OPTIMAL CLUSTERING ALGORITHMS IN BLOCK MARKOV CHAINS

By Jaron Sanders and Alexandre Proutière

KTH Royal Institute of Technology, Stockholm, Sweden

This paper considers cluster detection in Block Markov Chains (BMCs). These Markov chains are characterized by a block structure in their transition matrix. More precisely, the n possible states are divided into a finite number of K groups or clusters, such that states in the same cluster exhibit the same transition rates to other states. One observes a trajectory of the Markov chain, and the objective is to recover, from this observation only, the (initially unknown) clusters. In this paper we devise a clustering procedure that accurately, efficiently, and provably detects the clusters. We first derive a fundamental information-theoretical lower bound on the detection error rate satisfied under any clustering algorithm. This bound identifies the parameters of the BMC, and trajectory lengths, for which it is possible to accurately detect the clusters. We next develop two clustering algorithms that can together accurately recover the cluster structure from the shortest possible trajectories, whenever the parameters allow detection. These algorithms thus reach the fundamental detectability limit, and are optimal in that sense.

1. Introduction. The ability to accurately discover all hidden relations between items that share similarities is of paramount importance to a wide range of disciplines. Clustering algorithms in particular are employed throughout social sciences, biology, computer science, economics, and physics. The reason these techniques have become prevalent is that once clusters of similar items have been identified, any subsequent analysis or optimization procedure benefits from a powerful reduction in dimensionality.

The canonical Stochastic Block Model (SBM), originally introduced in [1], has become the benchmark to investigate the performance of cluster detection algorithms. This model generates random graphs that contain groups of similar vertices. Vertices within the same group are similar in that they share the same average edge densities to the other vertices. More precisely, if the set of n vertices $\mathcal{V}$ is for example partitioned into two groups $\mathcal{V}_1$ and $\mathcal{V}_2$, an edge is drawn between two vertices $x, y \in \mathcal{V}$ with probability $p \in (0,1)$ if they belong to the same group, and with probability $q \in (0,1)$, $p \neq q$, if they belong to different groups. Edges are drawn independently of all other edges. Within the context of the SBM and its generalizations, the problem of cluster detection is to infer the clusters from observations of a realization of the random graph with the aforementioned structure.

This paper generalizes the problem of cluster detection when the observation is the sample path of a Markov chain over the set of vertices. Specifically, we introduce the Block Markov Chain (BMC), which is a Markov chain characterized by a block structure in its transition matrix. States that are in the same cluster are similar in the sense that they have the same transition rates. The goal is to detect the clusters from an observed sample path $X_0, X_1, \ldots, X_T$ of the Markov chain (Figure 1). This extension is mathematically challenging because consecutive samples of the random walk are *not* independent: besides noise, there is bias in a sample path. Intuitively though there is hope for accurate cluster detection if the Markov chain can get close to stationarity within T steps. Indeed, as we will show, the mixing time [2] of the BMC plays a crucial role in the detectability of the clusters.

Clustering in BMCs is motivated by reinforcement learning problems [3] with large state spaces. These problems are concerned with the control of dynamical systems modeled as Markov

MSC 2010 subject classifications: Primary 62H30, 60J10; secondary 05C85

Keywords and phrases: cluster detection, Markov chains, asymptotic analysis, mixing times

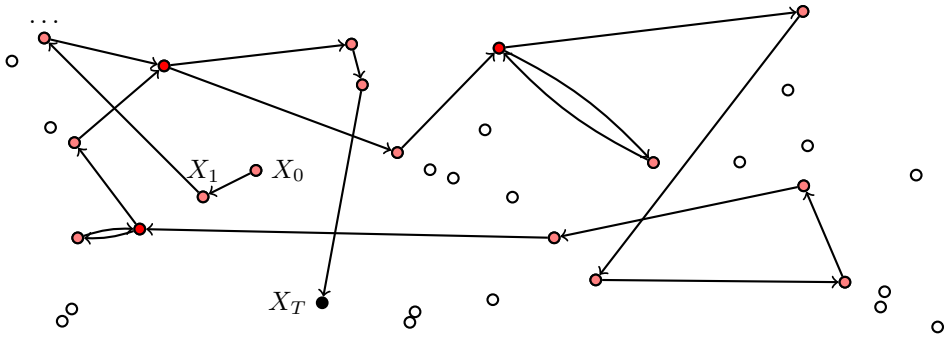


Fig 1: The goal of this paper is to infer the hidden cluster structure underlying a Markov chain $\{X_t\}_{t\geq 0}$, from one observation of a sample path $X_0, X_1, \ldots, X_T$ of length T.

chains whose transition kernels are initially unknown. The objective is to identify an optimal control policy as early as possible by observing the trajectory of a Markov chain generated under some known policy. The time it takes to learn efficient policies using standard algorithms such as Q-learning dramatically increases with the number of possible states, so that these algorithms become useless when the state space is prohibitively large. In most practical problems however, different states may yield similar reward and exhibit similar transition probabilities to other states, i.e., states can be grouped into clusters. In this scenario it becomes critical to learn and leverage this structure in order to speed up the learning process. In this paper we consider *uncontrolled* Markov chains, and we aim to identify clusters of states of quickly as possible. In the future we hope to extend the techniques developed here for an uncontrolled BMC to the more general case of controlled Markov chains, and hence to devise reinforcement learning algorithms that will efficiently exploit an underlying cluster structure.

This paper answers two major questions for the problem of cluster detection on BMCs. First, we derive a fundamental information-theoretical clustering error lower bound. The latter allows us to identify the parameters of the BMC and the sample path lengths T for which it is theoretically possible to accurately detect the underlying cluster structure. By accurately, we mean that the proportion of misclassified states vanishes as n grows large. Second, we develop two clustering algorithms that when combined, are able to accurately detect the underlying cluster structure from the shortest possible sample paths, whenever the parameters of the BMC allow detection, and that provably work as $n \to \infty$. These algorithms thus reach the fundamental detectability limit, and are optimal in that sense.

1.1. *Related work.* Significant advances have been made on cluster recovery within the context of the SBM and its generalizations. We defer the reader to [4] for an extensive overview. Substantial focus has in particular been on characterizing the set of parameters for which some recovery objectives can be met.

In the *sparse* regime, i.e., when the average degree of vertices is $O(1)$, necessary and sufficient conditions on the parameters have been identified under which it is possible to extract clusters that are positively correlated with the true clusters [5–7]. More precisely, for example if $p = \frac{a}{n}$ and $q = \frac{b}{n}$ and in the case of two clusters of equal sizes, it was conjectured in [5] that $a-b \geq \sqrt{2(a+b)}$ is a necessary and sufficient condition for the existence of algorithms that can *detect* the clusters (in the sense that they perform better than a random assignment of items to clusters). This result was established in [7] (necessary condition) and in [6] (sufficient condition).

In the *dense* regime, i.e., when the average degree is $\omega(1)$, it is possible to devise algorithms under which the proportion of misclassified vertices vanishes as the size of the graph grows large [8]. In this case, one may actually characterize the *minimal* asymptotic (as n grows large) classification error, and develop clustering algorithms achieving this fundamental limit [9]. We may further establish conditions under which asymptotic *exact* cluster recovery is possible [9–16].

This paper draws considerable inspiration from [8–10]. Over the course of these papers, the authors consider the problem of clustering in the Labeled Stochastic Block Model (LSBM), which is a generalization of the SBM. They identify the set of LSBM-parameters for which the clusters can be detected using change-of-measure arguments, and develop algorithms based on spectral methods that achieve this fundamental performance limit. Our contributions in this paper include the extension of the approaches to the context of Markov chains. This required us in particular to design novel changes-of-measure, carefully incorporate the effect of mixing, deal with new and non-convex log-likelihood functions, and widen the applicability of spectral methods to random matrices with bias.

1.2. *Methodology.* Similar to the extensive efforts for the SBM, we will identify parameters of the BMC for which it is theoretically possible to detect the clusters. We will furthermore provide a clustering algorithm that achieves this fundamental limit. The key difference between clustering in SBMs and clustering in BMCs is that instead of observing (the edges of) a random graph, we now try to infer the cluster structure from an as short as possible sample path. The sample path will be inherently noisy and biased by construction, and this necessitates a careful analysis of the mixing time of the Markov chain [2]. The mixing time is a measure for how close the Markov chain is to stationarity, intuitively a prerequisite for successful clustering. By analyzing the mixing time, we are able to use powerful concentration inequalities [17] that can deal with the bias inherent in the sample path. One difficulty we encounter is that a BMC is not necessarily reversible, forcing us to employ a nonstandard mixing time bound [18].

