k \widehat{n}_k(\widehat{n}_k - 1)} \\ &\overset{(a)}{\le} 2\frac{\sqrt{2(P_l \vee Q_l)}\sqrt{\widetilde{C}_1\gamma K\log\frac{1}{\gamma}}}{\sqrt{n-K}} + 2\frac{\widetilde{C}_1\gamma K\log\frac{1}{\gamma}}{n-K} \\ &\overset{(b)}{\le} 4\sqrt{\frac{P_l \vee Q_l}{n}}\sqrt{\widetilde{C}_1 K\gamma\log\frac{1}{\gamma}} + 4\frac{\widetilde{C}_1 K\gamma\log\frac{1}{\gamma}}{n},\end{aligned}$$

where (a) uses inequality (B.2) and (b) uses the assumption $n - K \geq \frac{n}{2}$.

To further simplify the expression, note that $P_l \vee Q_l \geq \frac{c}{n}$ implies $\frac{1}{n} \leq \frac{1}{\sqrt{c}}\sqrt{\frac{P_l \vee Q_l}{n}}$, so with probability at least $1 - \exp(-C_1 \gamma n \log \frac{1}{\gamma} - (3 + \delta_p) \log n)$, we have

$$|\widehat{P}_l - \mathbb{E}(\widehat{P}_l)| \leq \sqrt{\frac{P_l \vee Q_l}{n}} \left(C_1' \sqrt{\gamma \log \frac{1}{\gamma}} + C_2' \gamma \log \frac{1}{\gamma} \right), \tag{B.4}$$

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

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

When γ is sufficiently small, the first term dominates, so we may take choose the right-hand sides to be $\eta_2 \sqrt{\frac{P_l \vee Q_l}{n}}$, where $\eta_2 = C_3 \sqrt{\gamma \log \frac{1}{\gamma}}$.

Finally, note that there are at most $\binom{n}{\gamma n} K^{\gamma n}$ possible σ's satisfying the error bound. We have

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

for a suitable constant C_1. Taking a union bound across all cluster assignments, we then conclude that the probability of inequality (B.4) holding simultaneously for all labels l is at least $1 - Ln^{-(3+\delta_p)}$.

Combining inequalities (B.3) and (B.4), we arrive at the bound

$$\max \left\{ |P_l - \widehat{P}_l|,\ |Q_l - \widehat{Q}_l| \right\} \leq \eta_1 \Delta_l + \eta_2 \sqrt{\frac{P_l \vee Q_l}{n}}. \tag{B.5}$$

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

$$\max \left\{ |P_l - \widehat{P}_l|,\ |Q_l - \widehat{Q}_l| \right\} \leq \eta_1 \Delta_l + \eta_2 \Delta_l = (\eta_1 + \eta_2) \Delta_l,$$

whereas if $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$, we have

$$\max \left\{ |P_l - \widehat{P}_l|,\ Q_l - \widehat{Q}_l| \right\} \leq \eta_1 \sqrt{\frac{P_l \vee Q_l}{n}} + \eta_2 \sqrt{\frac{P_l \vee Q_l}{n}} = (\eta_1 + \eta_2) \sqrt{\frac{P_l \vee Q_l}{n}}.$$

Since $\eta_2 = C_3 \sqrt{\gamma \log \frac{1}{\gamma}}$ dominates $\eta_1 = C_2 \gamma$ for sufficiently small γ, the desired bounds follow. □

B.2 Analysis of spectral clustering

Proposition B.2. *Suppose an unweighted adjacency matrix A is drawn from a homogeneous stochastic block model with probabilities p and q and cluster imbalance factor β. Suppose $p, q \geq \frac{c}{n}$. Then there exists a constant C such that if*

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

the output σ of Algorithm 4.4 with parameters $\mu \geq 32C^2\beta$ and $\tau = \overline{d}$ satisfies

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

with probability at least $1 - n^{-C'}$, where $C' > 4$.

Proof. Note that the trim parameter τ is a random variable, since the average degree $\overline{d}$ is random. Since the community sizes are bounded by $\frac{n}{\beta K}$, we may find constants $C_{d_1} < C_{d_2} = 1$, depending 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 Hoeffding's inequality, we conclude that with probability at least $1 - \exp(-nC_{\overline{d}})$, for some constant $C_{\overline{d}}$, we have

$$\frac{C_{d_1}}{2} n(p \vee q) \leq \overline{d} \leq 2C_{d_2} n(p \vee q). \tag{B.6}$$

We now apply the following lemma:

Lemma B.1. *(Lemma 5 of Gao et al. [7]) Let $P \in [0,1]^{n \times n}$ be a symmetric matrix, and let $p_{max} := \max_{u \geq v} P_{uv}$. Let A be an adjacency matrix such that $A_{uu} = 0$ and $A_{uv} \sim Ber(P_{uv})$ for $u < v$. For any $C' > 0$ and $0 < C_1 < C_2$, there exists some $C > 0$ such that*

$$\|T_\tau(A) - P\|_2 \leq C\sqrt{np_{max} + 1}, \qquad \forall \tau \in [C_1(np_{max} + 1), C_2(np_{max} + 1)],$$

with probability at least $1 - n^{-C'}$.

Remark B.1. Lemma 5 from Gao et al. [7] is stated slightly differently—for any $C' > 0$, there exist constants c, C_1, and C_2 such that the result holds with probability at least $1 - n^{-C'}$. However, our restatement follows immediately from slight modifications of the proof. Furthermore, note that the statement of Lemma B.1 refers to the output of spectral clustering with respect to a fixed trim parameter, but we will apply it in a setting where τ is random.

Using Lemma B.1 with a fixed $C' > 4$ and constants $C_1 = \frac{C_{d_1}}{2}$ and $C_2 = 2C_{d_2}$, we conclude that there exists a constant $C > 0$ such that

$$\|T_\tau(A) - P\|_2 \leq C\sqrt{n(p \vee q)}, \qquad \forall \tau \in [C_1, C_2], \tag{B.7}$$

with probability at least $1 - n^{-C'}$. In particular, this inequality will also hold for the random choice $\tau = \overline{d}$, by inequality (B.6) above. Furthermore, we may assume that $C \geq 1$, since inequality (B.7) also holds with C replaced by $\max(1, C)$. Thus,

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

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. This implies that

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

We now denote the K distinct rows of P by $\{\mathcal{Z}_i\}_{1 \leq i \leq K}$, and for a vertex u, denote the row P_u by $\mathcal{Z}(u)$. Note that

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

since each cluster contains at least $\frac{n}{\beta K}$ vertices.

A vertex u is considered *valid* if $\|\widehat{A}_u - \mathcal{Z}_i\|_2^2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$ for some $\mathcal{Z}_i$; otherwise, u is *invalid.* Also 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 P closest to $\widehat{A}_u$. Note 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 show that the set S constructed in Algorithm 4.4 satisfies the following properties:

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

We first prove that the proposition follows from the claims. We denote the rows of $\widehat{A}$ corresponding to the members of S by $\mathcal{S}_i$, assigning indices so that $\mathcal{S}_i$ is the surrogate for $\mathcal{Z}_i$ (i.e., both $\mathcal{S}_i$ and $\mathcal{Z}_i$ are associated to a common vertex u). In particular, note that

$$\|\mathcal{S}_i - \mathcal{Z}_i\|_2 \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n, \qquad \forall 1 \leq i \leq K,$$

since S only contains valid points. Let $\mathcal{S}(u)$ be the surrogate of $\mathcal{Z}(u)$, and denote $\mathcal{S}^*(u) = \arg\min_{\mathcal{S}_i} \|\widehat{A}_u - \mathcal{S}_i\|^2$; i.e., the member of S that is closest to $\widehat{A}_u$. We say that a valid point u is *misclassified* if $\mathcal{S}^*(u) \neq \mathcal{S}(u)$. The number of mistakes we make is thus bounded by the number of invalid points plus the number of misclassified valid points. Note that if u is invalid, we have $\|\widehat{A}_u - P_u\|_2^2 \geq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. We claim that the same inequality holds for any misclassified valid point u.

Consider such a point u. Since u is valid, there exists $\mathcal{Z}_i$ such that

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

We claim that $\mathcal{S}^*(u) = \mathcal{S}_i$. For any $j \neq i$, we have

$$\begin{aligned}\|\widehat{A}_u - \mathcal{S}_j\|_2 &\geq \|\mathcal{Z}_i - \mathcal{Z}_j\|_2 - \|\mathcal{Z}_j - \mathcal{S}_j\|_2 - \|\mathcal{Z}_i - \widehat{A}_u\|_2 \\ &\geq \sqrt{\frac{2}{\beta K}(p-q)^2 n} - 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} \\ &> 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n}.\end{aligned}$$

Furthermore,

$$\|\widehat{A}_u - \mathcal{S}_i\|_2 \leq \|\widehat{A}_u - \mathcal{Z}_i\|_2 + \|\mathcal{S}_i - \mathcal{Z}_i\|_2 \leq 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n}.$$

Thus, for any $j \neq i$, we have

$$\|\widehat{A}_u - \mathcal{S}_j\|_2 > \|\widehat{A}_u - \mathcal{S}_i\|_2,$$

implying that $\mathcal{S}^*(u) = \mathcal{S}_i$.

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

$$\|\widehat{A}_u - \mathcal{Z}(u)\|_2 = \|\widehat{A}_u - \mathcal{Z}_j\|_2$$

$$\begin{aligned}
&\geq \|\widehat{A}_u - \mathcal{S}_j\|_2 - \|\mathcal{S}_j - \mathcal{Z}_j\|_2 \\
&\geq 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} - \sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} \\
&= \sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n},
\end{aligned}$$

which is the bound we wanted to prove.

Finally, we conclude that the number of mistakes incurred by algorithm 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 the notation $N(u) = \{v : \|\widehat{A}_u - \widehat{A}_v\|_2^2 \leq \mu K^2 \frac{\overline{d}}{n}\}$. Furthermore, by a Chernoff bound, we have $\overline{d} \leq 2(p\vee q)n$ with probability at least $1 - \exp(-C_{\overline{d}} n)$. 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.

By the 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 the triangle inequality, we then have

$$\begin{aligned}
\|\widehat{A}_v - \mathcal{Z}(v)\|_2 &\geq \|\widehat{A}_u - \mathcal{Z}(v)\|_2 - \|\widehat{A}_u - \widehat{A}_v\|_2 \\
&\geq \sqrt{\frac{1}{16\beta K}(p-q)^2 n} - \sqrt{2\mu K^2 (p\vee q)} \\
&\overset{(a)}{\geq} \sqrt{\frac{1}{16\beta K}(p-q)^2 n} - \sqrt{\frac{1}{64\beta K}(p-q)^2 n} = \sqrt{\frac{1}{64\beta K}(p-q)^2 n},
\end{aligned}$$

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

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

By assumption, this quantity is less than $\frac{1}{\mu}\frac{n}{K}$.

Proof of Claim 2: We first claim that in every cluster, at least half the points u satisfy $\|\widehat{A}_u - P_u\|_2^2 \leq \frac{1}{4}\mu K^2 (p\vee q)$. This is because the total error is bounded by $\sum_{u=1}^n \|\widehat{A}_u - P_u\|_2^2 \leq 4KC^2 n(p\vee q)$, so the total number of points that violate the condition is at most $\frac{4KC^2 n(p\vee q)}{\frac{1}{4}\mu K^2 (p\vee q)} \leq \frac{n}{2\beta K}$, using the assumption that $\mu \geq 32C^2\beta$.

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

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

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

On the other hand, because S does not yet have cardinality K, some $\mathcal{Z}_j$ must exist that 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 $N(u) \geq \frac{1}{\mu}\frac{n}{K}$ and

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

where the second inequality follows by assumption. Since $\mathcal{Z}_j \neq \mathcal{Z}(s_i)$ for any $1 \leq i \leq r$, we have $\|\mathcal{Z}_j - \mathcal{Z}(s_i)\|_2^2 \geq 2\frac{1}{\beta K}(p-q)^2 n$, for all $1 \leq i \leq r$. By Claim 1, all s_i's are valid, so

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

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

B.3 Choosing the label l^*

First, we show that for sufficiently well-separated labels, $\widehat{I}_l$ is close to $\frac{(P_l - Q_l)^2}{P_l \vee Q_l}$. If the probabilities are not well-separated, we claim that $\widehat{I}_l$ is negligibly small.

Proposition B.3. *Suppose $\frac{1}{\rho_L} \leq \frac{P_l}{Q_l} \leq \rho_L$ for all l. Let σ^l be the output of spectral clustering based on $\widetilde{A}_{ij} = \mathbf{1}(A_{ij} = l)$, and let $\widehat{P}_l$ and $\widehat{Q}_l$ be estimates of P_l and Q_l constructed from σ^l. There exist positive constants C_{test}, C_1, C_2, C, and δ_p such that, with probability at least $1 - Ln^{-3+\delta_p}$, we have the following:*

1. *For all labels l satisfying $P_l \vee Q_l > \frac{c}{n}$ and $\Delta_l \geq \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$,*

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

2. *For all labels satisfying $P_l \vee Q_l > \frac{c}{n}$ and $\Delta_l < \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$,*

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

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 $\Delta_l^2 \geq \frac{P_l \vee Q_l}{n}$, the estimated probabilities $\widehat{P}_l$ and $\widehat{Q}_l$ satisfy

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

with probability at least $1 - n^{-(3+\delta_p)}$. We first pick a value of γ such that $\eta < \frac{1}{4}$. We now ensure that the error rate obtained from Proposition B.2 matches our choice of γ. Recall that Proposition B.2 states that under the assumptions $P_l \vee Q_l > \frac{c}{n}$ and

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

we have

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

for appropriate constants C_1 and C_2. In particular, for $C_{test} \geq 1$ sufficiently large,

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

for all labels l such that $\Delta_l \geq \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$. Next, note that

$$|\widehat{P}_l - \widehat{Q}_l| \leq |\widehat{P}_l - P_l| + |P_l - Q_l| + |\widehat{Q}_l - Q_l| \leq 2\eta\Delta_l + \Delta_l \leq \frac{3}{2}\Delta,$$

and

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

Furthermore,

$$\widehat{P}_l \vee \widehat{Q}_l \leq (P_l \vee Q_l) + \eta\Delta_l \leq (P_l \vee Q_l) + \eta(P_l \vee Q_l) \leq \frac{5}{4}(P_l \vee Q_l),$$

and

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

We conclude that

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

Now suppose $\Delta_l^2 < C_{test}\frac{P_l \vee Q_l}{n}$. 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, so with probability at least $1 - Ln^{-(3+\delta_p)}$,

$$|\widehat{P}_l - P_l| \leq \eta\left(\Delta_l \vee \sqrt{\frac{P_l \vee Q_l}{n}}\right) \leq \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}},$$

using the choice $\eta \leq \frac{1}{4}$. An analogous bound holds for $|\widehat{Q}_l - Q_l|$. Hence,

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

Now note that

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

for some constant C'. It follows that

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

□

We apply Proposition B.3 to conclude that Algorithm 4.3 succeeds in choosing a color l^* for which $\frac{n(P_{l^*}-Q_{l^*})^2}{P_{l^*}\vee Q_{l^*}}$ is arbitrarily large:

Proposition B.4. *Suppose $a_n := \frac{nI_L}{L\rho_L^4} \to \infty$. For sufficiently large n, with probability at least $1-2Ln^{-(3+\delta_p)}$, we have $\frac{n(P_{l^*}-Q_{l^*})^2}{(P_{l^*}\vee Q_{l^*})\rho_L^4} \geq Ca_n$, for some constant C.*

Proof. Let C_{test} be the constant in Proposition B.3. By Lemma B.6, we know that I_L is of the same order as $\sum_{l\in L_1}\frac{\Delta_l^2}{P_l\vee Q_l}$, implying the existence of a label l such that $\Delta_l \geq C_{test}\sqrt{\frac{P_l\vee Q_l}{n}}$ and $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq C\frac{nI_L}{L} = Ca_n\rho_L^4$, for a constant C. Suppose the event of Proposition B.3 holds, which happens with probability at least $1-Ln^{-(3+\delta_p)}$.