Our analysis consists of two parts. In the first part we use techniques from information theory to derive a lower bound on the number of misclassified states that holds for any classification algorithm. This relies on a powerful change-of-measure argument, originally explored in [19] in the context of online stochastic optimization. First, we relate the probability of misclassifying a state in the BMC to a log-likelihood ratio that the sample path was generated by a perturbed Markov chain instead. Then, given any BMC, we show how to construct a perturbed Markov chain that assigns a nonzero probability to the event that all clustering algorithms misclassify at least one particular state. Finally, we maximize over all possible perturbations to get the best possible lower bound that holds for any algorithm. The second part consists of developing a clustering algorithm that can asymptotically detect the clusters. Specifically, we develop a two-step procedure. The first step consists in applying a classical Spectral Clustering Algorithm. This algorithm essentially creates a rank-K approximation of a random matrix corresponding to the empirical transition rates between any pair of states, and then uses a K-means algorithm [20] to cluster all states. We show that this first step clusters the majority of states roughly correctly. Next, we introduce the *Cluster Improvement Algorithm*. This algorithm uses the rough structure learned from the Spectral Clustering Algorithm, together with the sample path, to move each individual state into the cluster the state most likely belongs to. This is achieved through a recursive, local maximization of a log-likelihood ratio.

In our derivations, we encounter two challenging issues. The noise and bias within the sample path have to first be related to the spectrum of the random matrix, for which we use techniques of [21]. Then, because the entries of this random matrix are not independent, it is hard to quantify the concentration of its eigenvalues. While concentration of the eigenvalues of a random matrix has been actively investigated when the entries are independent or satisfy a weak condition of dependence [22–27], or when the transition matrix of the Markov chain itself is random [28, 29], we have been unable to find work related to the case when the entries are dictated by a Markov chain with a fixed transition matrix with a block structure. That is why we choose, for now, to settle for asymptotic results. We believe that studying the concentration of noisy, *biased* random matrices is an important open problem, refer to Section 8 for a detailed discussion.

1.3. *Overview.* This paper is structured as follows. We introduce the BMC in Section 2. Section 3 provides an overview of our results and our algorithms. We subsequently prove our

results by first deriving an information lower bound and developing an optimal change-of-measure in Section 4, and then by developing the Spectral Clustering Algorithm in Section 5 and the Cluster Improvement Algorithm in Section 6. We assess the performance of both algorithms, i.e., we quantify their asymptotic error rates. Section 7 discusses several numerical experiments designed to test the algorithms. Section 8 discusses an open problem of spectral norm concentration of a biased random matrix.

Notation. For any two sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$ we define their symmetric difference by $\mathcal{A} \triangle \mathcal{B} = \{\mathcal{A} \backslash \mathcal{B}\} \cup \{\mathcal{B} \backslash \mathcal{A}\}$. For any two numbers $a, b \in \mathbb{R}$ we introduce the shorthand notations $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For any n-dimensional vector $x = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$, we define its l_p norms by

$$\|x\|_p = \Big(\sum_{r=1}^{n} |x_i|^p \Big)^{1/p}$$

The n-dimensional unit vector of which the r-th component equals 1 will be denoted by $e_{n,r}$, and the n-dimensional vector for which all elements $r \in \mathcal{A} \subseteq \{1, \ldots, n\}$ equal 1 will be denoted by $1_{\mathcal{A}}$. For any $m \times n$ matrix $A \in \mathbb{R}^{m \times n}$, we indicate its rows by $A_{r,\cdot}$ for $r = 1, \ldots, m$ and its columns by $A_{\cdot,c}$ for $c = 1, \ldots, n$. We also introduce the short-hand notation $A_{\mathcal{A},\mathcal{B}} = \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} A_{x,y}$ for all subsets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$. Its Frobenius norm and spectral norm are defined by

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{r=1}^{m} \sum_{c=1}^{n} A_{r,c}^2}, \quad \|A\| = \sup_{b \in \mathbb{S}^{n-1}} \{\|Ab\|_2\},$$

respectively. Here, $\mathbb{S}^{n-1} = \{x = (x_1, \ldots, x_n) \in (0,1)^n : \|x\|_2 = 1\}$ denotes the n-dimensional unit sphere. We define the probability simplex of dimension $n-1$ by $\Delta^{n-1} = \{x \in (0,1)^n : \|x\|_1 = 1\}$ as well as the set of left stochastic matrices $\mathbb{\Delta}^{(n-1)\times n} = \{((x_{1,1}, \ldots, x_{1,n}), \ldots, (x_{n,1}, \ldots, x_{n,n})) \in [0,1]^{n \times n} : \sum_{c=1}^{n} x_{r,c} = 1 \text{ for } r = 1, \ldots, n\}$ similarly.

In our asymptotic analyses, we write $f(n) \sim g(n)$ if $\lim_{n\to\infty} f(n)/g(n) = 1$, $f(n) = o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$ and $f(n) = O(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) < \infty$. Whenever $\{X_n\}_{n=1}^{\infty}$ is a sequence of real-valued random variables and $\{a_n\}_{n=1}^{\infty}$ a deterministic sequence, we write

$$X_n = o_{\mathbb{P}}(a_n) \Leftrightarrow \lim_{n\to\infty} \mathbb{P}\Big[\Big|\frac{X_n}{a_n}\Big| \geq \delta\Big] = 0 \, \forall_{\delta > 0} \Leftrightarrow \forall_{\varepsilon,\delta} \exists_{N_{\varepsilon,\delta}} : \mathbb{P}\Big[\Big|\frac{X_n}{a_n}\Big| \geq \delta\Big] \leq \varepsilon \, \forall_{n > N_{\varepsilon,\delta}},$$
$$\text{and} \quad X_n = O_{\mathbb{P}}(a_n) \Leftrightarrow \forall_{\varepsilon} \exists_{\delta_\varepsilon, N_\varepsilon} : \mathbb{P}\Big[\Big|\frac{X_n}{a_n}\Big| \geq \delta_\varepsilon\Big] \leq \varepsilon \, \forall_{n > N_\varepsilon}.$$

Similarly, $X_n = \Omega_{\mathbb{P}}(a_n)$ denotes $\forall_\varepsilon \exists_{\delta_\varepsilon, N_\varepsilon} : \mathbb{P}[|X_n/a_n| \leq \delta_\varepsilon] \leq \varepsilon \, \forall_{n > N_\varepsilon}$, and $X_n \asymp_{\mathbb{P}} (a_n)$ means $\forall_\varepsilon \exists_{\delta_\varepsilon^-, \delta_\varepsilon^+, N_\varepsilon} : \mathbb{P}[\delta_\varepsilon^- \leq |X_n/a_n| \leq \delta_\varepsilon^+] \geq 1 - \varepsilon \, \forall_{n > N_\varepsilon}$.

2. Block Markov Chains (BMCs). We assume that we have n states $\mathcal{V} = \{1, \ldots, n\}$, each of which is associated to one of K clusters. This means that the set of states is partitioned so that $\mathcal{V} = \cup_{k=1}^{K} \mathcal{V}_k$ with $\mathcal{V}_k \cap \mathcal{V}_l = \emptyset$ for all $k \neq l$. Let $\sigma(v)$ denote the cluster of a state $v \in \mathcal{V}$. We also assume that there exist constants $\alpha \in \Delta^{K-1}$ so that $\lim_{n\to\infty} |\mathcal{V}_k|/(n\alpha_k) = 1$.

For any $\alpha \in \Delta^{K-1}$ and $p \in \Delta^{(K-1)\times K}$, we define the BMC $\{X_t\}_{t \geq 0}$ as follows. Its transition matrix $P \in \mathbb{\Delta}^{(n-1)\times n}$ will be defined as

$$P_{x,y} \triangleq \frac{p_{\sigma(x),\sigma(y)}}{|\mathcal{V}_{\sigma(y)}| - \mathbb{1}[\sigma(x) = \sigma(y)]} \mathbb{1}[x \neq y] \quad \text{for all} \quad x, y \in \mathcal{V}. \tag{1}$$

Note that this Markov chain is not necessarily reversible.