Step 1. We claim that l^* satisfies $\Delta_{l^*} \geq C_{test}\sqrt{\frac{P_l\vee Q_l}{n}}$. Let l be a label such that $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq Ca_n\rho_L^4$, and suppose the claim is false. By Proposition B.3 and the maximality of l^*, we have

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

Proposition B.3 also implies that

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

However, this is a contradiction, since $a_n \to \infty$ and $\rho_L \geq 1$.

Step 2: Again, let l be a label such that $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq Ca_n\rho_L^4$. By Proposition B.3, we then have

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

implying the desired result. □

B.4 Analysis of error probability for a single node

Proposition B.5. *Let u be an arbitrary fixed node, and let $\widetilde{\sigma}_u$ be the output of Algorithm 4.3. Suppose $\pi_u \in S_K$ satisfies*

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

where both l and d are taken with respect to the set $\{1,2,\ldots,n\}\setminus\{u\}$. Conditioned on the events that the error rate γ of $\widetilde{\sigma}_u$ satisfies $\gamma\rho_L^4 \to 0$, and also the event that the result of Proposition B.1 holds for a sequence η satisfying $\eta\rho_L^2 \to 0$, we have

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

with probability at least $1-(K-1)\exp\left(-(1-o(1))\frac{n}{\beta K}I_L\right)$.

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

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

or equivalently,

$$\sum_{v\,:\,\widetilde{\sigma}_u(v)=k}\overline{A}_{uv} - \sum_{v\,:\,\widetilde{\sigma}_u(v)=1}\overline{A}_{uv} \geq 0, \tag{B.10}$$

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 $m_k = |\{v : \widetilde{\sigma}_u(v)=k\}|$, and let $m_1' = \{v : \widetilde{\sigma}_u(v) = 1, \sigma_0(v)=1\}$ be the points correctly clustered by σ_u. Let $m_k' = \{v : \widetilde{\sigma}_u(v) \neq 1, \sigma_0(v)=k\}$ denote the points correctly classified by $\widetilde{\sigma}_u$ in community k. With these definitions, the probability of the bad event in equation (B.10) 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 and write $\widetilde{Y}_i$ and $\widetilde{X}_i$ in both bracketed terms. These random variables are not the same, but are independent and identical copies.) This is equal to 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 further bound this probability 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

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

$$= \left(\frac{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} P_l}{\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l} \right)^{m_k - m_k'} \left(\frac{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} Q_l}{\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l} \right)^{m_1 - m_1'} \tag{B.11}$$

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

We bound terms (B.11) and (B.12) separately. Loosely speaking, we will show that term (B.11) is bounded in magnitude by $\exp(o(I_L)\frac{n}{K})$, and term (B.12) is bounded by $\exp(-\frac{n}{\beta K}(1+o(1))I_L)$.

Bound for term (B.11). We derive a number of separate lemmas bounding various intermediate terms in the computation. In particular, we use the bounds from Lemmas B.5, B.4, and B.6 in the following sequence of inequalities:

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

In (a), we have used Lemma B.5. In (b), we have used the fact that $\sum_l \sqrt{P_l Q_l} \to 1$, so this sum exceeds $\frac{1}{2}$ when n is sufficiently large. In (c), we have employed Lemma B.4, which appropriately bounds the term $\left(\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} - 1 \right)$. Here, $\eta' = o(1)$. Inequality (d) follows from Lemma B.6. Finally, inequality (e) follows from the assumption that $\frac{I_L n}{L \rho_L^4} \to \infty$ (note that $\rho_L \ge 1$) and by appropriately redefining η'). Identical analysis shows that

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

Finally, note that $|x| \leq \exp(|1-x|)$, so term (B.11) may be bounded as

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

$$\leq \exp(CI_L\rho_L\gamma n).$$

Since $\gamma\rho_L = o(1)$, we conclude that term (B.11) is bounded by $\exp\left(\frac{n}{K}o(I_L)\right)$, as desired.

Bound for term (B.12). Let $\widehat{I} = -\log\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)$. With this definition, we have

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

We claim that the following statements are true:

1. $m_1, m_k \geq \frac{n}{\beta K}(1-\beta K\gamma)$.
2. $\widehat{I} \geq I_L(1+o(1))$.
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 assume these statements are true, and bound term (B.12). We have

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

where the last inequality holds because $\gamma = o(1)$. It remains to prove the three claims.

Claim 1: This is straightforward. The labeling $\widetilde{\sigma}_u$ has at most γn errors, so $m_1 \geq m_1' \geq \frac{n}{\beta K} - \gamma n$. A similar argument 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. We begin by writing

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

Let us first consider the numerator:

$$\left(\sum_l \sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} P_l\right)$$

$$= \left(\sum_l \sqrt{P_l Q_l} \sqrt{\frac{\widehat{P}_l}{P_l} \frac{Q_l}{\widehat{Q}_l}} \right) \left(\sum_l \sqrt{P_l Q_l} \sqrt{\frac{P_l}{\widehat{P}_l} \frac{\widehat{Q}_l}{Q_l}} \right)$$

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

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

where we define

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

Furthermore,

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

$$\geq -\log \left(1 + 4 \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 \right) \right) \quad \left(\text{assuming } \sum_l \sqrt{P_l Q_l} \geq 1/2\right)$$

$$\geq -4 \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} \left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 \right). \tag{B.13}$$

We now bound $|T_{l,l'} - 1|$:

$$|T_{l,l'} - 1| = \left| \frac{\widehat{P}_l}{P_l} \frac{Q_l}{\widehat{Q}_l} \frac{P_{l'}}{\widehat{P}_{l'}} \frac{\widehat{Q}_{l'}}{Q_{l'}} - 1 \right|$$

$$= \left| \left(1 - \frac{P_l - \widehat{P}_l}{P_l} \right) \left(1 - \frac{\widehat{Q}_l - Q_l}{\widehat{Q}_l} \right) \left(1 - \frac{\widehat{P}_{l'} - P_{l'}}{\widehat{P}_{l'}} \right) \left(1 - \frac{Q_{l'} - \widehat{Q}_{l'}}{Q_{l'}} \right) - 1 \right|$$

$$\overset{(a)}{\leq} 2 \left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{\widehat{Q}_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{\widehat{P}_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}} \right)$$

$$\overset{(b)}{\leq} 4 \left(\frac{|P_l - \widehat{P}_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l} + \frac{|\widehat{P}_{l'} - P_{l'}|}{P_{l'}} + \frac{|Q_{l'} - \widehat{Q}_{l'}|}{Q_{l'}} \right),$$

where (a) and (b) follow from Lemma B.3. Since we only work with pairs (l, l') such that $l' > l$, we may choose any ordering we like. Thus, suppose the l's are ordered in decreasing order of $\frac{|\widehat{P}_l - P_l|}{P_l} + \frac{|\widehat{Q}_l - Q_l|}{Q_l}$. For all pairs $l < l'$, we then have

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

By Proposition B.1, we have

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

for any $l \in L_1$. For $l \notin L_1$, we have

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

and similarly for the $\frac{|\widehat{Q}_l - Q_l|}{Q_l}$ term. Plugging these bounds into the previous derivation, we obtain

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

We now use the Taylor approximation of $\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2$ around $T_{l,l'} = 1$:

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

Continuing the bound (B.13), we then obtain

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

where (a) follows from the fact that $\sum_{l'} \sqrt{P_{l'} Q_{l'}} \leq 1$, the statement $\sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} = \Theta(I_L)$ from Lemma B.6, and our assumption that $\sum_{l \notin L_1} \frac{1}{n} \leq \frac{L}{n} = o(I_L)$. This proves the claim.

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

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

Assume $m_k \geq m_1$. The reverse case may be analyzed in an identical manner. We may 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 $\sum_l \sqrt{P_l Q_l} \to 1$, so Lemma B.5 implies that the denominators are $\Theta(1)$. We bound the numerator as follows:

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

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 $\eta\rho_L \to 0$, and step (d) follows from our assumption $\frac{L}{n} = o(I_L)$. Thus, we obtain

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

proving the claim.

Combining bounds for terms (B.11) and (B.12): Multiplying the bounds for terms (B.11) and (B.12) 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.5 Additional lemmas for Proposition 6.1

Lemma B.2. *Let L, P_l, Q_l, ρ_L, and I_L satisfy the assumptions in Proposition 6.1. Define the new probabilities of edge labels as follows:*

$$P_l' := P_l(1 - \delta) + \frac{\delta}{L + 1}, \quad \text{and} \quad Q_l' := Q_l(1 - \delta) + \frac{\delta}{L + 1},$$

for all $0 \le l \le L$, where $\delta = \frac{c(L+1)}{n}$. Let I'_L denote the Renyi divergence between P'_l and Q'_l. Then for all sufficiently large n, we have $P'_l, Q'_l > \frac{c}{n}$ for all $0 \le l \le L$, and

$$I'_L = I_L(1 + o(1)).$$

Proof. Clearly, $P'_l, Q'_l > \frac{c}{n}$. For the second part of the lemma, we begin by writing

$$I_L = -2\log \sum_{l=0}^{L} \sqrt{P_l Q_l} = -2\log\left(1 - \frac{1}{2}\sum_{l=0}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2\right) = \left(\sum_{l=0}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2\right)(1 + o(1)).$$

Similarly, we have $I'_L = \left(\sum_{l=0}^{L}(\sqrt{P'_l} - \sqrt{Q'_l})^2\right)(1 + o(1))$, so it is enough to show that

$$\left(\sum_{l=0}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2\right) = \left(\sum_{l=0}^{L}(\sqrt{P'_l} - \sqrt{Q'_l})^2\right)(1 + o(1)).$$

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

Let $\Delta_l := |P_l - Q_l|$ and $\Delta'_l := |P'_l - Q'_l|$. For labels in G_1, we have $\Delta_l \le \frac{a}{n}$. Thus,

$$(\sqrt{P_l} - \sqrt{Q_l})^2 = \frac{\Delta_l^2}{(\sqrt{P_l} + \sqrt{Q_l})^2} \le \Delta_l \le \frac{a}{n},$$

and

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

Therefore,

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

For labels in G_2, we may write

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

We analyze the term inside the absolute value as follows:

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

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

$$\frac{1}{n(P_l \vee Q_l)} < \frac{1}{a} = o(1), \quad \text{so} \quad \frac{1}{nP_l} \le \frac{\rho_L}{n(P_l \vee Q_l)} < \frac{\rho_L}{a} = o(1).$$

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

$$\left(\frac{\sqrt{P_l}+\sqrt{Q_l}}{\sqrt{P_l}\sqrt{1-\delta+\frac{c}{nP_l}}+\sqrt{Q_l}\sqrt{1-\delta+\frac{c}{nQ_l}}}\right)^2 = \left(\frac{\sqrt{P_l}+\sqrt{Q_l}}{\sqrt{P_l}(1+o(1))+\sqrt{Q_l}(1+o(1))}\right)^2 = 1+o(1).$$

Hence, we may conclude that

$$\sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2}\left|1-(1-\delta)^2\frac{(\sqrt{P_l}+\sqrt{Q_l})^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2}\right| = \sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2}\cdot o(1) = o(I_L).$$

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

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

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

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

In particular, for sufficiently small η, we have $\frac{1}{2}P_l \le \widehat{P}_l \le 2P_l$, and likewise for Q_l.

Proof. We prove the statement first for P_l; the same argument applies to Q_l. By Proposition B.1, we have that either $|\widehat{P}_l - P_l| \le \eta\Delta_l$ or $|\widehat{P}_l - P_l| \le \eta\sqrt{\frac{P_l\vee Q_l}{n}}$. First suppose $|\widehat{P}_l - P_l| \le \eta\Delta_l$. Then

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

If instead $|\widehat{P}_l - P_l| \le \eta\sqrt{\frac{P_l\vee Q_l}{n}}$, we have

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

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

Lemma B.4. *Suppose $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ and $P_l, Q_l \ge \frac{c}{n}$ for all l, where $\rho_L > 1$. Conditioned on the event that the conclusion of Proposition B.1 holds for a sequence η such that $\eta\rho_L^2 \to 0$:*

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 η' 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*

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

By symmetry, the same bounds also hold for $\sqrt{\frac{\widehat{Q}_l}{\widehat{P}_l}} - 1$.

Proof. First suppose l satisfies $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq 1$. By Lemma B.3, we have $\frac{\widehat{Q}_l - Q_l}{Q_l} = \eta'$, where $\eta' \to 0$. In the following derivation, we use η' to denote a sequence such that $\eta' = o(1)$; the actual value of η' may change from instance to instance. We use η to denote a sequence where $\eta\rho_L = o(1)$. We have

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

In (a), we have used the fact that $|\widehat{P}_l - P_l| \leq \eta\Delta_l$ and $|\widehat{Q}_l - Q_l| \leq \eta\Delta_l$ by Proposition B.1. Applying the inequality $|\sqrt{x} - 1| \leq |x - 1|$ then completes the proof of the first case.

The proof of the second case is almost identical. Suppose l satisfies $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$. Then

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

By Lemma B.3, we have $\frac{\widehat{Q}_l - Q}{Q_l} = \eta'$, where $\eta' = o(1)$. Hence,

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

Using the inequality $|\sqrt{1+x} - 1| \leq x$ for $x \geq 0$, 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 have assumed that $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$, implying

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

□

The following lemma provides a bound for $\sqrt{\frac{\widehat{P}_l}{\widehat{Q}_l}} Q_l$:

Lemma B.5. *Suppose*

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

where $\eta' = o(1)$. For all sufficiently small η, we have

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

Proof. We have the sequence of equalities

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

where the penultimate equality uses the fact that $\frac{\widehat{Q}_l - Q_l}{Q_l} = \eta' \to 0$. Taking η sufficiently small yields the upper bound.

For sufficiently small η, we also have $\widehat{P}_l \geq \frac{1}{2} P_l$ and $\widehat{Q}_l \geq \frac{1}{2} Q_l$, yielding the lower bound. □

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

$$C_1 \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} \leq I_L \leq C_2 \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l}, \tag{B.14}$$

for some constants C_1 and C_2.

Proof. Throughout the proof, let η' denote a sequence converging to 0, and let C denote a $\Theta(1)$ sequence that may 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 $n \to \infty$, so $\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 \to 0$, we have the following bound for sufficiently large n:

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

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

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

Now note that

$$\sum_{l \in L_1^c} \frac{\Delta_l^2}{P_l \vee Q_l} \le \frac{LC_{test}^2}{n},$$

and by our assumptions, $\frac{L}{n} = o(I_L)$. Hence, the sum over labels in L_1 must be $\Theta(I_L)$, which is equivalent to inequality (B.14). □

We state Lemma 4 from Gao et al. [7], which analyzes the consensus step of the algorithm:

Lemma B.7. *Let σ and σ' be two clusters such that, for some constant $C \ge 1$, the minimum cluster size is at least $\frac{n}{Ck}$. Define the map $\xi : [k] \to [k]$ according to*

$$\xi(k) = \arg\max_{k'} |\{v : \sigma(v) = k\} \cap \{v : \sigma'(v) = k'\}|.$$

If $\min_{\pi \in S_k} l(\pi(\sigma), \sigma') < \frac{1}{Ck}$, we have $\xi \in S_k$ 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 $d(\pi(\sigma), \sigma') < \frac{T}{2n}$ and $d(\xi(\sigma), \sigma') < \frac{T}{2n}$. Then $\pi = \xi$.*

Proof. Suppose the contrary, and choose any k such that $\pi(k) \neq \xi(k)$. We then have

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

implying that $|\{\sigma(u) = k\} \cap \{\sigma'(u) = \pi(k)\}| > \frac{T}{2}$. But then

$$n \cdot d(\xi(\sigma), \sigma') \ge |\{\sigma(u) = k\} \cap \{\sigma'(u) \neq \xi(k)\}| \ge |\{\sigma(u) = k\} \cap \{\sigma'(u) = \pi(k)\}| \ge \frac{T}{2},$$

a contradiction. □

C Proof of Proposition 6.3

Proof. We use the notation

$$\widetilde{P}_l := (1 - P_0) \int_{a_l}^{b_l} p(x)dx := (1 - P_0)P_l,$$

$$\widetilde{Q}_l := (1 - Q_0) \int_{a_l}^{b_l} q(x)dx := (1 - Q_0)Q_l.$$

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 claimed bounds. Consider an l such that $\text{bin}_l \cap R^c = \emptyset$. For all $x \in \text{bin}_l$, we have $\frac{1}{\rho} \le \frac{p(x)}{q(x)} \le \rho$ by Assumption C2'. It follows that $\frac{P_l}{Q_l} = \frac{\int_{\text{bin}l} p(x)dx}{\int_{\text{bin}_l} q(x)dx} \le \rho$. The lower bound is derived in the same manner. Since $\frac{1-P_0}{1-Q_0} \in \left[\frac{1}{c_0}, c_0\right]$, we conclude that $\frac{1}{\rho c_0} \le \frac{P_l}{Q_l} \le \rho c_0$.

Now suppose $\text{bin}_l \cap R^c \neq \emptyset$. Since R is an interval and that $\mu\{R^c\} = o(H)$, and since $L \leq \frac{2}{H}$, we conclude that $\mu\{R^c\} < \frac{1}{2L}$ for all sufficiently large n. Thus, only bins $[0, \frac{1}{L}]$ and $[1 - \frac{1}{L}, 1]$ can satisfy $\text{bin}_l \cap R^c \neq 0$. Note that Assumption C5' implies both $p(x)$ and $q(x)$ are increasing in $\left[0, \frac{1}{L}\right]$ and decreasing in $\left[1 - \frac{1}{L}, 1\right]$. Define $P'_l = \int_{\text{bin}_l \cap R} p(x)dx$ and $Q'_l = \int_{\text{bin}_l \cap R} q(x)dx$, and define $P''_l = \int_{\text{bin}_l \cap R^c} p(x)dx$ and $Q''_l = \int_{\text{bin}_l \cap R^c} q(x)dx$. Then $P_l = P'_l + P''_l$ and $Q_l = Q'_l + Q''_l$. Note that $\frac{1}{\rho} \leq \frac{P'_l}{Q'_l} \leq \rho$ by the same argument as before. Furthermore, using the monotonic properties of $p(x)$ and $q(x)$ in the relevant intervals, we have

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

where the first inequality follows because $\mu(R^c) \leq \frac{1}{2L}$, and the second inequality follows from Assumption C5'. Similarly, $Q'_l \geq Q''_l$. Thus,

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

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

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

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

Likewise,

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

The key step in completing our proof is the following proposition, proved in Appendix D.1:

Proposition C.1. *Under Assumptions C1'–C5', we have*

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

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

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

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

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

We first notice that $\{\text{bin}_l\} \cap R$ constitute an approximately-uniform binning of R; i.e., there exist constants c_{bin} and C_{bin} such that $\frac{c_{\text{bin}}}{L} \leq |\text{bin}_l \cap R| \leq \frac{C_{\text{bin}}}{L}$. This is reasoned as follows: Since R is an interval, we know that $\text{bin}_l \cap R$ is an interval, as well. The inequality $|\text{bin}_l \cap R^c| \leq \mu\{R^c\} \leq \frac{1}{2L}$ then implies $\frac{1}{2L} \leq |\text{bin}_l \cap R| \leq \frac{1}{L}$.

In a series of lemmas, we show that the approximately-uniform binning of R leads to several useful bounds on H^R and H_L^R. In particular, Lemma D.1 shows that as long as L grows, we have $d_L \to \frac{1}{4}$, where $d_L := \sum_l Q_l' \left(\frac{1}{2} \frac{\gamma_l'}{Q_l'} \right)$ and $\gamma_l' = Q_l' - P_l'$. In Lemma D.2, we show that $H^R = \frac{\alpha^2}{4}(1+\eta)$, where $\eta = \Theta(\alpha)$. Similarly, Lemma D.3 establishes that $H_L^R = d_L \alpha^2 (1 + \eta_L)$. We combine the results of Lemmas D.1, D.2, and D.3 in Lemma D.4, to show that $|H^R - H_L^R| = o(H^R)$. The last step is to bound the difference between the sums and integrals over R and the entire real line. □

D Appendix for Proposition 6.3

D.1 Proof of Proposition C.1

Let a_L be an $o(1)$ sequence such that $\mu(R^c) \leq a_L H$. We divide the set of bins into three subsets:

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

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

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

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

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

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

Here, inequality (a) holds by the following reasoning: By the AM-GM inequality, we have $2\sqrt{P_l' Q_l' P_l'' Q_l''} \leq P_l' Q_l'' + P_l'' Q_l'$. Thus,

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

Taking square roots, we conclude that $\sqrt{P_l' Q_l'} + \sqrt{P_l'' Q_l''} \leq \sqrt{P_l Q_l}$, yielding (a).

On the other hand, we have

$$\sqrt{P_l Q_l} - \sqrt{P_l' Q_l'} = \frac{P_l Q_l - P_l' Q_l'}{\sqrt{P_l Q_l} + \sqrt{P_l' Q_l'}}$$

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

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

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

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

We still need to bound the last term $\frac{Q_l''P_l''}{2\sqrt{P_l'Q_l'}}$. Since $l \in L_2$, either $P_l \geq 2Ca_LH$ or $Q_l \geq 2Ca_LH$. Suppose the former inequality holds; the latter case may be handled in an identical manner. Since $P_l'' \leq Ca_LH$ and $P_l \geq 2Ca_LH$, we have $P_l'' \leq P_l'$, so

$$\frac{Q_l''P_l''}{2\sqrt{P_l'Q_l'}} \leq Q_l''\frac{P_l'}{2\sqrt{P_l'Q_l'}} \leq \sqrt{\rho}Q_l''.$$

Putting everything together, we have

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

Thus,

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

Combining these two bounds yields

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

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

Step 2: In step 2, we verify that $\{\text{bin}_l\}_{l\in L_1} \cup \{\text{bin}_l \cap R\}_{l\in L_2} \cup \{\text{bin}_l \cap R\}_{l\in L_3}$ constitutes a valid approximately-uniform binning of R. First, since R is an interval, it is easy to see that $\text{bin}_l \cap R$ is also an interval. Second, we have $|\text{bin}_l \cap R^c| \leq \mu\{R^c\} \leq a_LH$. Since $\frac{1}{H} \leq L$ by assumption, we have $\mu\{R^c\} \leq \frac{a_L}{L}$, so there exists a constant 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 may bound $|H - H_L|$ as

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

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

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

D.2 Lemmas for Proposition 6.3

Lemma D.1. *Let $d_L = \sum_l Q'_l \left(\frac{\gamma'_l}{Q_l}\right)^2$. Suppose Assumptions C1'–C5' hold. Then $\lim_{L \to \infty} d_L = \frac{1}{4}$.*

Proof. Let $h(x)$ be as defined in Assumption C3'; in particular, $|h(x)| \ge \left|\frac{\gamma'(x)}{q(x)}\right| \vee \left|\frac{q'(x)}{q(x)}\right|$. For a parameter $0 < \tau < 1$ to be chosen later, we call bin$_l$ *good* if

$$
\sup_{x \in \mathrm{bin}_l} |h(x)| \le L^\tau.
$$

We first argue that the proportion of bad bins converges to 0 as $L \to \infty$. Since $h(x)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped, the set $\{x : |h(x)| \ge L^\tau\}$ is a union of at most two intervals, for all $L \ge C_s'^{1/\tau}$. Using the notation $B_l = b_l - a_l$, we have

$$
\sum_{l \in \{l \,:\, |h(x)| \ge L^\tau\}} B_l \le \mu\left(\{x : |h(x)| \ge L^\tau\}\right) + \frac{4C_{\mathrm{bin}}}{L} \overset{(a)}{\le} \frac{C}{L^{\tau t}} + \frac{4C_{\mathrm{bin}}}{L} \overset{(b)}{\le} \frac{C}{L^{\tau t}},
$$

where (a) follows because $\int_R |h(x)|^t dx < \infty$ by Assumption C4'; and (b) follows because $t \le 1$ by Assumption C4', and the fact that $\tau t < 1$ by choice. In particular, the number of bad bins may be bounded as follows:

$$
\#\{l : |h(x)| \ge L^\tau\} \le \frac{CL^{-\tau t} L}{C_{\mathrm{bin}}} \le CL^{1-\tau t},
$$

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

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

$$
\begin{aligned}
&\overset{(a)}{\leq} Q'_l \int_{\mathrm{bin}_l} \frac{q(x)}{Q'_l} \left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&\leq \int_{\mathrm{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&\overset{(b)}{\leq} \left(\int_{\mathrm{bin}_l} q(x) \left|\frac{\gamma(x)}{q(x)}\right|^r dx\right)^{2/r} \left(\int_{\mathrm{bin}_l} q(x) dx\right)^{(r-2)/r} \\
&\overset{(c)}{\leq} C(C_{\mathrm{bin}})^{(r-2)/r} L^{-(r-2)/r} = CL^{-(r-2)/r}.
\end{aligned}
$$

Here, (a) follows from Jensen's inequality, (b) follows from Hölder's inequality, and (c) follows because $\int_R q(x) \left|\frac{\gamma(x)}{q(x)}\right|^r dx < \infty$ by Assumption C3' and the fact that $\int_{\mathrm{bin}_l} q(x)dx \leq \frac{CC_{bin}}{L}$.

We now have

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

For each good bin l, define $x_l = \arg\max_{x \in \mathrm{bin}_l} |q(x)|$. The maximum is attainable since q is continuous and bounded. Furthermore,

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

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

$$
\begin{aligned}
B_l \left|\frac{\xi_l}{q(x_l)}\right| &\leq \frac{1}{B_l} \int_{a_l}^{b_l} \left|\frac{q'(c_x)}{q(x_l)}\right| |x - x_l| dx \overset{(a)}{\leq} \frac{1}{B_l} \int_{a_l}^{b_l} \left|\frac{q'(c_x)}{q(c_x)}\right| |x - x_l| dx \\
&\overset{(b)}{\leq} \frac{1}{B_l} \int_{a_l}^{b_l} L^\tau |x - x_l| dx \leq C_{\mathrm{bin}} L^{\tau - 1},
\end{aligned}
$$

where (a) follows because $q(c_x) \leq q(x_l)$, and (b) follows because l is a good bin, so $\left|\frac{q'(c_x)}{q(c_x)}\right| \leq L^\tau$. The last inequality follows because $B_l \leq \frac{C_{\mathrm{bin}_l}}{L}$. We may perform a similar analysis on γ:

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

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

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

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 good bin l:

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

To arrive at the last equality, we assume that $L \geq C_{\text{bin}}^{1/(1-\tau)}$. Then $\left|B_l \frac{\xi'_l}{q(x_l)}\right| \leq \frac{1}{2}$, so we may take a Taylor approximation. Here, η_l is a constant satisfying $|\eta_l| \leq 16$. Expanding the right-hand side, we have

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

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 $\frac{C_{\text{bin}}}{2} L^{\tau-1} \leq \frac{1}{4}$. Then

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

We now bound

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

The third term is bounded by $C_1 L^{2\tau-3}$, for a suitable constant C_1. To bound the second term, we split into two cases:

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

Case 2: $\left|\frac{\gamma(x_l)}{q(x_l)}\right| \leq 1$. 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 either case, we have

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

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

$$\begin{aligned}
|d_L - d_R| &= \left|\sum_l Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2 - \sum_{l\, good} B_l q(x_l) \left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| \\
&\leq \sum_{l\, good} \left|Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2 - B_l q(x_l) \left(\frac{\gamma(x_l)}{q(x_l)}\right)^2\right| + C_{M,M',K,C} L^{\frac{2}{\tau} - \tau t}
\end{aligned}$$

$$\begin{aligned}
&\leq C_3 d_R L^{\tau-1} + L \cdot C_4 L^{\tau-2} + C_5 L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\tau-1} + C_4 L^{\tau-1} + C_5 L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\frac{2-rt}{r(1+t)}} + C_6 L^{\frac{2-rt}{r(1+t)}},
\end{aligned}$$

where we have made the choice $\tau = \frac{2+r}{r(1+t)}$ in the last inequality to balance $L^{\tau-1}$ and $L^{2/r-\tau t}$. Notice that $0 < \tau < 1$ by Assumption C4', since $rt > 2$. Furthermore, $|d_L - d_R| = o(d_R) + o(1)$.

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

$$\begin{aligned}
d = \int_R q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx &= \sum_{l=1}^{L} \int_{\text{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&= \sum_{l\, good} \int_{\text{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx + \sum_{l\, bad} \int_{\text{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx \\
&\leq \sum_{l\, good} \int_{\text{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx + |\{l \,:\, l\, \text{bad}\}| C L^{-\frac{2}{r}} \\
&\leq \sum_{l\, good} \int_{\text{bin}_l} q(x) \left(\frac{\gamma(x)}{q(x)}\right)^2 dx + C L^{\frac{2}{r}-\tau t}.
\end{aligned}$$

The bound on the second term follows from the previous analysis. For the first term, note that for all $x \in \text{bin}_l$, we have

$$q(x) = q(x_l) + q'(c_l)(x - x_l), \quad \text{and} \quad \gamma(x) = \gamma(x_l) + \gamma'(c_l')(x - x_l),$$

where $c_l, c_l' \in \text{bin}_l$ depend implicitly on x. For bin_l, we have

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

Denote

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

Observe that $|x - x_l| \leq B_l$ and

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

Similarly, $\left|\frac{q'(c_l)}{q(x_l)}\right| \leq L^{\tau}$. Hence, $|T_1|, |T_2| \leq C_{\text{bin}} L^{\tau-1}$. Now suppose $C_{\text{bin}} L^{\tau-1} \leq \frac{1}{2}$, which is satisfied if $L \geq (2C_{\text{bin}})^{\frac{1}{1-\tau}}$. 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 satisfying $|\eta| \leq 16$. Thus,

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

The same analysis used to bound $|d_L - d_R|$ implies that $|d - d_R| = o(d_R) + o(1)$. Since $d = \frac{1}{4}$, we have $d_R \to \frac{1}{4}$, which in turn implies $d_L \to \frac{1}{4}$. This completes the proof. □

Lemma D.2. *Let*

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

Suppose Assumptions C1'–C5' hold. Then $H^R = \frac{\alpha^2}{4}(1+\eta)$, where $|\eta| \leq C(\alpha+\alpha^2)$ for some constant C. In particular, if $H^R \to 0$, then $\alpha \to 0$ and $\eta \to 0$.