2.1. *Equilibrium behavior.* We assume that the stochastic matrix p is such that the equilibrium distribution of $\{X_t\}_{t\geq 0}$ exists, and we will denote it by Π_x for $x \in \mathcal{V}$. By symmetry, $\Pi_x = \Pi_y \triangleq \bar{\Pi}_k$ for any two states $x, y \in \mathcal{V}_k$ for all $k = 1, \dots, K$. Consider the scaled quantity

$$\pi_k \triangleq \lim_{n\to\infty} \sum_{x\in\mathcal{V}_k} \Pi_x = \lim_{n\to\infty} |\mathcal{V}_k| \bar{\Pi}_k \quad \text{for} \quad k = 1, \dots, K.$$

PROPOSITION 1. *The quantity π solves $\pi^{\mathrm{T}} p = \pi^{\mathrm{T}}$, and is therefore the equilibrium distribution of a Markov chain with transition matrix p and state space $\Omega = \{1, \dots, K\}$.*

PROOF. We first prove that π is a probability distribution. This follows by (i) definition of π, (ii) symmetry of all states in the same cluster, and (iii) because Π is a probability distribution:

$$\sum_{k=1}^{K} \pi_k \overset{\text{(i)}}{=} \sum_{k=1}^{K} \lim_{n\to\infty} \bar{\Pi}_k |\mathcal{V}_k| \overset{\text{(ii)}}{=} \lim_{n\to\infty} \sum_{k=1}^{K} \sum_{x\in\mathcal{V}_k} \Pi_x = \lim_{n\to\infty} \sum_{x\in\mathcal{V}} \Pi_x \overset{\text{(iii)}}{=} 1.$$

Next, we show that the balance equations hold. For $k = 1, \dots, K$ it follows by symmetry of any two states $x, z \in \mathcal{V}_k$ that $\Pi_x = \Pi_z = \bar{\Pi}_k$. Hence for any $y \in \mathcal{V}_l$, by (iv) global balance

$$\Pi_y = \bar{\Pi}_l \overset{\text{(iv)}}{=} \sum_{k=1}^{K} \sum_{x\in\mathcal{V}_k} \Pi_x P_{x,y} = \sum_{k=1}^{K} \bar{\Pi}_k (|\mathcal{V}_k| - \mathbb{1}[k = l]) \frac{p_{k,l}}{|\mathcal{V}_l| - \mathbb{1}[k = l]}.$$

Letting $n \to \infty$, we find that $\pi_l = \sum_{k=1}^{K} \pi_k p_{k,l}$ for all k, l. This completes the proof. □

2.2. *Mixing time.* Proposition 2 gives a bound on the mixing time $t_{\mathrm{mix}} \in (0, \infty)$, which is defined by

$$d(t) \triangleq \max_{x\in\mathcal{V}} \{d_2(P^t_{x,\cdot}, \Pi)\} \quad \text{and} \quad t_{\mathrm{mix}}(\varepsilon) \triangleq \min\{t \geq 0 : d(t) \leq \varepsilon\}, \tag{2}$$

where

$$d_p(\mu, \nu) \triangleq \Big(\sum_{x\in\mathcal{V}} \Big| \frac{\mu_x}{\Pi_x} - \frac{\nu_x}{\Pi_x} \Big|^p \Pi_x \Big)^{1/p} \quad \text{for} \quad p \in [1, \infty). \tag{3}$$

PROPOSITION 2. *There exists a strictly positive constant $c_{\mathrm{mix}} > 0$ independent of n such that $t_{\mathrm{mix}}(\varepsilon) \leq -c_{\mathrm{mix}} \ln \varepsilon$.*

Proposition 2 implies that the mixing times are short enough so that our results will hold *irrespective* of whether we assume that the Markov chain is initially in equilibrium. We will show in Section 4.4 that what is important is that the chain reaches stationarity *within* T steps (the length of the observed trajectory), and consequentially, T needs to be chosen sufficiently large with respect to n to ensure that this occurs. Throughout this paper we therefore assume for simplicity that the chain is started from equilibrium. This eliminates the need of tracking higher order correction terms. For the proof of Proposition 2, see Appendix C.

Examples. Figure 2 illustrates the structure of a BMC when there are $K = 2$ groups. We find after solving the balance equations that the limiting equilibrium behavior is given by $\pi_1 = p_{21}/(p_{12} + p_{21})$ and $\pi_2 = p_{12}/(p_{12} + p_{21})$.

For $K = 3$, we find after solving the balance equations that the limiting equilibrium behavior is given by

$$\pi_1 = \frac{p_{23}p_{31} + p_{21}(p_{31} + p_{32})}{Z(p)}, \quad \pi_2 = \frac{p_{13}p_{32} + p_{12}(p_{31} + p_{32})}{Z(p)}, \quad \pi_3 = \frac{p_{12}p_{23} + p_{13}(p_{21} + p_{23})}{Z(p)},$$

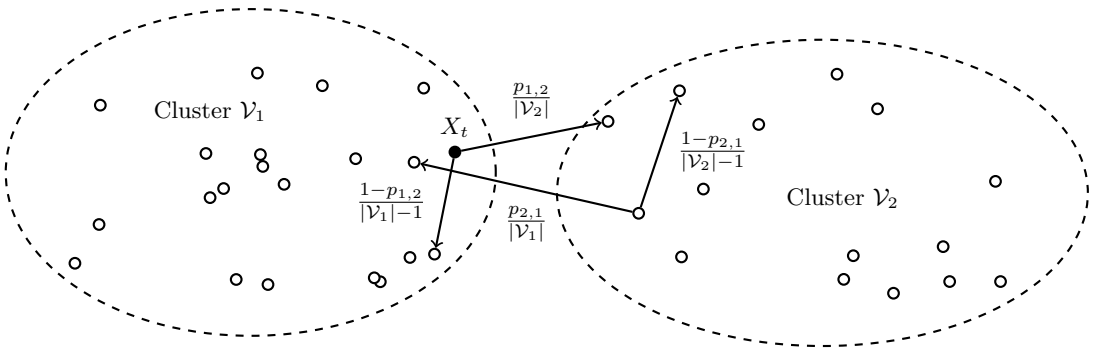


Fig 2: In the BMC with $K = 2$ groups $\mathcal{V}_1 \cup \mathcal{V}_2 = \mathcal{V}$, whenever the Markov chain is at some state $X_t \in \mathcal{V}_1$, it will next jump with probability $p_{1,2}$ to cluster $\mathcal{V}_2$, and with probability $1 - p_{1,2}$ to some other state in cluster $\mathcal{V}_1$. Similarly, if $X_t \in \mathcal{V}_2$, it would next jump to cluster $\mathcal{V}_1$ with probability $p_{2,1}$, or stay within its own cluster with probability $1 - p_{2,1}$.

with $Z(p) = (p_{21} + p_{23})(p_{13} + p_{31}) + (p_{13} + p_{21})p_{32} + p_{12}(p_{23} + p_{31} + p_{32})$. Let us also illustrate the structure of the transition matrix when $\alpha = (2/10, 3/10, 5/10)$ and $n = 10$:

$$P = \begin{pmatrix}
0 & p_{1,1} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} \\
p_{1,1} & 0 & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & 0 & \frac{p_{2,2}}{2} & \frac{p_{2,2}}{2} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & \frac{p_{2,2}}{2} & 0 & \frac{p_{2,2}}{2} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & \frac{p_{2,2}}{2} & \frac{p_{2,2}}{2} & 0 & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & 0 & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & 0 & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & 0 & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & 0 & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & 0
\end{pmatrix} \tag{4}$$

3. Summary of our results. In this paper we obtain quantitative statements on the set of misclassified states,

$$\mathcal{E} \triangleq \min_{\text{all permutations } \gamma} \bigcup_{k=1}^{K} \hat{\mathcal{V}}_{\gamma(k)} \backslash \mathcal{V}_k.$$

Here, the sets $\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_K$ will always denote an approximate cluster assignment obtained from some clustering algorithm. For notational convenience we will always number the approximate clusters so as to minimize the number of misclassifications, allowing us to forego defining it formally via a permutation. We will now also note that in this paper, we restrict the analysis to the case that the number of clusters K is known. This reduces the complexity of the analysis. Based on the findings in [8–10] however, we are confident that this assumption can be relaxed in future work.

Let us now summarize our results: for the precise statements, see Sections 4–6. The results identify an important quantity $I(\alpha, p)$ that measures how difficult it is to cluster in a BMC.

DEFINITION. *For $\alpha \in \Delta^{K-1}$ and $p \in \mathbb{\Delta}^{(K-1)\times K}$, let*

$$I(\alpha, p) \triangleq \min_{a \neq b} \Big\{ \sum_{k=1}^{K} \frac{1}{\alpha_a} \Big(\pi_a p_{a,k} \ln \frac{p_{a,k}}{p_{b,k}} + \pi_k p_{k,a} \ln \frac{p_{k,a}\alpha_b}{p_{k,b}\alpha_a} \Big) + \Big(\frac{\pi_b}{\alpha_b} - \frac{\pi_a}{\alpha_a} \Big) \Big\}. \tag{5}$$

Here π denotes the solution to $\pi^{\mathrm{T}} p = \pi^{\mathrm{T}}$.

RESULT (Theorem 1). *If $I(\alpha,p) > 0$, then there exist strictly positive, finite constants $C, J(\alpha,p)$ independent of n such that*

$$\mathbb{E}_P\Big[\frac{|\mathcal{E}|}{n}\Big] \geq C\exp\Big(-J(\alpha,p)\frac{T}{n} + o\Big(\frac{T}{n}\Big)\Big)$$

for any clustering algorithm. As a consequence, a necessary condition for the existence of a clustering algorithm misclassifying a vanishing proportion of states in average, i.e., such that $\mathbb{E}_P[\frac{|\mathcal{E}|}{n}] = o(1)$, is $T = \omega(n)$. Similarly, a necessary condition for the existence of an asymptotically exact clustering algorithm, i.e., such that $\mathbb{E}_P[|\mathcal{E}|] = o(1)$, is $T = \omega(n\ln n)$.