Proof. We write

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

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

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

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

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

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 $\xi_2(x) \leq 2\left|\frac{\delta(x)}{q(x)}\right|$, implying that

$$|\eta| \leq \int_R q(x)\left(\frac{\gamma(x)}{q(x)}\right)^2 \left(4\left|\frac{\delta(x)}{q(x)}\right|^2 + 4\left|\frac{\delta(x)}{q(x)}\right|\right) dx$$

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

using the finiteness of integrals in Assumption C3'. □

Lemma D.3. *Let*

$$
\begin{aligned}
H_L^R &= \sum_{l=1}^{L} \left(\sqrt{P_l'} - \sqrt{Q_l'} \right)^2, \\
\delta(x) &= q(x) - p(x), \\
\alpha^2 &= \int_R q(x) \left(\frac{\delta(x)}{q(x)} \right)^2 dx, \\
\gamma(x) &= \frac{\delta(x)}{\alpha} dx.
\end{aligned}
$$

Suppose that Assumptions C1'–C5' hold. Then $H_L^R = d_L(1+\eta_L)$*, 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 $\delta_l = Q_l' - P_l'$. We have

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

where by convention, we define $\frac{\delta_l}{Q_l'} = 0$ when $Q_l', P_l' = 0$, and we use the shorthand $\xi_l = 1 - \frac{1}{2} \frac{\delta_l}{Q_l'} - \sqrt{1 - \frac{\delta_l}{Q_l'}}$. Hence,

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

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

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

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 $|\xi_{2l}| \leq 2 \left| \frac{\delta_l}{Q_l'} \right|$. 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 $\frac{1}{4}$ by Lemma D.1 and may be bounded by $1/(2C')$ for large enough L. To bound the numerator, note that for a single l, we have

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

Therefore,

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

implying that $|\eta_L| \leq (2\alpha + \alpha^2) 2C'M$. □

Lemma D.4. *Suppose Assumptions A1'–A4' hold. For any sequences $L_n, \alpha_n \to \infty$, we have $H_L^R = H^R(1 + o(1))$; i.e.,*

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

Proof. By Lemmas D.2 and D.3, we have

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

implying that

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

where $|\eta|, |\eta_L| \leq C_1(\alpha + \alpha^2)$ for all L. Thus,

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

Furthermore, by Lemma D.1, we have $|4d_L - 1| \to 0$, uniformly for all α. Thus, $\left|\frac{H_L^R}{H^R} - 1\right| \to 0$, completing the proof. □

E Proof of Proposition 6.2

First, we prove the bounds on $\frac{\widetilde{P}_l}{\widetilde{Q}_l}$. Define

$$R = \left\{x \in [0, 1] \,:\, \left|\log \frac{p(x)}{q(x)}\right| \leq C(2L)^{1/r}\right\},$$

where $C = \left(\int \left|\log \frac{p(x)}{q(x)}\right|^r dx\right)^{1/r}$ is a constant. Since $\int \left|\log \frac{p(x)}{q(x)}\right|^r dx < \infty$, Markov's Inequality implies $\mu\{R^c\} \leq \frac{1}{2L}$.

The remainder of the proof follows the argument used to prove Proposition 6.3, except for the final step, where we need to show that

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

We establish this fact in the following proposition:

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

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

Proof. We use a similar argument to the proof of Proposition D.1. First observe that

$$\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \int p(x)dx + \int q(x)dx - 2\int \sqrt{p(x)q(x)}dx = 2 - 2\int \sqrt{p(x)q(x)}$$

and

$$\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 = \sum_l P_l + \sum_l Q_l - 2\sum_l \sqrt{P_l Q_l}.$$

Thus, we only need to show that

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

We have $|h(x)| \geq \left|\frac{p'(x)}{p(x)}\right| \vee \left|\frac{q'(x)}{q(x)}\right|$. Let $0 < \tau < 1$. We call bin_l *good* if

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

We now argue that the proportion of bad bins converges to 0 as $L \to \infty$: Since $h(x)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped, the set $\{x : |h_n(x)| \geq L^\tau\}$ is a union of at most two intervals, for all $L \geq C'^{1/\tau}_s$. Hence,

$$\begin{aligned} \sum_{l \in \left\{l \,:\, \sup_{x\in\text{bin}_l} |h(x)| \geq L^\tau\right\}} B_l &\leq \mu\left(\left\{x \,:\, \sup_{x\in\text{bin}_l} |h(x)| \geq L^\tau\right\}\right) + 4C_{\text{bin}}L^{-1} \\ &\overset{(a)}{\leq} CL^{-\tau t} + 4C_{\text{bin}}L^{-1} \overset{(b)}{\leq} CL^{-\tau t}, \end{aligned}$$

where (a) follows because $\int_R |h(x)|^t dx < \infty$ by Assumption C3; and (b) follows because $t \leq 1$, so $\tau t < 1$ and the first term dominates. We now bound the number of bad bins:

$$\#\{l \,:\, |h(x)| \geq L^\tau\} \leq \frac{CL^{-\tau t}L}{C_{\text{bin}}} \leq CL^{1-\tau t}.$$

For a bad bin, we have $P_l, Q_l \leq \frac{CC_{\text{bin}}}{L}$ and $\int_{\text{bin}_l} (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \leq \frac{2CC_{\text{bin}}}{L}$.

We now 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 bounded. We have

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

where $\xi_l = \frac{1}{B_l^2}\int_{a_l}^{b_l} p'(c_x)(x - x_l)dx$. Furthermore,

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

Likewise, define $x_l' = \arg\max_{x\in \text{bin}_l} q(x)$. We have $Q_l = B_l q(x_l') + B_l^2 \xi_l'$, where

$$\xi_l' := \frac{1}{B_l}\int_{a_l}^{b_l} q'(c_x)(x - x_l')dx.$$

We can also bound $B_l \left|\frac{\xi_l'}{q(x_l')}\right| \leq C_{\text{bin}} L^{\tau-1}$. Thus,

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

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 as

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

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

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

and likewise,

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

Since

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

we may bound the nuisance term as follows:

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

The term $B_l\sqrt{p(x_l)q(x_l')}$ is bounded by $B_l C$. Hence,

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

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

$$\left| \sqrt{P_l Q_l} - \int_{a_l}^{b_l} \sqrt{p(x)q(x)}dx \right| \leq B_l C C_{\text{bin}} L^{\tau-1}.$$

Hence,

$$\begin{aligned}
\left| \sum_l \sqrt{P_l Q_l} - \int \sqrt{p(x)q(x)}dx \right| &\leq \sum_{l:\ l \text{ bad}} B_l C + \sum_{l:\ l \text{ good}} \left| \sqrt{P_l Q_l} - \int_{a_l}^{b_l} \sqrt{p(x)q(x)}dx \right| \\
&\leq C L^{-\tau t} + \sum_{l:\ l \text{ good}} B_l C C_{\text{bin}} L^{\tau-1} \\
&\leq C L^{-\tau t} + C C_{\text{bin}} L^{\tau-1}.
\end{aligned}$$

Setting $\tau = \frac{1}{1+t}$, we obtain

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

completing the proof. □

F Proofs of Theorems 5.2 and Theorem 5.1

We now outline the proofs of Theorems 5.2 and 5.1, with proofs of supporting propositions in the succeeding subsections.

F.1 Main argument: Proof of Theorem 5.2

By the argument outlined in Section 6.3, the divergences I and H do not change after transforming the densities $p(x)$ and $q(x)$ according to Φ. Proposition F.1 shows that under Assumptions A1'–A5', Assumptions C1'–C5' are also satisfied.

Furthermore, our assumption that $L = o(\frac{1}{H})$ implies $L \leq \frac{2}{H}$ for sufficiently large L. Hence, Proposition 6.3 applies, and we may conclude that after transformation and discretization, the label probabilities satisfy $\frac{1}{2c_0\rho} \leq \frac{P_l}{Q_l} \leq 2c_0\rho$, for all l. Using the assumption $L = o(nI)$ and the fact that $I_L = I(1+o(1))$ from Proposition 6.3, we also have $L = o(nI_L)$, so we may use Proposition 6.1 (with $\rho_L = 2c_0\rho$) to obtain

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

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

F.2 Main argument: Proof of Theorem 5.1

The proof parallels the argument for Theorem 5.2 outlined above. Proposition F.2 establishes that Assumptions A1–A4 imply Assumptions C1–C4. Hence, Proposition 6.2 applies, and we may conclude that after transformation and discretization, the label probabilities 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 $I_L = I(1 + o(1))$. Therefore, we may again apply Proposition 6.1 (with $\rho_L = 2c_0 \exp(L^{1/r})$) to obtain

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

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

F.3 Transformation Analysis

Proposition F.1. *Let $p(x)$ and $q(x)$ be densities over S, where $S = \mathbb{R}$ or $S = \mathbb{R}^+$, and let $\Phi : S \to [0, 1]$ be a CDF such that $\phi = \Phi'$ is positive and continuous. Suppose Assumptions A1'–A5' hold.*

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' $p_\Phi(z), q_\Phi(z) > 0$ on $(0, 1)$, and $\sup_z \{p_\Phi(z) \vee q_\Phi(z)\} < \infty$.

C2' There exists a subinterval $R \subseteq [0, 1]$ such that

(a) for all $z \in R$, $\frac{1}{\rho} \leq \left|\frac{p_\Phi(z)}{q_\Phi(z)}\right| \leq \rho$, where ρ is an absolute constant, and

(b) $\mu\{R^c\} = o(H)$, where μ is the Lebesgue measure.

C3' Let $\alpha^2 = \int_R \frac{(p_\Phi(z)-q_\Phi(z))^2}{q_\Phi(z)} dz$ and $\gamma_\Phi(z) = \frac{q_\Phi(z)-p_\Phi(z)}{\alpha}$. Then $\int_R q_\Phi(z) \left|\frac{\gamma(z)}{q_\Phi(z)}\right|^r dz < \infty$, for an absolute constant $r \geq 4$.

C4' There exists $h_\Phi(z)$ such that

(a) $h_\Phi(z) \geq \max\left\{ \left|\frac{\gamma'_\Phi(z)}{q_\Phi(z)}\right|, \left|\frac{q'_\Phi(z)}{q_\Phi(z)}\right| \right\}$,

(b) $h_\Phi(z)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped, for absolute constants c'_{s1}, c'_{s2}, and C'_s, and

(c) $\int_R |h_\Phi(z)|^t dz < \infty$ for an absolute constant $\frac{2}{r} < t < 1$.

C5' $p'_\Phi(z), q'_\Phi(z) \geq 0$ and for all $z < c'_{s1}$, and $p'_\Phi(z), q'_\Phi(z) \leq 0$ for all $z > c'_{s2}$.

Proof. **C1'** follows from A1' and the condition that ϕ is positive and continuous.

To prove **C2'**, assume that A2' is true, and let R be a subinterval of $\mathbb{R}$ such that $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$. Define $R_\Phi = \{z \in [0, 1] : \Phi^{-1}(z) \in R\}$, so $\mathbf{1}_{R_\Phi}(z) = \mathbf{1}_R(\Phi^{-1}(z))$. Then R_Φ is clearly an interval, and $\Phi\{R^c\} = \mu\{R^c_\Phi\}$.

C3' follows from A3' via a change of variables.

It remains to prove **C4'** and **C5'**. We first prove **C5'**. Note that

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

Therefore, $p'_\Phi(z) \geq 0$ if and only if $p'(x) \geq p(x)\frac{\phi'(x)}{\phi(x)}$, and likewise for $q'_\Phi(z)$.

Moving onto **C4'**, we first construct $h(z)$. For ease of presentation, let $x = \Phi^{-1}(z)$. We then have

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

implying that

$$\left|\frac{q'_\Phi(z)}{q_\Phi(z)}\right| \leq \left|\frac{q'(x)}{q(x)}\frac{1}{\phi(x)}\right| + \left|\frac{\phi'(x)}{\phi(x)}\frac{1}{\phi(x)}\right| \lesssim (h(x) + 1)\frac{1}{\phi(x)},$$

where the last inequality follows because $\left|\frac{\phi'(x)}{\phi(x)}\right|$ is bounded. Furthermore,

$$\begin{aligned}\frac{\gamma'_\Phi(z)}{q_\Phi(z)} &= \frac{1}{\alpha}\frac{p'(x) - p(x)\frac{\phi'(x)}{\phi(x)} - q'(x) + q(x)\frac{\phi'(x)}{\phi(x)}}{q(x)\phi(x)} \\ &= \left(\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)} - \frac{1}{\alpha}\frac{p(x) - q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right)\frac{1}{\phi(x)},\end{aligned}$$

so

$$\begin{aligned}\left|\frac{\gamma'_\Phi(z)}{q_\Phi(z)}\right| &\leq \left|\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)}\right|\frac{1}{\phi(x)} + \left|\frac{1}{\alpha}\frac{p(x) - q(x)}{q(x)}\right|\left|\frac{\phi'(x)}{\phi(x)}\right|\frac{1}{\phi(x)} \\ &= \left|\frac{\gamma'(x)}{q(x)}\right|\frac{1}{\phi(x)} + \left|\frac{\gamma(x)}{q(x)}\right|\left|\frac{\phi'(x)}{\phi(x)}\right|\frac{1}{\phi(x)} \\ &\overset{(a)}{\lesssim} h(x)\frac{1}{\phi(x)},\end{aligned}$$

where (a) follows because $\left|\frac{\phi'(x)}{\phi(x)}\right|$ is bounded. We want to take $h_\Phi(z) \simeq (h(x)+1)\frac{1}{\phi(x)}$, but we use a modified upper bound to ensure that $h_\Phi(z)$ is bowl-shaped. Let $\psi(x) = \max\left\{\frac{1}{\phi(c_{s1})}, \frac{1}{\phi(c_{s2})}, \frac{1}{\phi(x)}\right\}$. We then take

$$h_\Phi(z) \simeq h(x)\psi(x) = h(\Phi^{-1}(z))\psi(\Phi^{-1}(z)).$$

Note that $(h(x)+1)$ is (c_{s1}, c_{s2}, C_s+1)-bowl-shaped, and ϕ is unimodal, so $\frac{1}{\phi(x)}$ is quasi-convex. Hence, $\psi(x)$ is quasi-convex and has a mode lying in $[c_{s1}, c_{s2}]$. Therefore, $(h(x)+1)\psi(x)$ is (c_{s1}, c_{s2}, C'_s)-bowl-shaped, where $C'_s \simeq (C_s+1)\left(\frac{1}{\phi(c_{s1})} \vee \frac{1}{\phi(c_{s2})}\right)$. This shows that $h_\Phi(z)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped for $c'_{s1} = \Phi(c_{s1})$ and $c'_{s2} = \Phi(c_{s2})$.

Finally, we need to verify the integrability conditions:

$$\begin{aligned}\int |h_\Phi(z)|^t dz &\simeq \int (h(\Phi^{-1}(z))+1)^t \psi(\Phi^{-1}(z))^t dz \\ &\overset{(a)}{=} \int_S (h(x)+1)^t \psi(x)^t \phi(x) dx \\ &\leq \left\{\int_S (h(x)+1)^{2t}\phi(x)dx\right\}^{1/2}\left\{\int_S \psi(x)^{2t}\phi(x)dx\right\}^{1/2},\end{aligned}$$

where (a) follows from a change of variables. To bound the first term, note that

$$\begin{aligned}\int_S (h(x)+1)^{2t}\phi(x)dx &\leq \int_S h(x)^{2t}\phi(x)dx + \int_S \phi(x)dx \\ &\leq \int_S h(x)^{2t}\phi(x)dx + 1.\end{aligned}$$

The first inequality follows since $2t < 1$ by assumption. Note that $\int_S h(x)^{2t}\phi(x)dx < \infty$.

We now bound the second term:

$$\int_S \psi(x)^{2t}\phi(x)dx \leq \int_S \phi(c_{s1})^{-2t}\phi(x)dx + \int_S \phi(c_{s2})^{-2t}\phi(x)dx + \int_S \phi(x)^{-2t}\phi(x)dx.$$

The first two terms are constants. The last term is $\int_S \phi(x)^{1-2t}dx$, which is finite because $1-2t > 0$ and ϕ is a valid transformation function. □

Proposition F.2. *Suppose Assumptions A1–A4 hold. 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 $p_\Phi(z), q_\Phi(z) > 0$ *on* $(0,1)$*, and* $\sup_z \{p_\Phi(z) \vee q_\Phi(z)\} < \infty$.

C2 For some $r > 2$, $\int \left|\log \frac{p_\Phi(z)}{q_\Phi(z)}\right|^r dz < \infty$.

C3 There exists $h_\Phi(z)$ *such that*

- *(a)* $h_\Phi(z) \geq \max\left\{\left|\frac{p'_\Phi(z)}{p_\Phi(z)}\right|, \left|\frac{q'_\Phi(z)}{q_\Phi(z)}\right|\right\}$,
- *(b)* $h_\Phi(z)$ *is* (c'_{s1}, c'_{s2}, C'_s)*-bowl-shaped, and*
- *(c)* $\int_R |h_{\Phi,n}(z)|^t dz < \infty$*, for some constant* t *such that* $\frac{2}{r} \leq t \leq 1$.

C4 We have that $p'_\Phi(z), q'_\Phi(z) \geq 0$ *for all* $z < c'_{s1}$*, and* $p'_\Phi(z), q'_\Phi(z) \leq 0$ *for all* $z > c'_{s2}$.

Proof. The proof is identical to that of Proposition F.1, so we omit the details. □

G Proof of Proposition 5.1

First suppose $\|\theta_1 - \theta_0\| \to 0$. In Lemma G.2, we show that $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$. Assumptions **A1'** and **A5'** follow directly from Assumptions B1 and B5, respectively.

We now prove **A2'**. Let ρ be a constant and define

$$R = \left\{x \,:\, g_1(x) \leq \frac{\log \rho}{\|\theta_1 - \theta_0\|}\right\}, \tag{G.1}$$

where $g_1(x)$ is the upper bound on $\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(x)\|$ defined in Assumption B3. Since $g_1(x)$ is bowl-shaped, we conclude that R is an interval if $\log \rho \geq C_s \mathrm{diam}(\Theta)$. Note that

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

This implies

$$\left|\log \frac{p(x)}{q(x)}\right| \leq \|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\| \leq \|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\| \leq \|\theta_1 - \theta_0\| \cdot g_1(x),$$

where $\overline{\theta}$ is a convex combination of θ_0, θ_1. Therefore, for all $x \in R$, we have $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$.

Since we know from Assumption B3 that $\int |g_1(x)|^r \phi(x) dx < \infty$, Markov's inequality gives

$$\Phi(R^c) = \Phi\left\{x \,:\, g_1(x) > \frac{\log \rho}{\|\theta_1 - \theta_0\|}\right\} \leq C \frac{\|\theta_1 - \theta_0\|^r}{(\log \rho)^r} \overset{(a)}{=} \Theta(H^{r/2}) = o(H),$$

where (a) follows from Lemma G.2. The last equality follows from the assumption that $r > 2$. This proves A2'.

Now we move on to **A3'**. By Lemma G.1, we have

$$\begin{aligned}
\frac{1}{\alpha}\left|\frac{p(x) - q(x)}{q(x)}\right| &\lesssim \frac{1}{\|\theta_1 - \theta_0\|}\left|\frac{p(x)}{q(x)} - 1\right| \\
&= \frac{1}{\|\theta_1 - \theta_0\|}\left|\exp(f_{\theta_1}(x) - f_{\theta_0}(x)) - 1\right| \\
&\overset{(a)}{=} \frac{1}{\|\theta_1 - \theta_0\|}\left|(\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x)\right| \exp(f_{\overline{\theta}}(x) - f_{\theta_0}(x))
\end{aligned}$$

$$\begin{aligned}
&\leq \|\nabla_\theta f_{\overline{\theta}}(x)\| \exp(f_{\overline{\theta}}(x) - f_{\theta_0}(x)) \\
&\overset{(b)}{=} \|\nabla_\theta f_{\overline{\theta}}(x)\| \exp\Big((\overline{\theta} - \theta_1)^\mathsf{T} \nabla_\theta f_{\widetilde{\theta}}(x)\Big) \\
&\leq \|\nabla_\theta f_{\overline{\theta}}(x)\| \exp\Big(\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\widetilde{\theta}}(x)\|\Big),
\end{aligned}$$

where in (a), $\overline{\theta}$ is a convex combination of θ_1 and θ_0, and in (b), $\widetilde{\theta}$ is a convex combination of $\overline{\theta}$ and θ_0. Assumption B3 implies that both $\|\nabla_\theta f_{\overline{\theta}}(x)\|$ and $\|\nabla_\theta f_{\widetilde{\theta}}(x)\|$ are upper-bounded by $g_1(x)$, so

$$\frac{1}{\alpha}\left|\frac{p(x) - q(x)}{q(x)}\right| \lesssim g_1(x) \exp\big(\|\theta_0 - \theta_1\| g_1(x)\big). \tag{G.2}$$

Therefore,

$$\begin{aligned}
\int_R \left(\frac{1}{\alpha}\left|\frac{p(x)}{q(x)} - 1\right|\right)^r q(x)dx &\lesssim \int_R g_1(x)^r \exp\big(r\|\theta_0 - \theta_1\| g_1(x)\big) q(x)dx \\
&\overset{(a)}{\leq} \int_R g_1(x)^r \rho^r q(x)dx \\
&\overset{(b)}{\lesssim} \int_R g_1(x)^r \phi(x)dx,
\end{aligned}$$

where (a) follows from the definition of R and (b) follows because $\frac{q(x)}{\phi(x)}$ is bounded. This proves A3'.

To prove **A4'**, we first construct $h(x)$. By equation G.2, we have

$$\left|\frac{\gamma(x)}{q(x)}\right| \lesssim g_1(x) \exp\Big(\|\theta_0 - \theta_1\| g_1(x)\Big).$$

By Assumption B3, we also have $\left|\frac{q'(x)}{q(x)}\right| = |f'_{\theta_0}(x)| \leq g_{2,\theta_0}(x)$, and

$$\begin{aligned}
\left|\frac{\gamma'(x)}{q(x)}\right| &= \left|\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)}\right| \\
&\lesssim \frac{1}{\|\theta_0 - \theta_1\|}\left|\frac{p'(x) - q'(x)}{q(x)}\right| \\
&= \frac{1}{\|\theta_0 - \theta_1\|}\left|f'_{\theta_1}\frac{p(x)}{q(x)} - f'_{\theta_0}(x)\right| \\
&= \frac{1}{\|\theta_0 - \theta_1\|}\left|(f'_{\theta_1}(x) - f'_{\theta_0}(x))\frac{p(x)}{q(x)} + f'_{\theta_0}\left(\frac{p(x)}{q(x)} - 1\right)\right| \\
&\leq \|\nabla_\theta f'_{\overline{\theta}}(x)\|\frac{p(x)}{q(x)} + \frac{1}{\|\theta_1 - \theta_0\|}\left|\frac{p(x)}{q(x)} - 1\right| |f'_{\theta_0}(x)|,
\end{aligned}$$

where $\overline{\theta}$ is a convex combination of θ_0 and θ_1.

Using Assumption B3 and inequality (G.2), we have

$$\left|\frac{\gamma'(x)}{q(x)}\right| \lesssim g_{2,\overline{\theta}}(x) \exp\Big(\|\theta_0 - \theta_1\| g_1(x)\Big) + g_1(x) \exp\Big(\|\theta_0 - \theta_1\| g_1(x)\Big) g_{2,\theta_0}(x).$$

Hence, we may choose choose

$$\begin{aligned}
h(x) \simeq\ & g_{2,\overline{\theta}}(x) \exp\Big(\|\theta_0 - \theta_1\| g_1(x)\Big) + g_1(x) \exp\Big(\|\theta_0 - \theta_1\| g_1(x)\Big) g_{2,\theta_0}(x) \\
&+ g_1(x) \exp\Big(\|\theta_0 - \theta_1\| g_1(x)\Big).
\end{aligned}$$

Since all the component functions are $(c_{s1}, c_{s2}, \widetilde{C}_s)$ bowl-shaped, $h(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped, where $C_s = 3\widetilde{C}_s^2 \exp(\operatorname{diam}(\Theta)\widetilde{C}_s)$. Furthermore,

$$\begin{aligned}
\int_R h(x)^{2t}\phi(x)dx &\overset{(a)}{\lesssim} \int_R \left(g_{2,\bar{\theta}}^{2t}\rho^{2t} + g_1(x)^{2t}\rho^{2t} g_{2,\theta_0}^{2t} + g_1(x)^{2t}\rho^{2t} \right)\phi(x)dx \\
&\lesssim \int_R g_{2,\bar{\theta}}(x)^{2t}\phi(x)dx + \int_R g_1(x)^{2t} g_{2,\theta_0}^{2t}\phi(x)dx + \int_R g_1(x)^{2t}\phi(x)dx,
\end{aligned}$$

where (a) follows because on R, we have $\|\theta_0 - \theta_1\| g_1(x) \le \log\rho$. By Assumption B3, the first and third terms are finite, uniformly over all $\theta_1, \theta_0 \in \Theta$. It is straightforward to show that the second term is also finite, by an application of the Cauchy-Schwartz inequality.

Now suppose $\|\theta_1 - \theta_0\| = \Theta(1)$. Lemma G.2 implies $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \Theta(1)$. Assumptions **A1** and **A4** follow directly from Assumptions B1 and B5.

To prove **A2**, note that from a previous derivation, we have

$$\left|\log \frac{p(x)}{q(x)}\right| \le \|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\| \le \|\theta_1 - \theta_0\| g_1(x).$$

Since $\|\theta_1 - \theta_0\| \le \operatorname{diam}(\Theta)$, which is a constant, and $\int g_1(x)^r \phi(x)dx < \infty$ by Assumption B3, we obtain A2.

To prove **A3**, note that

$$\frac{q'(x)}{q(x)} = f'_{\theta_0}(x), \quad \text{and} \quad \frac{p'(x)}{p(x)} = f'_{\theta_1}(x).$$

Therefore, the choice $h(x) = g_{2,\theta_0}(x) + g_{2,\theta_1}(x)$ upper-bounds $\left|\frac{q'(x)}{q(x)}\right|$ and $\left|\frac{p'(x)}{p(x)}\right|$, by Assumption B3. Furthermore, $h(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped, where $C_s = 2\widetilde{C}_s$.

To prove the last integrability condition, note that

$$\int h(x)^{2t}\phi(x)dx \le \int g_{2,\theta_0}(x)^{2t}\phi(x)dx + \int g_{2,\theta_1}(x)^{2t}\phi(x)dx.$$

Hence,

$$\int h(x)^{2t}\phi(x)dx \le 2\sup_{\theta\in\Theta} \int g_{2,\theta}(x)^{2t}\phi(x)dx.$$

G.1 Supporting lemmas

Lemma G.1. *Under Assumptions B1–B5, we have $\alpha \asymp \|\theta_1 - \theta_0\|$.*

Proof. We write

$$\begin{aligned}
\alpha^2 &= \int_R \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x)dx \\
&= \int_R \left|\exp\left(f_{\theta_1}(x) - f_{\theta_0}(x)\right) - 1\right|^2 q(x)dx \\
&= \int_R \left((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\bar{\theta}}(x) \exp\left(f_{\bar{\theta}}(x) - f_{\theta_0}(x)\right)\right)^2 q(x)dx.
\end{aligned}$$

First we show an upper bound:

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

$$\leq \int_R \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp\Big(\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\widetilde{\theta}}(x)\|\Big) \exp(f_{\overline{\theta}}(x))dx.$$

On R, we have $\|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\| \leq \log \rho$. Hence,

$$\begin{aligned}\alpha^2 &\leq \|\theta_1 - \theta_0\|^2 \int_R \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 e^{\log \rho} \exp(f_{\overline{\theta}}(x))dx \\ &\leq \|\theta_1 - \theta_0\|^2 \rho \int_{-\infty}^{\infty} \|\nabla_\theta f_{\overline{\theta}}(x)\|^2 \exp(f_{\overline{\theta}}(x))dx \\ &\overset{(a)}{\lesssim} \|\theta_1 - \theta_0\|^2,\end{aligned}$$

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

$$\begin{aligned}\alpha^2 &\geq \int_R \big((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x)\big)^2 \exp\Big(-|f_{\overline{\theta}}(x) - f_{\theta_0}(x)|\Big) \exp(f_{\overline{\theta}}(x))dx \\ &\geq \int_R \big((\theta_1 - \theta_0)^\mathsf{T} \nabla_\theta f_{\overline{\theta}}(x)\big)^2 \exp\Big(-\|\theta_1 - \theta_0\| \|\nabla_\theta f_{\overline{\theta}}(x)\|\Big) \exp(f_{\overline{\theta}}(x))dx \\ &\overset{(a)}{\geq} \frac{1}{\rho}(\theta_1 - \theta_0)^\mathsf{T} \Big(\int_R (\nabla_\theta f_{\overline{\theta}}(x))(\nabla_\theta f_{\overline{\theta}}(x))^\mathsf{T} \exp(f_{\overline{\theta}}(x))dx\Big)(\theta_1 - \theta_0),\end{aligned}$$

where (a) follows from Assumption B3. Define

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

As ρ increases, $R \to S$. Therefore, there exists an absolute constant such that for all ρ greater than or equal to this constant, we have $\lambda_{min}(\widetilde{G}_{\overline{\theta}}) > \frac{1}{2}\lambda_{min}(G_{\overline{\theta}}) > 0$. Hence, $\alpha^2 \gtrsim \|\theta_1 - \theta_0\|^2$. □

Lemma G.2. *The Hellinger distance satisfies the bound*

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

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

Proof. Expanding the left-hand side, we have

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

Let $h(\theta) = \exp\left(\frac{f_\theta(x)}{2} - \frac{f_{\theta_0}(x)}{2}\right)$. 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:

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

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

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

where $\Theta \subseteq \mathbb{R}^{d_\Theta}$. Furthermore,

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

□

G.2 Proof of examples

Proposition G.1. *Let* $\exp(f(x))$ *be a positive density over* $\mathbb{R}$*, where*

(a) $|f^{(k)}(x)|$ *is bounded for some* $k \ge 2$*, and*

(b) there exist constants c *and* M *such that* $f'(x) > M$ *for* $x < -c$ *and* $f'(x) < -M$ *for* $x > c$*.*

Let $\theta = (\mu, \sigma)$ *and* $\Theta = [-C_\mu, C_\mu] \times [\frac{1}{c_\sigma}, c_\sigma]$*, for some absolute constants* C_μ *and* c_σ*, and let*

$$f_\theta(x) = f\left(\frac{x - \mu}{\sigma} \right) - \log \sigma.$$

Then $\{f_\theta(x)\}_{\theta \in \Theta}$ *satisfies Assumptions B1–B4 with respect to* ϕ *defined in equation* (5.4)*.*

Proof. Before we prove the claims, let us derive some useful properties of f.

First, for any $x > c$, we have

$$f(x) = \int_0^x f'(t) dt = \int_0^c f'(t) dt + \int_c^x f'(t) dt \lesssim 1 - \int_c^x M dt \lesssim 1 - x.$$

Similarly, for any $x < -c$, we have $f(x) \lesssim 1 + x$. Therefore, $f(x) \lesssim 1 - |x|$.

Likewise, we have

$$f\left(\frac{x - \mu}{\sigma} \right) \lesssim 1 - \left| \frac{x - \mu}{\sigma} \right| \lesssim 1 - \left| \frac{x}{\sigma} \right| + \frac{\mu}{\sigma} \overset{(a)}{\lesssim} 1 - |x|,$$

where (a) follows because $\sigma \ge \frac{1}{c_\sigma}$ and $|\mu| \le C_\mu$, for some absolute constants c_σ and C_μ. Thus, the density $\exp f\left(\frac{x-\mu}{\sigma} \right)$ is sub-exponential.