The proof of Theorem 1 relies on a change-of-measure argument, and on relating the probability of misclassifying a state in the BMC to a log-likelihood ratio that the sample path was generated by a perturbed Markov chain instead. The key challenge is to construct the appropriate change-of-measures and perturbed Markov chains, which is discussed in detail in Section 4.

RESULT (Theorems 2, 3). *If $I(\alpha,p) > 0$, if $\exists_{0<\eta\neq 1} : \max_{a,b,c=1,\dots,K}\{p_{b,a}/p_{c,a}, p_{a,b}/p_{a,c}\} \leq \eta$, $\|\hat{N} - N\| = O_{\mathbb{P}}(f(n,T))$ for some $f(n,T) = o(T/n)$, and if $\|\hat{P} - P\| = O_{\mathbb{P}}(g(n,T))$ for some $g(n,T) = o(1)$, then there exists a clustering algorithm that misclassifies $|\mathcal{E}| = o_{\mathbb{P}}(1)$ states.*

Our proofs of Theorems 2, 3 are constructive, in that we create actual clustering procedures and perform asymptotic analyses of their performances. As we explained in the introduction, we have developed a two-step clustering procedure. First, our Spectral Clustering Algorithm roughly clusters most states accurately (Algorithm 1). Next, this approximate assignment is recursively used to obtain improved cluster assignments using our Cluster Improvement Algorithm (Algorithm 2). Their asymptotic behaviors are analyzed in Sections 5–6, and simulation results of their performance can be found in Section 7.

Input: A trajectory $X_0, X_1, \dots, X_T$
Output: An approximate cluster assignment $\hat{\mathcal{V}}_1^{[0]}, \dots, \hat{\mathcal{V}}_K^{[0]}$, and matrices $\hat{P}$, $\hat{N}$
begin
  for $x \leftarrow 1$ **to** n **do**
    for $y \leftarrow 1$ **to** n **do**
      $\hat{N}_{x,y} \leftarrow \sum_{t=0}^{T-1} \mathbb{1}[X_t = x, X_{t+1} = y]$;
      $\hat{P}_{x,y} \leftarrow \big(\sum_{t=0}^{T-1} \mathbb{1}[X_t = x, X_{t+1} = y]\big)/\big(\sum_{t=0}^{T-1} \mathbb{1}[X_t = x]\big)$;
    end
  end
  Calculate the Singular Value Decomposition (SVD) $U\Sigma V^{\mathrm{T}}$ of either $\hat{P}$ or $\hat{N}$;
  Order U, Σ, V s.t. the singular values $\sigma_1 \geq \sigma \geq \dots \geq \sigma_n \geq 0$ are in descending order;
  Construct the rank-K approximation $\hat{R} = \sum_{k=1}^{K} \sigma_k U_{\cdot,k} V_{\cdot,k}{}^{\mathrm{T}}$;
  Apply a K-means algorithm to $\hat{R}$ to determine $\hat{\mathcal{V}}_1^{[0]}, \dots, \hat{\mathcal{V}}_K^{[0]}$;
end

Algorithm 1: Pseudo-code for the Spectral Clustering Algorithm.

4. The information bound and the change of measure. In this section, we prove Theorem 1. As a consequence of this result, a necessary condition for the existence of a clustering algorithm that misclassifies $\mathbb{E}_P[|\mathcal{E}|] = o(s)$ vertices is $T = \omega(n\ln(n/s))$.

THEOREM 1. *If $I(\alpha,p) > 0$, then there exist strictly positive and finite constant C independent of n such that: for any clustering algorithm*

$$\mathbb{E}_P\Big[\frac{|\mathcal{E}|}{n}\Big] \geq C\exp\Big(-J(\alpha,p)\frac{T}{n} + o\Big(\frac{T}{n}\Big)\Big),$$

Input: An approximate cluster assignment $\hat{\mathcal{V}}_1^{[t]}, \dots, \hat{\mathcal{V}}_K^{[t]}$, and matrices $\hat{P}$, $\hat{N}$
Output: A revised assignment $\hat{\mathcal{V}}_1^{[t+1]}, \dots, \hat{\mathcal{V}}_K^{[t+1]}$

begin
  for $a \leftarrow 1$ **to** K **do**
    $\hat{\pi}_a \leftarrow \hat{N}_{\hat{\mathcal{V}}_a^{[t]},\mathcal{V}}/T$, $\alpha_a \leftarrow |\hat{\mathcal{V}}_a^{[t]}|/n$, $\hat{\mathcal{V}}_a^{[t+1]} \leftarrow \emptyset$;
    for $b \leftarrow 1$ **to** K **do**
      $\hat{p}_{a,b} \leftarrow (|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[a=b])\hat{P}_{\hat{\mathcal{V}}_a^{[t]},\hat{\mathcal{V}}_b^{[t]}}/(|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|)$;
    end
  end
  for $x \leftarrow 1$ **to** n **do**
    $c_x^{\mathrm{opt}} \leftarrow \arg\max_{c=1,\dots,K}\Big\{\sum_{k=1}^K \big(\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} \ln \hat{p}_{c,k} + \hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} \ln \frac{\hat{p}_{k,c}}{\hat{\alpha}_c}\big) - \frac{T}{n}\cdot\frac{\hat{\pi}_c}{\hat{\alpha}_c}\Big\}$;
    $\mathcal{V}_{c_x^{\mathrm{opt}}}^{[t+1]} \leftarrow \mathcal{V}_{c_x^{\mathrm{opt}}}^{[t+1]} \cup \{x\}$;
  end
end

Algorithm 2: Pseudo-code for the Cluster Improvement Algorithm.

where

$$0 < J(\alpha, p) \triangleq \min_{k\neq l} \min_{q\in\mathcal{Q}(k,l)} \Big(\frac{\alpha_k}{\alpha_k+\alpha_l} I_k(q||p) + \frac{\alpha_l}{\alpha_k+\alpha_l} I_l(q||p)\Big) < \infty.$$

Here

$$I_c(q||p) \triangleq \sum_{k=1}^K \Big((\sum_{l=1}^K \pi_l q_{l,0}) q_{0,k} \ln \frac{q_{0,k}}{p_{c,k}} + \pi_k q_{k,0} \ln \frac{q_{k,0}\alpha_c}{p_{k,c}}\Big) + \Big(\frac{\pi_c}{\alpha_c} - \sum_{k=1}^K \pi_k q_{k,0}\Big) \tag{6}$$

for $c = 1, \dots, K$, *and*

$$\mathcal{Q}(k,l) \triangleq \big\{q \in \mathcal{Q} \big| I_k(q||p) = I_l(q||p)\big\} \neq \emptyset \quad \textit{for all} \quad k \neq l,$$
$$\mathcal{Q} \triangleq \big\{(q_{k,0}, q_{0,k})_{k=0,\dots,K} \in (0,\infty) \big| q_{0,0} = 0, \sum_{l=1}^K q_{0,l} = 1\big\}. \tag{7}$$

4.1. *Change-of-measure argument.* We now proceed with the change-of-measure argument. It consists in considering that the observations $X_0, \dots, X_T$ are generated by a slightly different stochastic model than the true model defined by the clusters and the transition matrix P. We denote by Φ (resp. by Ψ) the true (resp. modified) model, and by $\mathbb{P}_\Phi$ (resp. by $\mathbb{P}_\Psi$) the probability measure corresponding to Φ (resp. Ψ). The modified model is obtained by randomly choosing a state V^* (this choice will be made precise later on) and by constructing a transition matrix Q depending on V^* that is slightly different from P.

Given a sample path $X_0, X_1, \dots, X_T \in \mathcal{V}$, the argument revolves around the quantity

$$L \triangleq \ln \frac{\mathbb{P}_Q[X_0, X_1, \dots, X_T]}{\mathbb{P}_P[X_0, X_1, \dots, X_T]} \tag{8}$$

which resembles a log-likelihood ratio. Here,

$$\mathbb{P}_P[X_0, X_1, \dots, X_T] = \prod_{t=1}^T P_{X_{t-1},X_t}, \quad \text{and} \quad \mathbb{P}_Q[X_0, X_1, \dots, X_T] = \prod_{t=1}^T Q_{X_{t-1},X_t}$$

such that $L = \sum_{t=1}^T \ln\left(Q_{X_{t-1},X_t}/P_{X_{t-1},X_t}\right)$. Note that L is random because it depends on the observations, but also on V^*. We now prove the following information bound.

PROPOSITION 3. *If V^* is chosen uniformly at random from two clusters $a \neq b$, and if*

$$\exists \text{ an absolute constant } \delta \text{ s.t. } \quad \mathbb{P}_\Psi[V^* \in \mathcal{E}] \geq \delta > 0 \quad \text{for any classification algorithm,} \tag{9}$$

then there exists a strictly positive constant $C > 0$ independent of n such that

$$\frac{\mathbb{E}_\Phi[|\mathcal{E}|]}{n} \geq C \exp\left(-\mathbb{E}_\Psi[L] - \sqrt{\frac{2}{\delta}}\sqrt{\mathrm{Var}_\Psi[L]}\right) \tag{10}$$

for any clustering algorithm.

PROOF. Select a state V^* uniformly at random from any two specific clusters $a, b \in \{1, \ldots, K\}$, $a \neq b$. We are going to bound

$$\mathbb{P}_\Psi[L \leq f(n,T)] = \mathbb{P}_\Psi[L \leq f(n,T), V^* \in \mathcal{E}] + \mathbb{P}_\Psi[L \leq f(n,T), V^* \notin \mathcal{E}]. \tag{11}$$

for any function $f : \mathbb{N}_+^2 \to \mathbb{R}$.

The first term of (11) can be bounded using our change of measure formula (8). Namely,

$$\mathbb{P}_\Psi[L \leq f(n,T), V^* \in \mathcal{E}] \overset{(8)}{\leq} \mathrm{e}^{f(n,T)}\mathbb{P}_\Phi[L \leq f(n,T), V^* \in \mathcal{E}] \leq \mathrm{e}^{f(n,T)}\mathbb{P}_\Phi[V^* \in \mathcal{E}]. \tag{12}$$

Because V^* is selected from $\mathcal{V}_a \cup \mathcal{V}_b$ uniformly at random, we have by Lemma 13, see Appendix A, that for any V selected uniformly at random from *all* vertices $\mathcal{V}$,

$$\mathbb{P}_\Phi[V^* \in \mathcal{E}] = \mathbb{P}_\Phi[V \in \mathcal{E} | V \in \mathcal{V}_a \cup \mathcal{V}_b] = \frac{\mathbb{P}_\Phi[V \in \mathcal{E}, V \in \mathcal{V}_a \cup \mathcal{V}_b]}{\mathbb{P}_\Phi[V \in \mathcal{V}_a \cup \mathcal{V}_b]} \leq \frac{\mathbb{P}_\Phi[V \in \mathcal{E}]}{\alpha_a + \alpha_b}.$$

Subsequently by Lemma 14, see Appendix A,

$$\mathbb{P}_\Phi[V^* \in \mathcal{E}] \leq \frac{\mathbb{E}_\Phi[|\mathcal{E}|]}{(\alpha_a + \alpha_b)n}. \tag{13}$$

Substituting (13) into (12), we obtain

$$\mathbb{P}_\Psi[L \leq f(n,T), V^* \in \mathcal{E}] \leq \mathrm{e}^{f(n,T)} \frac{\mathbb{E}_\Phi[|\mathcal{E}|]}{(\alpha_a + \alpha_b)n}. \tag{14}$$

The second term of (11) can be bounded using Assumption (9):

$$\mathbb{P}_\Psi[L \leq f(n,T), V^* \notin \mathcal{E}] \leq \mathbb{P}_\Psi[V^* \notin \mathcal{E}] = 1 - \mathbb{P}_\Psi[V^* \in \mathcal{E}] \leq 1 - \delta < 1. \tag{15}$$

Now using (14) and (15) to bound (11), we arrive at

$$\mathbb{P}_\Psi[L \leq f(n,T)] \leq \mathrm{e}^{f(n,T)} \frac{\mathbb{E}_\Phi[|\mathcal{E}|]}{(\alpha_a + \alpha_b)n} + 1 - \delta \tag{16}$$

with $\delta > 0$. This strict separation will become important in a moment.

We now prepare for an application of Chebyshev's inequality. First note using (16) that

$$\mathbb{P}_\Psi[L \geq f(n,T)] = 1 - \mathbb{P}_\Psi[L \leq f(n,T)] \geq \delta - \mathrm{e}^{f(n,T)} \frac{\mathbb{E}_\Phi[|\mathcal{E}|]}{(\alpha_a + \alpha_b)n}.$$

Specify $f(n,T) = \ln(\delta/2) + \ln\left((\alpha_a + \alpha_b)n / \mathbb{E}_\Phi[|\mathcal{E}|]\right)$, so that

$$\mathbb{P}_\Psi\Big[L \geq \ln\frac{\delta}{2} + \ln\frac{(\alpha_a + \alpha_b)n}{\mathbb{E}_\Phi[|\mathcal{E}|]}\Big] \geq \frac{\delta}{2}.$$

Since $\delta > 0$, we can apply (i) Chebyshev's inequality to conclude

$$\mathbb{P}_\Psi\Big[L \geq \mathbb{E}_\Psi[L] + \sqrt{\frac{2}{\delta}}\sqrt{\mathrm{Var}_\Psi[L]}\Big] \overset{\text{(i)}}{\leq} \frac{\delta}{2} \overset{(4.1)}{\leq} \mathbb{P}_\Psi\Big[L \geq \ln\frac{\delta}{2} + \ln\frac{(\alpha_a+\alpha_b)n}{\mathbb{E}_\Phi[|\mathcal{E}|]}\Big].$$

Comparing the events in the left member and the right member, we then must have

$$\ln\frac{\delta}{2} + \ln\frac{(\alpha_a+\alpha_b)n}{\mathbb{E}_\Phi[|\mathcal{E}|]} \leq \mathbb{E}_\Psi[L] + \sqrt{\frac{2}{\delta}}\sqrt{\mathrm{Var}_\Psi[L]}.$$

Rearranging gives (10) with $C = (\alpha_a+\alpha_b)\delta/2 > 0$. This completes the proof. □

In order to further lower bound $\mathbb{E}_\Phi[|\mathcal{E}|]$, we proceed by constructing a change of measure that satisfies condition (9), and we then calculate the leading order behavior of $\mathbb{E}_\Psi[L|\sigma(V^*)]$ and upper bound $\mathrm{Var}_\Psi[L|\sigma(V^*)]$.

4.2. *The perturbed BMC given V^* and $q \in \mathcal{Q}$.* We now construct a transition matrix Q that resembles P, but differs in that V^* is placed in its own cluster with its own specific intra-cluster jump rates. We will label this extra cluster by 0 to indicate its special status. Asymptotically Q will resemble P as we are going to move probability mass away from and towards the other entries. While most entries will individually be perturbed slightly only, we still need to carefully analyze their collective contribution as this will be significant.

Define

$$q_{k,l} \triangleq p_{k,l} - \frac{q_{k,0}}{Kn} \quad \text{for} \quad k,l = 1,\dots,K, \tag{17}$$

and assume that $n > \lceil \max_{k,l=1,\dots,K}\{q_{k,0}/(Kp_{k,l})\}\rceil$ so that the entries of (17) are strictly positive. This assumption is not restrictive because the right-hand side is independent of n and finite. Note that the collection $\{q_{k,l}\}_{k,l\in\{0,1,\dots,K\}}$ does *not* constitute a stochastic matrix, but does resemble the transition matrix p for sufficiently large n. Now define component-wise

$$Q_{x,y} \triangleq \frac{q_{\omega(x),\omega(y)}\mathbb{1}[x \neq y]}{|\mathcal{W}_{\omega(y)}| - \mathbb{1}[\omega(x) = \omega(y)]}, \quad Q_{x,V^*} \triangleq \frac{q_{\omega(x),0}}{n} \quad \text{for} \quad x \in \mathcal{V}, y \neq V^*, \tag{18}$$

where

$$\omega(x) \triangleq \begin{cases} 0 & \text{if } x = V^*, \\ \sigma(x) & \text{if } x \neq V^*, \end{cases} \quad \text{and} \quad \mathcal{W}_k \triangleq \begin{cases} \{V^*\} & \text{if } k = 0, \\ \mathcal{V}_k\backslash\{V^*\} & \text{if } k = 1,\dots,K, \end{cases}$$

for notational convenience. This has the added benefit of giving (18) a similar form as (1).

Q is by construction a stochastic matrix (see Appendix D). Note furthermore that because Q is constructed from P, which by assumption describes an irreducible Markov chain, and because the entries $(q_{k,0}, q_{0,k})^{k=1,\dots,K}$ are all strictly positive, Q also describes an irreducible Markov chain.