Since $f^{(k)}(x)$ is bounded, L'Hopital's rule implies $|f'(x)| \lesssim |x|^{k-1} + 1$ and $|f''(x)| \lesssim |x|^{k-2} + 1$. Furthermore,

$$f'\left(\frac{x - \mu}{\sigma} \right) \lesssim \left| \frac{x - \mu}{\sigma} \right|^{k-1} + 1 \overset{(a)}{\lesssim} \left| \frac{x}{\sigma} \right|^{k-1} + \left| \frac{\mu}{\sigma} \right|^{k-1} + 1 \overset{(b)}{\lesssim} |x|^{k-1} + 1, \tag{G.3}$$

where (a) follows because k is a constant and (b) follows because $|\mu| \le C_\mu$ and $\sigma \ge \frac{1}{c_\sigma}$, by assumption.

Now we prove the first claim **B1**. We have

$$\begin{aligned}
\log \phi(x) - f_\theta(x) &= \log \frac{e}{8} - \sqrt{|x|+1} - f\left(\frac{x-\mu}{\sigma}\right) - \log \sigma \\
&\ge -\sqrt{|x|+1} - f\left(\frac{x-\mu}{\sigma}\right) - \log \frac{1}{c_\sigma} + \log \frac{e}{8} \\
&\ge -\sqrt{|x|+1} - C(1-|x|) - \log \frac{1}{c_\sigma} + \log \frac{e}{8} > -\infty.
\end{aligned}$$

Moving on to **B2**, we have

$$\nabla f_\theta(x) = \begin{bmatrix} -\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \\ -\left(\frac{x-\mu}{\sigma^2}\right) f'\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{\sigma} \end{bmatrix} = -\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \begin{bmatrix} 1 \\ \frac{x-\mu}{\sigma} + 1 \end{bmatrix}.$$

To show that $\lambda_{\max}(G_\theta) < \infty$, it is sufficient to show that

$$\begin{aligned}
&\int \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)^2 \left(\frac{x-\mu}{\sigma} + 1\right) \exp f\left(\frac{x-\mu}{\sigma}\right) dx < \infty, \quad \text{and} \\
&\int \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)^2 \left(\frac{x-\mu}{\sigma} + 1\right)^2 \exp f\left(\frac{x-\mu}{\sigma}\right) dx < \infty.
\end{aligned}$$

Since $\left|f'\left(\frac{x-\mu}{\sigma}\right)\right| \lesssim \left|\frac{x-\mu}{\sigma}\right|^{k-1} + 1$ and $\exp f\left(\frac{x-\mu}{\sigma}\right)$ is sub-exponential with all moments finite, we conclude that both integrals converge.

To show that $\lambda_{\min}(G_\theta) > 0$, we need to show that $\det(G_\theta) > 0$. Let $g(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)^2 \exp f\left(\frac{x-\mu}{\sigma}\right)$, and note that g is positive and integrable. The integral of g is not 0, since $|f'(x)| \ge M$ for all $|x| > c$. Thus, g may be normalized to a density $\overline{g}$.

Showing that $\det(G_\theta) > 0$ is equivalent to showing that

$$\int g(x)dx \int \left(\frac{x-\mu}{\sigma} + 1\right)^2 g(x)dx > \left(\int \left(\frac{x-\mu}{\sigma} + 1\right) g(x)dx\right)^2,$$

which is equivalent to showing that

$$\mathbb{E}_{\overline{g}}\left[\left(\frac{X-\mu}{\sigma} + 1\right)^2\right] - \left(\mathbb{E}_{\overline{g}}\left[\frac{X-\mu}{\sigma} + 1\right]\right)^2 > 0,$$

or $\mathrm{Var}_{\overline{g}}(X) > 0$. This follows because $g(x) \neq 0$.

To verify **B3**, note that

$$\begin{aligned}
\|\nabla f_\theta\| &= \left|\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)\right| \sqrt{1 + ((x-\mu)/\sigma + 1)^2} \\
&\lesssim (1 + |x|^{k-1})(1 + |(x-\mu)/\sigma|) \\
&\lesssim 1 + |x|^k.
\end{aligned}$$

Thus, we set $g_1(x) = C(1+|x|^k)$ for some absolute constant C. Note that $g_1(x)$ is clearly bowl-shaped and $\int g_1(x)^r \phi(x)dx$ is finite, since all moments of ϕ are finite. To construct $g_{2,\theta}$, note that

$$f'_\theta(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right), \quad \text{and} \quad \nabla f'_\theta(x) = \begin{bmatrix} -\frac{1}{\sigma^2} f''\left(\frac{x-\mu}{\sigma}\right) \\ -\frac{1}{\sigma^2} f'\left(\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma^3} f''\left(\frac{x-\mu}{\sigma}\right) \end{bmatrix}.$$

Therefore, $|f'_\theta(x)| \lesssim 1 + |x|^{k-1}$, and

$$\|\nabla f'_\theta(x)\| \le \frac{1}{\sigma^2}\left|f''\left(\frac{x-\mu}{\sigma}\right)\right| + \frac{1}{\sigma^2}\left|f'\left(\frac{x-\mu}{\sigma}\right)\right| + \frac{1}{\sigma^2}\left|f''\left(\frac{x-\mu}{\sigma}\right)\right|\left|\frac{x-\mu}{\sigma}\right|$$

$$\overset{(a)}{\lesssim} (1+|x|^{k-2}) + (1+|x|^{k-1}) + (1+|x|^{k-2})(1+|x|)$$
$$\lesssim 1+|x|^{k-1},$$

where (a) follows because $|f''(x)| \lesssim 1+|x|^{k-2}$. Thus, we may take $g_{2,\theta}(x) = C(1+|x|^{k-1})$. This is clearly bowl-shaped and integrable, as well.

Finally, we prove **B4**. We have

$$(\log\phi)'(x) = \begin{cases} \frac{1}{2}\frac{1}{\sqrt{1-x}}, & \text{if } x<0 \\ -\frac{1}{2}\frac{1}{\sqrt{1+x}}, & \text{if } x>0. \end{cases}$$

In particular, $(\log\phi)'(x) \to 0$ as $|x| \to \infty$.

We also know that $f'(x) \geq M$ for all $x \leq -c$, and $f'(x) \leq -M$ for all $x \geq c$. If $x \leq -\frac{c}{c_\sigma} - C_\mu$, then $\frac{x-\mu}{\sigma} \leq -c$ and

$$f'_\theta(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \geq \frac{M}{c_\sigma}.$$

If $x \geq \frac{c}{c_\sigma} + C_\mu$, then $\frac{x-\mu}{\sigma} \geq c$ and

$$f'_\theta(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \leq -\frac{M}{c_\sigma}.$$

Thus, there exist $c_{s1} < 0$ and $c_{s2} > 0$ such that B4 holds.

□

Proposition G.2. *Let $\exp(f(x))$ be a positive density over $\mathbb{R}^+$, where*

(a) $|f^{(k)}(x)|$ is bounded for some $k \geq 2$, and

(b) there exist constants c and M such that $f'(x) < -M$ for $x > c$.

Let $\theta = \sigma$ and $\Theta = [\frac{1}{c_\sigma}, c_\sigma]$ for some absolute constant c_σ, and let

$$f_\theta(x) = f\left(\frac{x}{\sigma}\right) - \log\sigma.$$

Then $\{f_\theta(x)\}_{\theta\in\Theta}$ satisfies Assumptions B1–B4 with respect to ϕ defined in equation (5.3).

The proof is almost identical to that of proposition G.1.

Proposition G.3. *Let $\theta = (\alpha, \beta)$ and $\Theta = [\frac{1}{c}, c]^2$ for some constant c, and let*

$$f_\theta = (\alpha-1)\log x - \beta x + \alpha\log\beta - \log\Gamma(\alpha).$$

Then $\{f_\theta(x)\}_{\theta\in\Theta}$ satisfies Assumptions B1–B4 with respect to ϕ defined in equation (5.3).

Proof. We first prove **B1**. We have $\log\phi(x) = \log\frac{e}{4} - \sqrt{x+1}$, so

$$\log\phi(x) - f_\theta(x) = -\sqrt{x+1} - (\alpha-1)\log x + \beta x + \log\frac{e}{4} - \alpha\log\beta - \log\Gamma(\alpha) > -\infty.$$

To prove **B2**, note that $G_\theta = \int (H_\theta f_\theta(x)) \exp f_\theta(x) dx$, where H_θ is the Hessian operator. Hence,

$$\nabla f_\theta(x) = \begin{bmatrix} \log x + \log\beta - d_\alpha \log\Gamma(\alpha) \\ -x + \frac{\alpha}{\beta} \end{bmatrix},$$

impyling that

$$H_\theta f_\theta(x) = \begin{bmatrix} d_\alpha^2 \log\Gamma(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & -\frac{\alpha}{\beta^2} \end{bmatrix}.$$

Therefore, $G_\theta = H_\theta f_\theta(x)$ which is clearly full-rank.

To prove **B3**, we write

$$\begin{aligned} \|\nabla f_\theta(x)\| &\leq |\log x| + x + |\log\beta| + |d_\alpha \log\Gamma(\alpha)| + \frac{\alpha}{\beta} \\ &\overset{(a)}{\leq} |\log x| + x + C, \end{aligned}$$

where (a) follows because $c \geq \beta, \alpha \geq \frac{1}{c}$. Therefore, we take $g_1(x) = |\log x| + x + C$. Then

$$\begin{aligned} \int g_1(x)^r \phi(x) dx &= \int_0^\infty (|\log x| + x + C)^r \phi(x) dx \\ &\lesssim \int_0^\infty |\log x|^r \phi(x) + \int_0^\infty x^r \phi(x) dx. \end{aligned}$$

Observe that both terms are finite for our choice of ϕ. Furthermore,

$$f_\theta'(x) = \frac{(\alpha-1)}{x} - \beta, \quad \text{and} \quad \nabla f_\theta'(x) = \begin{bmatrix} \frac{1}{x} \\ -1 \end{bmatrix},$$

implying that $|f_\theta'(x)| \lesssim 1 + x^{-1}$ and $\|\nabla f_\theta'(x)\| \lesssim 1 + x^{-1}$. Thus, $g_{2,\theta} = C(1 + x^{-1})$ satisfies

$$\int_0^\infty g_{2,\theta}(x)^{4t} \phi(x) dx \lesssim 1 + \int_0^\infty x^{-4t} \phi(x) dx.$$

Since $4t < 1$ by assumption, the integral converges.

B4 readily follows because $\beta \geq \frac{1}{c} > 0$. □

H Appendix for Theorem 5.3

We begin by defining some notation. Let $\widehat{\sigma}$ denote a clustering algorithm and A denote a weighted network such that $\widehat{\sigma}(A) : [n] \to [K]$ is the clustering obtained by $\widehat{\sigma}$ based on the input A. Let

$$S_K[\widehat{\sigma}(A), \sigma_0] := \operatorname*{arg\,min}_{\rho \in S_K} d_H(\rho \circ \widehat{\sigma}(A), \sigma_0),$$

where $d_H(\cdot, \cdot)$ denotes the Hamming distance, and define

$$\mathcal{E}[\widehat{\sigma}(A), \sigma_0] := \Big\{ v \, : \, (\rho \circ \widehat{\sigma}(A))(v) \neq \sigma_0(v), \text{ for some } \rho \in S_K[\widehat{\sigma}(A), \sigma_0] \Big\}. \tag{H.1}$$

When $S_K[\widehat{\sigma}(A), \sigma_0]$ is a singleton, the set $\mathcal{E}[\widehat{\sigma}(A), \sigma_0]$ contains all nodes misclustered by $\widehat{\sigma}(A)$ in relation to σ_0. When $S_K[\widehat{\sigma}(A), \sigma_0]$ contains multiple elements, we continue to call $\mathcal{E}[\widehat{\sigma}(A), \sigma_0]$ the set of *misclustered* nodes.

H.1 Proof of Theorem 5.3

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

$$\widetilde{l}(\widehat{\sigma}(A), \sigma_0) = \frac{1}{n} \sum_{v=1}^n \mathbb{1}\{v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]\},$$

where $\mathcal{E}[\widehat{\sigma}(A), \sigma_0]$ is defined in equation (H.1). In particular, note that if $|S_K[\widehat{\sigma}(A), \sigma_0]| = 1$, we have $\widetilde{l} = l$. We have the following claims:

Claim 1: If $\frac{nI}{K} \to \infty$, then $\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0) \geq C \exp\left(-(1+o(1))\frac{nI}{\beta K}\right)$.
Claim 2: If $\frac{nI}{K} \leq c < \infty$, then $\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0) \geq c' > 0$, for some constants c and c'.

We first prove that the theorem follows from the claims. If $P\left(l(\widehat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) \geq \frac{1}{2}\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0)$, we have

$$\mathbb{E}l(\widehat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K} P\left(l(\widehat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) \geq \frac{1}{4\beta K}\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0).$$

On the other hand, if $P\left(l(\widehat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) < \frac{1}{2}\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0)$, we have

$$\begin{aligned}
\mathbb{E}l(\widehat{\sigma}(A), \sigma_0) &\geq \mathbb{E}\left[l(\widehat{\sigma}(A), \sigma_0) \,\middle|\, l(\widehat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right] P\left(l(\widehat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right) \\
&\overset{(a)}{=} \mathbb{E}\left[\widetilde{l}(\widehat{\sigma}(A), \sigma_0) \,\middle|\, l(\widehat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right] P\left(l(\widehat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right) \\
&= \mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0) - \mathbb{E}\left[\widetilde{l}(\widehat{\sigma}(A), \sigma_0) \,\middle|\, l(\widehat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right] P\left(l(\widehat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) \\
&\geq \mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0) - \frac{1}{2}\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0) \\
&= \frac{1}{2}\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0),
\end{aligned}$$

where (a) holds by invoking Lemma B.8. Thus, any lower bound on $\mathbb{E}\widetilde{l}(\widehat{\sigma}(A), \sigma_0)$ translates into a lower bound on $\mathbb{E}l(\widehat{\sigma}(A), \sigma_0)$ scaled by a suitable constant, implying the desired result.

We now focus on proving the claims. Without loss of generality, suppose clusters 1 has size $\frac{n}{\beta K}+1$ and cluster 2 has size $\frac{n}{\beta K}$. Also suppose nodes 1 and 2 are such that $\sigma_0(1) = 1$ and $\sigma_0(2) = 2$. Let $C_i = \{u : \sigma_0(u) = i\}$ denote the i^{th} cluster.

Let $\sigma_0^1 := \sigma_0$, and let $\sigma_0^2 : [n] \to [K]$ be the cluster assignment satisfying $\sigma_0^2(v) = \sigma_0(v)$ for all $v \neq 1$, and $\sigma_0^2(1) = 2$. Let σ^* be a random cluster assignment, where

$$\sigma^* = \begin{cases} \sigma_0^1, & \text{with probability } \frac{1}{2}, \\ \sigma_0^2, & \text{with probability } \frac{1}{2}. \end{cases}$$

Let Φ denote the probability measure on $(\sigma^*, A, \widehat{\sigma}(A))$, defined by

$$P_\Phi(\sigma^*, A, \widehat{\sigma}(A)) = P(\sigma^*) P_{SBM}(A \,|\, \sigma^*) P_{alg}(\widehat{\sigma}(A) \,|\, A),$$

where $P_{SBM}(A \,|\, \sigma^*)$ is the measure on the weighted graph defined by the weighted SBM treating σ^* as the true cluster assignment, and P_{alg} represents any randomness in the clustering algorithm. Let Ψ denote an alternative probability measure defined by

$$P_\Psi(\sigma^*, A, \widehat{\sigma}(A)) = P(\sigma^*) P_\Psi(A \,|\, \sigma^*) P_{alg}(\widehat{\sigma}(A) \,|\, A),$$

where $P_\Psi(A \,|\, \sigma^*)$ is defined as follows:

1. If $u, v \neq 1$, then A_{uv} is distributed just as in $P_{SBM}(A \,|\, \sigma^*)$.

2. If $v = 1$ and $u \notin C_1 \cup C_2$, then A_{uv} is distributed just as in $P_{SBM}(A \,|\, \sigma^*)$.

3. If $v = 1$ and $u \in C_1 \cup C_2$, then A_{uv} is distributed as Y^*, where Y^* is the distribution that minimizes D in Lemma H.2; i.e., $Y_0^* \propto (P_0 Q_0)^{1/2}$ and $(1 - Y_0^*) y^*(x) \propto \sqrt{(1-P_0)p(x)(1-Q_0)q(x)}$.