4.2.1. *Asymptotic behavior of the equilibrium distribution.* Let $\Pi^{(Q)}$ denote the equilibrium distribution of a Markov chain with transition matrix Q, i.e., the solution to $\Pi^{(Q)\mathrm{T}}Q = \Pi^{(Q)\mathrm{T}}$. By symmetry of states in the same cluster $\Pi^{(Q)}_x = \Pi^{(Q)}_y \triangleq \bar{\Pi}^{(Q)}_k$ for any two states $x, y \in \mathcal{W}_k$ and all $k \in \{1,\dots,K,0\}$. Define

$$\gamma^{[0]}_k \triangleq \lim_{n\to\infty}\sum_{x\in\mathcal{W}_k}\Pi^{(Q)}_x = \lim_{n\to\infty}|\mathcal{W}_k|\bar{\Pi}^{(Q)}_k \quad \text{for} \quad k \in \{0,1,\dots,K\}. \tag{19}$$

We can expect $\gamma^{[0]}_0$ to be zero, because by our construction of Q we can expect that $\Pi^{(Q)}_x = O(1/n)$ for all $x \in \mathcal{V}$ (including V^*). We therefore also define its higher order statistic

$$\gamma^{[1]}_0 \triangleq \lim_{n\to\infty} n\Pi^{(Q)}_{V^*}.$$

The following proposition relates these scaled quantities to the parameters of our BMC $\{X_t\}_{t\geq 0}$. The proof is deferred to Appendix E.

PROPOSITION 4. *For $k = 1, \dots, K$, $\gamma_k^{[0]} = \pi_k$. Furthermore, $\gamma_0^{[0]} = 0$ and $\gamma_0^{[1]} = \sum_{k=1}^K \pi_k q_{k,0}$.*

4.2.2. *Mixing time.* We now crucially note that Proposition 2 holds for a Markov chain with Q as its transition matrix as well. This follows when applying the exact same proof.

Example. It is illustrative to explicitly write down at least one example kernel Q. For $K = 3$, $\alpha = (2/10, 3/10, 5/10)$ and $n = 10$, $V^* = 7$, it is given by

$$Q = \begin{pmatrix}
0 & p_{1,1} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,3}}{4} & \frac{q_{1,0}}{10} & \frac{p_{1,3}}{4} & \frac{p_{1,3}}{4} & \frac{p_{1,3}}{4} \\
p_{1,1} & 0 & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,3}}{4} & \frac{q_{1,0}}{10} & \frac{p_{1,3}}{4} & \frac{p_{1,3}}{4} & \frac{p_{1,3}}{4} \\
\hdashline
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & 0 & \frac{p_{2,2}}{2} & \frac{p_{2,2}}{2} & \frac{p_{2,3}}{4} & \frac{q_{2,0}}{10} & \frac{p_{2,3}}{4} & \frac{p_{2,3}}{4} & \frac{p_{2,3}}{4} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & \frac{p_{2,2}}{2} & 0 & \frac{p_{2,2}}{2} & \frac{p_{2,3}}{4} & \frac{q_{2,0}}{10} & \frac{p_{2,3}}{4} & \frac{p_{2,3}}{4} & \frac{p_{2,3}}{4} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & \frac{p_{2,2}}{2} & \frac{p_{2,2}}{2} & 0 & \frac{p_{2,3}}{4} & \frac{q_{2,0}}{10} & \frac{p_{2,3}}{4} & \frac{p_{2,3}}{4} & \frac{p_{2,3}}{4} \\
\hdashline
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & 0 & \frac{q_{3,0}}{10} & \frac{p_{3,3}}{3} & \frac{p_{3,3}}{3} & \frac{p_{3,3}}{3} \\
\frac{q_{0,1}}{2} & \frac{q_{0,1}}{2} & \frac{q_{0,2}}{3} & \frac{q_{0,2}}{3} & \frac{q_{0,2}}{3} & \frac{q_{0,3}}{4} & 0 & \frac{q_{0,3}}{4} & \frac{q_{0,3}}{4} & \frac{q_{0,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{3} & \frac{q_{3,0}}{10} & 0 & \frac{p_{3,3}}{3} & \frac{p_{3,3}}{3} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{3} & \frac{q_{3,0}}{10} & \frac{p_{3,3}}{3} & 0 & \frac{p_{3,3}}{3} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{3} & \frac{q_{3,0}}{10} & \frac{p_{3,3}}{3} & \frac{p_{3,3}}{3} & 0
\end{pmatrix}$$

$$- \frac{1}{3 \cdot 10} \begin{pmatrix}
0 & q_{1,0} & \frac{q_{1,0}}{3} & \frac{q_{1,0}}{3} & \frac{q_{1,0}}{3} & \frac{q_{1,0}}{4} & 0 & \frac{q_{1,0}}{4} & \frac{q_{1,0}}{4} & \frac{q_{1,0}}{4} \\
q_{1,0} & 0 & \frac{q_{1,0}}{3} & \frac{q_{1,0}}{3} & \frac{q_{1,0}}{3} & \frac{q_{1,0}}{4} & 0 & \frac{q_{1,0}}{4} & \frac{q_{1,0}}{4} & \frac{q_{1,0}}{4} \\
\hdashline
\frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & 0 & \frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & \frac{q_{2,0}}{4} & 0 & \frac{q_{2,0}}{4} & \frac{q_{2,0}}{4} & \frac{q_{2,0}}{4} \\
\frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & 0 & \frac{q_{2,0}}{2} & \frac{q_{2,0}}{4} & 0 & \frac{q_{2,0}}{4} & \frac{q_{2,0}}{4} & \frac{q_{2,0}}{4} \\
\frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & \frac{q_{2,0}}{2} & 0 & \frac{q_{2,0}}{4} & 0 & \frac{q_{2,0}}{4} & \frac{q_{2,0}}{4} & \frac{q_{2,0}}{4} \\
\hdashline
\frac{q_{3,0}}{2} & \frac{q_{3,0}}{2} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & 0 & 0 & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{q_{3,0}}{2} & \frac{q_{3,0}}{2} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & 0 & 0 & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} \\
\frac{q_{3,0}}{2} & \frac{q_{3,0}}{2} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & 0 & \frac{q_{3,0}}{3} & 0 & \frac{q_{3,0}}{3} \\
\frac{q_{3,0}}{2} & \frac{q_{3,0}}{2} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & 0 & \frac{q_{3,0}}{3} & \frac{q_{3,0}}{3} & 0
\end{pmatrix}. \tag{20}$$

Here, we have indicated the original cluster structure in dashed lines, and we have colored the row and column corresponding to the modified cluster behavior of state V^*. Comparing (20) to (4) helps understanding how Q is constructed and how Q compares to P. Note in particular the minor changes in the normalizations of all entries.

4.3. *Leading order behavior of $\mathbb{E}_Q[L|\sigma(V^*)]$.*

PROPOSITION 5. *For any given $V^* \in \mathcal{V}$ and $q \in \mathcal{Q}$, it holds that*

$$\mathbb{E}_Q[L|\sigma(V^*)] = \frac{T}{n} I_{\sigma(V^*)}(q||p) + o\Big(\frac{T}{n}\Big).$$

Here, $I_{\sigma(V^*)}(q||p)$ is defined in (6).

PROOF. Define $R_{x,y} \triangleq \ln(Q_{x,y}/P_{x,y})$ for notational convenience: we refer to Appendix F for its asymptotic behavior. Since the Markov chain is started from equilibrium,

$$\mathbb{E}_Q[L|\sigma(V^*)] = T \sum_{x\in\mathcal{V}} \sum_{y\in\mathcal{V}} \Pi_x^{(Q)} Q_{x,y} \ln R_{x,y}. \tag{21}$$

The largest individual contributions to the expectation in (21) are by jumps to and from V^*, since this is where the change of measure is modified most. Jumps not involving V^* contribute

less individually, but there are many of such jumps. We therefore separate out the jumps to and from V^*, i.e.,

(22)
$$\frac{\mathbb{E}_Q[L|\sigma(V^*)]}{T} = \sum_{y\neq V^*} \Pi^{(Q)}_{V^*} Q_{V^*,y} \ln R_{V^*,y} + \sum_{x\neq V^*} \Pi^{(Q)}_{x} Q_{x,V^*} \ln R_{x,V^*} + \sum_{x,y\neq V^*} \Pi^{(Q)}_{x} Q_{x,y} \ln R_{x,y}.$$

We now calculate the leading order behavior of each term.