Note that $P_\Psi(A \,|\, \sigma^*) = P_\Psi(A)$ actually does not depend on whether $\sigma^* = \sigma_0^1$ or σ_0^2.

Furthermore, we have

$$\begin{aligned}\mathcal{Q} := \log \frac{dP_\Psi}{dP_\Phi} &= \log \frac{dP_{SBM}(A \,|\, \sigma^*)}{dP_\Psi(A \,|\, \sigma^*)} \\ &= \sum_{u \in C_{\sigma^*(1)}} \log \frac{Y(A_{u,1})}{P(A_{u,1})} + \sum_{u \in C_1 \cup C_1 \setminus C_{\sigma^*(1)}} \log \frac{Y(A_{u,1})}{Q(A_{u,1})},\end{aligned}$$

where we use the notation $P(A_{u,1}) = P_0$ if $A_{u,1} = 0$ and $P(A_{u,1}) = (1 - P_0)p(A_{u,1})$ if $A_{u,1} \neq 0$, and similarly for Y. Let

$$E = \Big\{ 1 \notin \mathcal{E}[\widehat{\sigma}(A), \sigma^*] \text{ and } \widetilde{l}(\widehat{\sigma}(A), \sigma^*) \leq \frac{1}{4\beta K} \Big\}.$$

For an arbitrary function $f(n)$ to be defined later, we may write

$$P_\Psi(\mathcal{Q} \leq f(n)) = P_\Psi(\mathcal{Q} \leq f(n),\, \neg E) + P_\Psi(\mathcal{Q} \leq f(n),\, E). \tag{H.2}$$

We bound the first term as follows:

$$\begin{aligned}P_\Psi(\mathcal{Q} \leq f(n), \neg E) &= \int_{\mathcal{Q} \leq f(n), \neg E} dP_\Psi = \int_{\mathcal{Q} \leq f(n), \neg E} \exp(\mathcal{Q}) dP_\Phi \\ &\leq \exp(f(n)) P_\Phi(\mathcal{Q} \leq f(n), \neg E) \\ &\leq \exp(f(n)) P_\Phi(\neg E) \\ &\leq \exp(f(n)) \left(P_\Phi(1 \in \mathcal{E}[\widehat{\sigma}(A), \sigma^*]) + P_\Phi\left(\widetilde{l}(\widehat{\sigma}(A), \sigma^*) \geq \frac{1}{4\beta K} \right) \right).\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*) &= \frac{1}{n} \sum_{v=1}^{n} P_\Phi(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma^*]) \\ &\geq \frac{1}{n} \sum_{v \in C_{\sigma^*(1)}} P_\Phi(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma^*]) \\ &\overset{(a)}{=} \frac{|C_{\sigma^*(1)}|}{n} P_\Phi(1 \in \mathcal{E}[\widehat{\sigma}(A), \sigma^*]) \\ &\geq \frac{1}{\beta K} P_\Phi(1 \in \mathcal{E}[\widehat{\sigma}(A), \sigma^*]),\end{aligned}$$

where (a) follows from Corollary H.1, and

$$\begin{aligned}\mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*) &\geq \mathbb{E}_\Phi \Big[\widetilde{l}(\widehat{\sigma}(A), \sigma^*) \,\Big|\, \widetilde{l}(\widehat{\sigma}(A), \sigma^*) \geq \frac{1}{4\beta K} \Big] P_\Phi\left(\widetilde{l}(\widehat{\sigma}(A), \sigma^*) \geq \frac{1}{4\beta K} \right) \\ &\geq \frac{1}{4\beta K} P_\Phi\left(\widetilde{l}(\widehat{\sigma}(A), \sigma^*) \geq \frac{1}{4\beta K} \right).\end{aligned}$$

so we have the bound

$$P_\Psi(\mathcal{Q} \leq f(n), \neg E) \leq \exp(f(n)) \cdot 5\beta K \cdot \mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*).$$

We now turn to the second term in equation (H.2). We have

$$P_\Psi(E) = \frac{1}{2} P_\Psi \left(1 \notin \mathcal{E}[\widehat{\sigma}(A), \sigma_0^1] \text{ and } \widetilde{l}(\widehat{\sigma}(A), \sigma_0^1) \leq \frac{1}{4\beta K} \right)$$

$$+\,\frac{1}{2}P_\Psi\left(1 \notin \mathcal{E}[\widehat{\sigma}(A), \sigma_0^2] \text{ and } \widetilde{l}(\widehat{\sigma}(A), \sigma_0^2) \le \frac{1}{4\beta K}\right). \quad \text{(H.3)}$$

If $l(\widehat{\sigma}(A), \sigma_0^1) \le \widetilde{l}(\widehat{\sigma}(A), \sigma_0^1) \le \frac{1}{4\beta K}$, Lemma B.8 implies $S_K[\widehat{\sigma}(A), \sigma_0^1]$ contains only one element, which we denote by ρ. Since $d_H(\sigma_0^1, \sigma_0^2) = 1$, we have $\frac{1}{n}d_H(\rho \circ \widehat{\sigma}(A), \sigma_0^2) \le \frac{1}{4\beta K} + \frac{1}{n} \le \frac{1}{2\beta K}$, so applying Lemma B.8 again, we conclude that $\rho \in S_K[\widehat{\sigma}(A), \sigma_0^2]$, as well. However, $(\rho \circ \widehat{\sigma}(A))(1)$ cannot be equal to both $\sigma_0^1(1) = 1$ and $\sigma_0^2(1) = 2$. Hence, we cannot simultaneously have $1 \notin \mathcal{E}[\widehat{\sigma}(A), \sigma_0^1]$ and $1 \notin \mathcal{E}[\widehat{\sigma}(A), \sigma_0^2]$. In particular, the two events in equation (H.3) are disjoint, so

$$P_\Psi(\mathcal{Q} \le f(n), E) \le P_\Psi(E) \le \frac{1}{2}.$$

Plugging back into equation (H.2), we conclude that

$$P_\Psi(\mathcal{Q} \le f(n)) \le \exp(f(n)) \cdot 5\beta K \cdot \mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*) + \frac{1}{2},$$

so setting $f(n) = \log \frac{1}{20\beta K \mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*)}$, we have

$$P_\Psi\left(\mathcal{Q} \le \log \frac{1}{20\beta K \mathbb{E}_\Phi l(\widehat{\sigma}, \sigma^*)}\right) \le \frac{3}{4}.$$

By Chebyshev's inequality, we also have

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

where $V_\Psi(Q)$ is the variance of Q under Ψ. Hence, $\log \frac{1}{20\beta K \mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma_0)} \le \mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}$, or equivalently,

$$\mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*) \ge \frac{1}{20\beta K} \exp\Big(- (\mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})})\Big).$$

We now compute $\mathbb{E}_\Psi \mathcal{Q}$ and $V_\Psi(\mathcal{Q})$. Note that

$$\mathbb{E}_\Psi \mathcal{Q} = \frac{1}{2}\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^1] + \frac{1}{2}\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^2].$$

Furthermore, by Lemma H.2, we have

$$\begin{aligned}
\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^1] &= \mathbb{E}_\Psi\left[\sum_{u:\, u \ne 1,\, \sigma_0^1(u) = 1} \log \frac{Y(A_{u,1})}{P(A_{u,1})} + \sum_{u:\, \sigma_0^1(u) = 2} \log \frac{Y(A_{u,1})}{Q(A_{u,1})}\right] \\
&= \frac{n}{\beta K} \int \log \frac{dY}{dP} dY + \frac{n}{\beta K} \int \log \frac{dY}{dQ} dY \\
&= \frac{n}{\beta K} 2D = \frac{n}{\beta K} I.
\end{aligned}$$

Similarly, we have $\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^2] = \frac{nI}{\beta K}$, so $\mathbb{E}_\Psi \mathcal{Q} = \frac{nI}{\beta K}$. We show in Lemma H.3 that the following bound holds for the variance:

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

Now note that if $\frac{nI}{\beta K} \to \infty$, we have $\sqrt{\frac{nI}{\beta K}} = o\left(\frac{nI}{\beta K}\right)$, so $\sqrt{5V_\Psi(\mathcal{Q})} = o\left(\frac{nI}{\beta K}\right)$. Therefore,

$$\mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*) \ge C \exp\left(-(1 + o(1))\frac{nI}{\beta K}\right).$$

If instead $\frac{nI}{K} \to c < \infty$, then $\mathbb{E}_\Psi \mathcal{Q} = c(1+o(1))$ and $\sqrt{5V_\Psi(\mathcal{Q})} \leq C(1+o(1))$, so

$$\mathbb{E}_\Phi \widetilde{l}(\widehat{\sigma}(A), \sigma^*) \geq c' > 0,$$

for some constant c'.

Now define two measures P_1, P_2 on $(A, \widehat{\sigma}(A))$, as follows:

$$\begin{aligned} P_1(A, \widehat{\sigma}(A)) &= P_{SBM}(A \,|\, \sigma_0^1) P_{alg}(\widehat{\sigma}(A) \,|\, A), \\ P_2(A, \widehat{\sigma}(A)) &= P_{SBM}(A \,|\, \sigma_0^2) P_{alg}(\widehat{\sigma}(A) \,|\, A). \end{aligned}$$

Note that $\mathbb{E}_\Phi[\widetilde{l}(\widehat{\sigma}, \sigma^*) \,|\, \sigma^* = \sigma_0^1] = \mathbb{E}_1 \widetilde{l}(\widehat{\sigma}, \sigma_0^1)$ and $\mathbb{E}_\Phi[\widetilde{l}(\widehat{\sigma}, \sigma^*) \,|\, \sigma^* = \sigma_0^2] = \mathbb{E}_2 \widetilde{l}(\widehat{\sigma}, \sigma_0^2)$, where $\mathbb{E}_1$ and $\mathbb{E}_2$ are expectations taken with respect to P_1 and P_2, respectively. We claim that $\mathbb{E}_1 \widetilde{l}(\widehat{\sigma}, \sigma_0^1) = \mathbb{E}_2 \widetilde{l}(\widehat{\sigma}, \sigma_0^2)$, in which case $\mathbb{E}_\Phi l(\widehat{\sigma}, \sigma^*) = \mathbb{E}_1 \widetilde{l}(\widehat{\sigma}, \sigma_0^1) = \mathbb{E} \widetilde{l}(\widehat{\sigma}, \sigma_0)$ and the claims follow.

Define a permutation $\pi \in S_n$ that swaps $\{2, \dots, \frac{n}{\beta K} + 1\}$ with $\{\frac{n}{\beta K} + 2, \dots, 2\frac{n}{\beta K} + 1\}$ and satisfies $\pi(u) = u$ for $u = 1$ and $u \geq 2\frac{n}{\beta K} + 2$. Clearly, $\sigma_0^2 = \tau \circ \sigma_0^1 \circ \pi^{-1}$, where $\tau \in S_K$ swaps cluster labels 1 and 2. Now let A be fixed and let $\rho \in S_K$ be arbitrary. We have

$$\begin{aligned} d_H(\rho \circ \widehat{\sigma}(A), \sigma_0^1) &= d_H(\rho \circ \widehat{\sigma}(A), \tau^{-1} \circ \sigma_0^2 \circ \pi) \\ &= d_H(\rho \circ \widehat{\sigma}(A) \circ \pi^{-1}, \tau^{-1} \circ \sigma_0^2) \\ &= d_H(\rho \circ \xi^{-1} \circ \widehat{\sigma}(\pi A), \tau^{-1} \circ \sigma_0^2) \\ &= d_H(\tau \circ \rho \circ \xi^{-1} \circ \widehat{\sigma}(\pi A), \sigma_0^2). \end{aligned}$$

Therefore, $\rho \mapsto \tau \circ \rho \circ \xi^{-1}$ is a bijection between $S_K[\widehat{\sigma}(A), \sigma_0^1]$ and $S_K[\widehat{\sigma}(\pi A), \sigma_0^2]$. Furthermore, if v is a node such that $(\rho \circ \widehat{\sigma}(A))(v) \neq \sigma_0^1(v)$, we equivalently have

$$(\rho \circ \widehat{\sigma}(A))(v) \neq (\tau^{-1} \circ \sigma_0^2 \circ \pi)(v) \iff (\rho \circ \widehat{\sigma}(A) \circ \pi^{-1})(u) \quad \neq (\tau^{-1} \circ \sigma_0^2)(u), \quad \text{where } \pi(v) = u,$$

so $(\tau \circ \rho \circ \xi^{-1} \circ \widehat{\sigma}(\pi A))(u) \neq \sigma_0^2(u)$. Thus, $v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0^1]$ if and only if $\pi(v) \in \mathcal{E}[\widehat{\sigma}(\pi A), \sigma_0^2]$. Finally, we conclude that

$$\begin{aligned} \mathbb{E}_1 \widetilde{l}(\widehat{\sigma}(A), \sigma_0^1) &= \frac{1}{n} \sum_{v=1}^n P_1(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0^1]) \\ &= \frac{1}{n} \sum_{v=1}^n P_1(\pi(v) \in \mathcal{E}[\widehat{\sigma}(\pi A), \sigma_0^2]) \\ &\overset{(a)}{=} \frac{1}{n} \sum_{v=1}^n P_2(\pi(v) \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0^2]) \\ &= \mathbb{E}_2 \widetilde{l}(\widehat{\sigma}(A), \sigma_0^2), \end{aligned}$$

where (a) follows because $[\pi A]_{ij} = A_{\pi^{-1}(i), \pi^{-1}(j)}$, implying that if A is distributed according to $P_{SBM}(A \,|\, \sigma_0^1)$, then πA is distributed according to $P_{SBM}(A \,|\, \sigma_0^2)$. This concludes the proof.

H.2 Properties of permutation-equivariant estimators

Permutation-equivariant estimators possess symmetry properties. The following lemma formalizes one symmetry property useful for the proof of Theorem 5.3:

Lemma H.1. *Let the true clustering σ_0 be arbitrary. Suppose the weight matrix A is drawn from an arbitrary probability measure and $\widehat{\sigma}$ is any permutation-equivariant estimator. Let u and v be two nodes such that there exists $\pi \in S_n$ satisfying*

(1) $\pi(u) = v$,

(2) π *is measure-preserving; i.e.,* $A \overset{d}{=} \pi A$, *and*

(3) π *preserves the true clustering; i.e., there exists* $\tau \in S_K$ *such that* $\tau \circ \sigma_0 \circ \pi^{-1} = \sigma_0$.

Then

$$P(u \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]) = P(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]).$$

Proof. Since $\widehat{\sigma}(A) \overset{d}{=} \widehat{\sigma}(\pi A)$, we have

$$P\Big(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\widehat{\sigma}(\pi(A)), \sigma_0]\Big).$$

We claim that $u \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]$ if and only if $v \in \mathcal{E}[\widehat{\sigma}(\pi A), \sigma_0]$, implying the desired result:

$$P\Big(u \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\widehat{\sigma}(\pi A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]\Big).$$

Consider a fixed matrix A, and let $\tau \in S_K$ satisfy $\tau \circ \sigma_0 \circ \pi^{-1} = \sigma_0$. Let $\xi \in S_K$ be the permutation such that $\widehat{\sigma}(\pi A) = \xi \circ \widehat{\sigma}(A) \circ \pi^{-1}$. For any $\rho \in S_K$, we have

$$\begin{aligned} d_H(\rho \circ \widehat{\sigma}(A),\ \sigma_0) &= d_H(\tau \circ \rho \circ \xi^{-1} \circ \xi \circ \widehat{\sigma}(A) \circ \pi^{-1},\ \tau \circ \sigma_0 \circ \pi^{-1}) \\ &= d_H(\tau \circ \rho \circ \xi^{-1} \circ \widehat{\sigma}(\pi A),\ \sigma_0). \end{aligned}$$

Therefore, $\rho \in S_K[\widehat{\sigma}(A), \sigma_0]$ if and only if $\tau \circ \rho \circ \xi^{-1} \in S_K[\widehat{\sigma}(\pi(A)), \sigma_0]$. In particular, if $v \in \mathcal{E}[\widehat{\sigma}(\pi A), \sigma_0]$, we have $\tau \circ \rho \circ \xi^{-1} \circ \widehat{\sigma}(\pi A)(v) \neq \sigma_0(v)$ for some $\rho \in S_K[\widehat{\sigma}(A), \sigma_0]$. Then

$$\begin{aligned} \widehat{\sigma}(A)(u) = \widehat{\sigma}(A)(\pi^{-1}(v)) &= \xi^{-1} \circ \xi \circ \widehat{\sigma}(A) \circ \pi^{-1}(v) = \xi^{-1} \circ \widehat{\sigma}(\pi A)(v) \\ &\neq \rho^{-1} \circ \tau^{-1} \circ \sigma_0(v) = \rho^{-1} \circ \tau^{-1} \circ \sigma_0(\pi(u)) = \rho^{-1}(\sigma_0(u)). \end{aligned}$$

Thus, $u \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]$. Similar reasoning shows that if $u \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]$, then $v \in \mathcal{E}[\widehat{\sigma}(\pi A), \sigma_0]$. □

Corollary H.1. *Let the true clustering* σ_0 *be arbitrary. Suppose the weight matrix* A *is drawn from an arbitrary probability measure and* $\widehat{\sigma}$ *is any permutation-equivariant estimator. Let* u *and* v *be two nodes lying in equally-sized clusters. Then*

$$P\Big(u \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\widehat{\sigma}(A), \sigma_0]\Big).$$

Proof. By Lemma H.1, it suffices to construct a permutation $\pi \in S_n$ satisfying conditions (1)–(3).

First suppose u and v lie in the same cluster. It is easy to see that the conditions are satisfied when π is the permutation that transposes u and v and τ is the identity.

If u and v lie in different clusters, suppose without loss of generality that u is in cluster 1 and v is in cluster 2, where clusters 1 and 2 have the same size. Let π be the permutation that exchanges all nodes in cluster 1 with all nodes in cluster 2. The conditions are satisfied when τ is the permutation that transposes cluster labels 1 and 2. □

H.3 Properties of Renyi divergence

We first state a lemma that provides an alternative characterization of the Renyi divergence:

Lemma H.2. *Let* P *and* Q *be two probability measures on* $\mathbb{R}$ *that are absolutely continuous with respect to each other, with respective point masses* P_0 *and* Q_0 *at zero. The Renyi divergence satisfies* $I = 2D$, *where*

$$D := \inf_{Y \in \mathcal{P}} \max \left\{ \int \log \frac{dY}{dP} dY,\ \int \log \frac{dY}{dQ} dY \right\},$$

and $\mathcal{P}$ *denotes the set of probability measures absolutely continuous with respect to both* P *and* Q.

Proof. First, note that D is finite by making the choice $Y = P$. We claim that

$$D = \inf_{Y \in \mathcal{P}} \left\{ \int \log \frac{dY}{dP} dY : \int \log \frac{dP}{dQ} dY = 0 \right\}. \tag{H.4}$$

This is because for any $Y \in \mathcal{P}$ such that $\int \log \frac{dP}{dQ} dY \neq 0$, we have $\int \log \frac{dY}{dP} dY \neq \int \log \frac{dY}{dQ} dY$. Suppose without loss of generality that the first quantity is larger. Then 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, so the infimum in the definition of D could not have been achieved.

Since the new formulation (H.4) is convex in Y, we may solve to obtain the optimal $Y^* \in \mathcal{P}$, defined by $Y_0^* = \frac{P_0^{1/2} Q_0^{1/2}}{Z}$ and $(1 - Y_0^*) y^*(x) = \frac{((1-P_0)\cdot p(x))^{1/2}((1-Q_0)\cdot q(x))^{1/2}}{Z}$. The quantity Z is the normalization term: $Z = P_0^{1/2} Q_0^{1/2} + \int \sqrt{(1-P_0)p(x)(1-Q_0)q(x)} dx$. Then

$$\begin{aligned}
\int \log \frac{dY^*}{dP} dY^* &= \log \frac{1}{Z} \left\{ \left(\frac{Q_0}{P_0} \right)^{1/2} Y_0^* + \int \left(\frac{(1-P_0)p(x)}{(1-Q_0)q(x)} \right)^{1/2} (1 - Y_0^*) y^*(x) dx \right\} \\
&= \log \frac{dP}{dQ} dY^* - \log Z = - \log Z.
\end{aligned}$$

It is straightforward to verify that $-2 \log Z = I$. □

H.4 Bounding the variance

Lemma H.3. *For a suitable constant C, we have $\sqrt{5 V_\Psi(\mathcal{Q})} \leq C \sqrt{\frac{nI}{\beta K}}$.*

Proof. We begin with the decomposition

$$\begin{aligned}
V_\Psi(\mathcal{Q}) &= V(\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^*]) + E[V_\Psi(\mathcal{Q} \,|\, \sigma^*)] \\
&= E[V_\Psi(\mathcal{Q} \,|\, \sigma^*)] \\
&= \frac{1}{2} V_\Psi(\mathcal{Q} \,|\, \sigma^* = \sigma_0^1) + \frac{1}{2} V_\Psi(\mathcal{Q} \,|\, \sigma^* = \sigma_0^2).
\end{aligned}$$

Then

$$\begin{aligned}
V_\Psi(\mathcal{Q} \,|\, \sigma^* = \sigma_0^1) &= \sum_{u:\, u \neq 1, \sigma_0^1(u) = 1} V_\Psi \left(\log \frac{Y(A_{v_1 u})}{P(A_{v_1 u})} \right) + \sum_{u:\, \sigma_0^1(u) = 2} V_\Psi \left(\log \frac{Y(A_{v_1 u})}{Q(A_{v_1 u})} \right) \\
&\leq \frac{n}{\beta K} \mathbb{E}_\Psi \left(\log \frac{Y(A_{v_1 u})}{P(A_{v_1 u})} \right)^2 + \frac{n}{\beta K} \mathbb{E}_\Psi \left(\log \frac{Y(A_{v_1 u})}{Q(A_{v_1 u})} \right)^2.
\end{aligned}$$

We will show that $\mathbb{E}_\Psi \left(\log \frac{Y(A_{v_1 u})}{P(A_{v_1 u})} \right)^2$ is bounded by CI, so $\sqrt{5 V_\Psi(\mathcal{Q})} \leq C \sqrt{\frac{nI}{\beta K}}$. We have

$$\begin{aligned}
\mathbb{E}_\Psi \left(\log \frac{Y(A_{uv^*})}{P(A_{uv^*})} \right)^2 &= \int \left(\log \frac{dY}{dP} \right)^2 dY \\
&= Y_0 \log^2 \frac{Y_0}{P_0} + (1 - Y_0) \int y(x) \log^2 \frac{(1-Y_0)y(x)}{(1-P_0)p(x)} dx.
\end{aligned} \tag{H.5}$$

To bound the first term, we write

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

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

where (a) follows from Lemma I.2 and the fact that $\frac{Q_0}{P_0}$ is bounded, and (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. Therefore,

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

where (a) follows because $(x+y)^2 \leq 2x^2+2y^2$, and (b) follows because $Y_0 = \frac{\sqrt{P_0Q_0}}{Z} = (1+o(1))CP_0$.

Since $I = o(1)$, we have $I^2 \leq I$. Also,

$$I \geq (1+o(1))(\sqrt{P_0}-\sqrt{Q_0})^2 = (1+o(1))\frac{(P_0-Q_0)^2}{(\sqrt{P_0}+\sqrt{Q_0})^2} = C(1+o(1))\frac{(P_0-Q_0)^2}{P_0},$$

from which we conclude that $Y_0 \log^2 \frac{Y_0}{P_0} \leq CI$.

Now we turn our attention to the second term in equation (H.5). We have

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

Therefore,

$$\begin{aligned}(1-Y_0)\int y(x)\left(\log\frac{1-Y_0}{1-P_0}\frac{y(x)}{p(x)}\right)^2 dx &\leq (1-Y_0)\int y(x)\left\{\frac{1}{2}\left|\log\frac{1-Q_0}{1-P_0}\right| + \frac{1}{2}\left|\log\frac{q(x)}{p(x)}\right| + \frac{I}{2}\right\}^2 dx \\ &\leq (1-Y_0)\int y(x)\left\{C\left|\log\frac{1-Q_0}{1-P_0}\right|^2 + C\left|\log\frac{q(x)}{p(x)}\right|^2 + CI^2\right\} dx,\end{aligned} \tag{H.6}$$

where we have used the fact that $(x+y+z)^2 \leq 9x^2+9y^2+9z^2$ in the last inequality. Define

$$\begin{aligned}\mathcal{A} &:= (1-Y_0)\int y(x)\left|\log\frac{1-Q_0}{1-P_0}\right|^2 dx, \\ \mathcal{B} &:= (1-Y_0)\int y(x)\left|\log\frac{q(x)}{p(x)}\right|^2 dx, \\ \mathcal{C} &:= (1-Y_0)\int y(x) I^2 dx.\end{aligned}$$

We bound each term separately, beginning with $\mathcal{A}$. Note that

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

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

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

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

$$I \ge (1+o(1))(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 = (1+o(1))\frac{(P_0-Q_0)^2}{(\sqrt{1-P_0}+\sqrt{1-Q_0})^2} = C(1+o(1))\frac{(P_0-Q_0)^2}{1-P_0},$$

implying that $\mathcal{A} \le C(1+o(1))I \le CI$.

Moving onto $\mathcal{B}$, first suppose $H = \Theta(1)$ and $\max\left\{\int p(x)\left|\log\frac{q(x)}{p(x)}\right|^2 dx,\ \int q(x)\left|\log\frac{q(x)}{p(x)}\right|^2 dx\right\} < \infty$. We have

$$\begin{aligned}
\mathcal{B} &\le C\int \frac{\sqrt{(1-P_0)p(x)(1-Q_0)q(x)}}{Z}\left|\log\frac{q(x)}{p(x)}\right|^2 dx \\
&\overset{(a)}{\le} C\sqrt{(1-P_0)(1-Q_0)}\int \sqrt{p(x)q(x)}\left|\log\frac{q(x)}{p(x)}\right|^2 dx \\
&\le C\sqrt{(1-P_0)(1-Q_0)}\int (p(x)+q(x))\left|\log\frac{q(x)}{p(x)}\right|^2 dx \\
&\overset{(b)}{\le} C\sqrt{(1-P_0)(1-Q_0)}H \le CI,
\end{aligned}$$

where (a) follows because $Z \to 1$ and (b) follows because $H = \Theta(1)$.

Next, we make no assumption on H but assume $\left|\log\frac{q(x)}{p(x)}\right|$ is bounded by a constant. We have

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

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 $\left|\frac{p(x)-q(x)}{p(x)}\right| = \left|1-\frac{q(x)}{p(x)}\right| \le 1 + \left|\frac{q(x)}{p(x)}\right| \le C$. Then

$$\mathcal{B} \le \frac{C}{Z}\int \sqrt{\frac{1-Q_0}{1-P_0}\frac{q(x)}{p(x)}}(1-P_0)p(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx$$

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

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

$$\mathcal{B} \leq C(1-P_0)H \leq C\sqrt{(1-P_0)(1-Q_0)}H \leq CI.$$

Finally, note that $\mathcal{C} = (1-Y_0)CI^2 \leq CI$. Substituting back into inequality (H.6), we therefore obtain

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

so substituting back into inequality (H.5), we obtain the desired bound. □

I Additional useful lemmas

Lemma I.1. *Let*

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

If $I^h < 2-2\epsilon$, then $I = I^h(1+\eta)$, where $|\eta| \leq \frac{I^h}{2\epsilon}$. Thus, $I \to 0$ if and only if $I^h \to 0$, in which case $I = I^h(1+o(1))$.

Proof. We have

$$\begin{aligned} I &= -2\log\left(\sqrt{P_0Q_0} + \int\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right) \\ &= -2\log\left(1 - \frac{1}{2}\left((\sqrt{P_0}-\sqrt{Q_0})^2 + \int(\sqrt{(1-P_0)p(x)}-\sqrt{(1-Q_0)q(x)})^2dx\right)\right) \\ &= -2\log\left(1-\frac{1}{2}I^h\right) \\ &= 2\cdot\frac{1}{2}I^h(1+\eta), \end{aligned}$$

where $|\eta| \leq \frac{I^h}{2\epsilon}$. The last equality follows from Lemma I.2. □

Lemma I.2. *Suppose $0 < \epsilon \leq 1$. For all $0 \leq x < 1-\epsilon$, we have $\log(1-x) = -(1+\eta)x$, 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. *Let $f(z) = \frac{1-\frac{z}{2}-\sqrt{1-z}}{z}$, for $z \leq 1$ and $z \neq 0$, and define $f(0) = 0$. Then $|f(z)| \leq |z|$, for all $z \leq 1$.*

Proof. Note that f is continuous, with derivative

$$f'(z) = -\frac{1}{z^2} - \frac{z-2}{2z^2\sqrt{1-z}}.$$

It is straightforward to check that $f'(z) \geq 0$ for all $z < 1$, and we may define $f'(0) = \frac{1}{4}$ such that $f'(z)$ is continuous. Therefore, $f(z)$ is monotonic and maximized at $z = 1$, yielding $f(1) = \frac{1}{2}$, and minimized at $\lim_{z\to-\infty} f(z) = -\frac{1}{2}$.

We now split into cases. If $z < -\frac{1}{2}$, then $|f(z)| \leq \frac{1}{2} < |z|$. If $-1/2 \leq z \leq 1/2$, a Taylor expansion gives

$$\sqrt{1-z} = 1 - \frac{1}{2}z - \frac{1}{8}z^2 - \frac{1}{16}z^3 - \cdots - \frac{(n+1)!!}{2^n n!}z^n - \cdots .$$

Hence,

$$\begin{aligned} \left|\sqrt{1-z} - \left(1 - \frac{z}{2}\right)\right| &\leq \frac{1}{8}(|z|^2 + |z|^3 + \cdots) \\ &\leq \frac{1}{8}|z|^2(1 + |z| + |z|^2 + \cdots) \\ &\leq \frac{1}{8}|z|^2\frac{1}{1-|z|} \leq \frac{1}{4}|z|^2, \end{aligned}$$

implying that $|f(z)| \leq \frac{1}{4}|z|$. Finally, if $z > 1/2$, we have $|f(z)| \leq \frac{1}{2} < z$. □