For the first term in (22) we have by (i) Lemma 16 in Appendix F and Q's definition, see (18), and (ii) Proposition 4,

$$\sum_{y\neq V^*} \Pi^{(Q)}_{V^*} Q_{V^*,y} \ln R_{V^*,y} \overset{\text{(i)}}{\sim} \sum_{k=1}^{K} \sum_{y\in\mathcal{W}_k} \Pi^{(Q)}_{V^*} \frac{q_{0,k}}{|\mathcal{W}_k|} \ln \frac{q_{0,\omega(y)}}{p_{\sigma(V^*),\omega(y)}} \overset{\text{(ii)}}{\sim} \frac{1}{n} \sum_{k=1}^{K} \gamma_0^{[1]} q_{0,k} \ln \frac{q_{0,k}}{p_{\sigma(V^*),k}}. \tag{23}$$

The second term in (22) handles similarly:

$$\sum_{x\neq V^*} \Pi^{(Q)}_{x} Q_{x,V^*} \ln R_{x,V^*} \overset{\text{(i)}}{\sim} \sum_{k=1}^{K} \sum_{x\in\mathcal{W}_k\setminus\{V^*\}} \bar{\Pi}^{(Q)}_{k} \frac{q_{k,0}}{n} \ln \frac{q_{k,0}\alpha_{\sigma(V^*)}}{p_{k,\sigma(V^*)}} \overset{\text{(ii)}}{\sim} \frac{1}{n} \sum_{k=1}^{K} \gamma_k^{[0]} q_{k,0} \ln \frac{q_{k,0}\alpha_{\sigma(V^*)}}{p_{k,\sigma(V^*)}}.$$

The third term in (22) requires (iii) a Taylor expansion of $\ln(1+x) = x + O(x^2)$ for $x \approx 0$ and (iv) the balance equations (61)–(62), so that

$$\begin{aligned}\sum_{x,y\neq V^*} \Pi^{(Q)}_{x} Q_{x,y} \ln R_{x,y} &\overset{\text{(iii)}}{\sim} \sum_{k,l\neq 0} \sum_{x\in\mathcal{W}_k} \sum_{y\in\mathcal{W}_l\setminus\{x\}} \bar{\Pi}^{(Q)}_{k} \frac{q_{k,l}}{|\mathcal{W}_l| - \mathbb{1}[k=l]} \cdot \frac{1}{n}\Big(\frac{\mathbb{1}[l=\sigma(V^*)]}{\alpha_l} - \frac{q_{k,0}}{p_{k,l}K}\Big) \\ &\overset{(17)}{\sim} \frac{1}{n} \sum_{k=1}^{K} \gamma_k^{[0]} \Big(\frac{q_{k,\sigma(V^*)}}{\alpha_{\sigma(V^*)}} - \sum_{l=1}^{K} \frac{1}{K} q_{k,0}\Big) \overset{\text{(iv)}}{=} \frac{1}{n}\Big(\frac{\gamma^{[0]}_{\sigma(V^*)}}{\alpha_{\sigma(V^*)}} - \gamma_0^{[1]}\Big).\end{aligned} \tag{24}$$

Substituting (23)–(24) into (22) gives

$$\mathbb{E}_Q[L|\sigma(V^*)] \sim \frac{T}{n} \sum_{k=1}^{K} \Big(\gamma_0^{[1]} q_{0,k} \ln \frac{q_{0,k}}{p_{\sigma(V^*),k}} + \gamma_k^{[0]} q_{k,0} \ln \frac{q_{k,0}\alpha_{\sigma(V^*)}}{p_{k,\sigma(V^*)}}\Big) + \frac{T}{n}\Big(\frac{\gamma^{[0]}_{\sigma(V^*)}}{\alpha_{\sigma(V^*)}} - \gamma_0^{[1]}\Big).$$

By now applying Proposition 4, we complete the proof. □

4.4. *Asymptotic negligibility of* $\mathrm{Var}_Q[L|\sigma(V^*)]$.

Proposition 6. *For any given $V^* \in \mathcal{V}$ and $q \in \mathcal{Q}$, it holds that if $T = \omega(1)$, then*

$$\mathrm{Var}_Q[L|\sigma(V^*)] = O\Big(\frac{T\ln T}{n}\Big).$$

As a consequence if $T = \omega(n)$, then $\mathrm{Var}_Q[L|\sigma(V^*)] = o(T^2/n^2)$.

Proof. Define $L_t \triangleq \ln(Q_{X_{t-1},X_t}/P_{X_{t-1},X_t})$. Expanding, we obtain

$$\mathrm{Var}_Q[L|\sigma(V^*)] = \mathrm{Var}_Q\Big[\sum_{t=1}^{T} L_t \Big| \sigma(V^*)\Big] = \sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{Cov}_Q[L_t, L_s|\sigma(V^*)]. \tag{25}$$

We now consider the cases $|t-s| \geq 2$ and $|t-s| \leq 1$, in that order. The idea is that we bound relatively crudely in the latter cases since there are only $O(T)$ such terms that contribute to the sum, and the former cases more sharply. Applying one rough bound on all terms of the former cases would not suffice, because there are as many as $O(T^2)$ such terms. As we will show for the former cases, we can derive a sharper bound when $|t-s| \gg t_{\mathrm{mix}}(\varepsilon)$ because Proposition 2 implies that the covariances decay quickly.

First note that because (i) the process is started from equilibrium, we have for any $t, s \in \{1, \dots, T\}$ that

$$\begin{aligned} \mathrm{Cov}_Q[L_t, L_s|\sigma(V^*)] &= \mathbb{E}_Q[L_t L_s|\sigma(V^*)] - \mathbb{E}_Q[L_t|\sigma(V^*)]\mathbb{E}_Q[L_s|\sigma(V^*)] \\ &\overset{\text{(i)}}{=} \mathbb{E}_Q[L_t L_s|\sigma(V^*)] - \mathbb{E}_Q[L_t|\sigma(V^*)]^2. \end{aligned} \tag{26}$$

Now consider the case $|t-s| \geq 2$. Define $S_{x,y,u,v} \triangleq (\ln R_{x,y})(\ln R_{u,v})$ for notational convenience: we refer to Appendix F for its asymptotic behavior. In this case the first term of (26) evaluates as

$$\begin{aligned} \mathbb{E}_Q[L_t L_s|\sigma(V^*)] &= \sum_{x,y,u,v} \mathbb{P}_Q[X_{t\wedge s-1} = x, X_{t\wedge s} = y, X_{t\vee s-1} = u, X_{t\vee s} = v|\sigma(V^*)] S_{x,y,u,v} \\ &= \sum_{x,y,u,v} \Pi_x^{(Q)} Q_{x,y} \Big(\sum_{z_{t\wedge s+1}, \dots, z_{t\vee s-2}} Q_{y,z_{t\wedge s+1}} Q_{z_{t\wedge s+1}, z_{t\wedge s+2}} \cdots Q_{z_{t\vee s-2},u} \Big) Q_{u,v} S_{x,y,u,v} \\ &= \sum_{x,y,u,v} \Pi_x^{(Q)} Q_{x,y} Q_{y,u}^{|t-s|-1} Q_{u,v} S_{x,y,u,v}. \end{aligned} \tag{27}$$

The second term of (26) expands as

$$\mathbb{E}_Q[L_t|\sigma(V^*)]^2 = \Big(\sum_{x,y} \Pi_x^{(Q)} Q_{x,y} \ln \frac{Q_{x,y}}{P_{x,y}} \Big)^2 = \sum_{x,y,u,v} \Pi_x^{(Q)} Q_{x,y} \Pi_u^{(Q)} Q_{u,v} S_{x,y,u,v}. \tag{28}$$

Substituting (27) and (28) into the last member of (26) gives

$$\mathrm{Cov}_Q[L_t, L_s|\sigma(V^*)] = \sum_{x,y,u,v} \Pi_x^{(Q)} Q_{x,y} (Q_{y,u}^{|t-s|-1} - \Pi_u^{(Q)}) Q_{u,v} S_{x,y,u,v}. \tag{29}$$

In order to bound (29), we need to take two effects into consideration: a filter effect that happens because the transition matrix Q is similar to the transition matrix P, and a concentration effect because the Markov chain moves closer to equilibrium as time progresses. The filter effect is quantified by Corollary 1 in Appendix F, because the latter implies that $\sum_{x,y,u,v} S_{x,y,u,v} \leq c_1 n^2$ for some absolute constant c_1 (even though $\sum_{x,y,u,v} 1 = n^4$). We can use the effect by for example bounding $\Pi_x^{(Q)} Q_{x,y} (Q_{y,u}^m - \Pi_u^{(Q)}) Q_{u,v} \leq c_2/n^4$ uniformly using another absolute constant, and then concluding that $\mathrm{Cov}_Q[L_t, L_s|\sigma(V^*)] \leq c_2 (T^2/n^4) \sum_{x,y,u,v} S_{x,y,u,v} \leq c_1 c_2 T^2/n^2$. However, this bound is not sufficiently sharp for our purposes: we need to provide a bound that is at least $o(T^2/n^2)$.

To arrive at a sharper bound, we use the concentration of the Markov chain. Apply the triangle inequality first, and then bound $\Pi_x^{(Q)} Q_{x,y} Q_{u,v} \leq c_1/n^3$ uniformly using an absolute constant c_1 to obtain

$$\Big| \sum_{t=1}^T \sum_{s=1}^T \mathbb{1}[|t-s| \geq 2] \mathrm{Cov}_Q[L_t, L_s|\sigma(V^*)] \Big| \leq \frac{2c_1}{n^3} \sum_{t=1}^T \sum_{s=t+2}^T \sum_{x,y,u,v} |Q_{y,u}^{|t-s|-1} - \Pi_u^{(Q)}| |S_{x,y,u,v}|. \tag{30}$$

Now let $m \in \mathbb{N}_+$. By nonnegativity of the summands and (3), $|Q_{x,y}^m - \Pi_y^{(Q)}| \leq \sum_y |Q_{x,y}^m - \Pi_y^{(Q)}| = d_1(Q_{x,\cdot}^m, \Pi^{(Q)})$. Note furthermore that Lemma 15, see Appendix B, implies that there exists an absolute constant c_2 so that $d_1(\mu, \nu) \leq c_2 d_2(\mu, \nu)$ for any two probability distributions μ, ν. This implies that we have $|Q_{x,y}^m - \Pi_y^{(Q)}| \leq c_2 d_2(Q_{x,\cdot}^m, \Pi^{(Q)})$ for all $x, y \in \mathcal{V}$. Recalling (2), it therefore follows after maximization over $x \in \mathcal{V}$ that

$$|Q_{x,y}^m - \Pi_y^{(Q)}| \leq c_2 d(m) \quad \text{for all} \quad x, y \in \mathcal{V} \text{ and } m \in \mathbb{N}_+. \tag{31}$$

Using (31) to bound the r.h.s. in (30), we obtain

$$\Big|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{1}[|t-s|\geq 2]\mathrm{Cov}_Q[L_t,L_s|\sigma(V^*)]\Big| \leq \frac{2c_1c_2}{n^3}\sum_{t=1}^{T}\sum_{s=t+2}^{T} d(|t-s|-1)\sum_{x,y,u,v}|S_{x,y,u,v}|$$

$$\overset{\text{(ii)}}{\leq} c_3\frac{T}{n}\sum_{m=1}^{T} d(m), \tag{32}$$

with $c_3 > 0$ an absolute constant. Here we have (ii) used the filter effect and extended the summation range, which is valid since $d(m) \geq 0$ for all $m \in \mathbb{N}_+$.

We now split the sum according to the mixing time. That is, for some absolute constant $c_4 > 0$

$$\sum_{m=1}^{T} d(m) = \sum_{m=1}^{t_{\mathrm{mix}}(\varepsilon)} d(m) + \sum_{m=t_{\mathrm{mix}}(\varepsilon)+1}^{T} d(m) \overset{\text{(iii)}}{\leq} c_4 t_{\mathrm{mix}}(\varepsilon) + T\varepsilon. \tag{33}$$

To obtain the inequality in (33) we used the facts (iii) that $d(m)$ is nonincreasing in $m \in \mathbb{N}_+$ [2], that for all $x \in \mathcal{V}$

$$d_2(Q_{x,\cdot},\Pi^{(Q)}) \overset{(3)}{=} \Big(\sum_{y\in\mathcal{V}}\frac{(Q_{x,y}-\Pi_y^{(Q)})^2}{\Pi_y^{(Q)}}\Big)^{1/2} \leq \frac{1}{\sqrt{\Pi_{\min}^{(Q)}}}\Big(\sum_{y\in\mathcal{V}} O\Big(\frac{1}{n^2}\Big)\Big)^{1/2} = O(1)$$

so that there exists an absolute constant such that $d(1) \leq c_4$, and that t_{mix} is such that $d(t) \leq \varepsilon$ for all $t \geq t_{\mathrm{mix}}$. By Proposition 2, there exists an absolute $c_{\mathrm{mix}} > 0$ such that $t_{\mathrm{mix}}(\varepsilon) \leq -c_{\mathrm{mix}}\ln\varepsilon$. After substituting (33) into (32) and then minimizing over ε, giving the optimal choice $\varepsilon = c_4c_{\mathrm{mix}}/T$, we obtain

$$\Big|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{1}[|t-s|\geq 2]\mathrm{Cov}[L_t,L_s]\Big| \leq c_3\frac{T}{n}\Big(c_5\ln T - c_5\ln c_5 + c_5\Big) = O\Big(\frac{T}{n}\ln T\Big), \tag{34}$$

where $c_5 = c_4c_{\mathrm{mix}} > 0$ is again an absolute constant.

Lastly we deal with the cases $|t-s| \leq 1$. When $|t-s| = 0$, or equivalently $t = s$, we have that (iv) because of Lemma 16 and Corollary 1 that there exist absolute constants $c_6,\ldots,c_9$ such that

$$\mathrm{Cov}[L_t,L_t|\sigma(V^*)] \overset{(26)}{\leq} \mathbb{E}_Q[L_t^2|\sigma(V^*)] = \sum_{x\in\mathcal{V}}\sum_{y\in\mathcal{V}}\Pi_x^{(Q)}Q_{x,y}(\ln R_{x,y})^2$$

$$\overset{\text{(iv)}}{\leq} \Pi_{V^*}^{(Q)}\sum_{y\neq V^*}Q_{V^*,y}c_6 + \sum_{x\neq V^*}\Pi_x^{(Q)}Q_{x,V^*}c_7 + \sum_{x\neq V^*}\sum_{y\neq V^*}\Pi_x^{(Q)}Q_{x,y}\frac{c_8}{n^2} \leq \frac{c_9}{n}$$

for all $t = 1,\ldots,T$. Therefore

$$\Big|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{1}[|t-s|=0]\mathrm{Cov}_Q[L_t,L_s|\sigma(V^*)]\Big| = \sum_{t=1}^{T}\mathrm{Var}_Q[L_t|\sigma(V^*)] = O\Big(\frac{T}{n}\Big). \tag{35}$$

When $|t-s| = 1$, there exists an absolute constant $c_{10} > 0$ such that

$$\mathrm{Cov}_Q[L_{t\wedge s},L_{t\wedge s+1}|\sigma(V^*)] \overset{(26)}{\leq} \mathbb{E}_Q[L_{t\wedge s}L_{t\wedge s+1}|\sigma(V^*)] \leq \sum_{x,y,z}\Pi_x^{(Q)}Q_{x,y}Q_{y,z}S_{x,y,y,z}$$

$$\leq \frac{c_{10}}{n^3}\sum_{x,y,z}S_{x,y,y,z}.$$

Invoking Corollary 1's filter effect implies that $\sum_{x,y,z}S_{x,y,y,z} = O(n^2)$. Therefore

$$\Big|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{1}[|t-s|=1]\mathrm{Cov}[L_t,L_s|\sigma(V^*)]\Big| \leq 2\sum_{t=1}^{T}|\mathrm{Cov}_Q[L_t,L_{t+1}|\sigma(V^*)]| = O\Big(\frac{T}{n}\Big). \tag{36}$$

Splitting (25) into the respective cases and then (v) substituting (34), (35), and (36) gives

$$\begin{aligned}\mathrm{Var}_Q[L|\sigma(V^*)] &= \sum_{t=1}^{T}\sum_{s=1}^{T}(\mathbb{1}[|t-s|=0]+\mathbb{1}[|t-s|=1]+\mathbb{1}[|t-s|\geq 2])\mathrm{Cov}_Q[L_t,L_s|\sigma(V^*)]\\ &\overset{(v)}{=} O\Big(\frac{T}{n}\ln T\Big),\end{aligned}$$

which completes the proof. □

4.5. *Appropriateness, deconditioning, and bound optimization.* Recall that the transition matrix Q is constructed given a state $V^* \in \mathcal{V}$. This implies in particular that Proposition 5 and Proposition 6 have determined the asymptotic behavior of the conditional expectation $\mathbb{E}_Q[L|\sigma(V^*)]$ and conditional variance $\mathrm{Var}_Q[L|\sigma(V^*)]$. The information bound Proposition 3 requires us however to analyze their unconditioned counterparts, which is the subject of this section. Importantly, we can limit the variance introduced by a random state selection by choosing $q \in \mathcal{Q}$ appropriately.

4.5.1. *Appropriateness.*

LEMMA 1. *If $I(\alpha,p)>0$, then for any two clusters $a\neq b$ there exists at least one finite point $\bar{q}\in\mathcal{Q}$ such that $I_a(\bar{q}||p)=I_b(\bar{q}||p)$. Furthermore, for any such $\bar{q}$ it holds that $0<I_a(\bar{q}||p)=I_b(\bar{q}||p)<\infty$.*

PROOF. Consider the points

$$q_c=\Big(\frac{p_{1,c}}{\alpha_c},\dots,\frac{p_{K,c}}{\alpha_c};p_{c,1},\dots,p_{c,K};0\Big)\in\mathcal{Q}\quad\text{where}\quad c\in\{1,\dots,K\}.$$

Let $a\neq b$. The points q_a, q_b have the following properties: (i) $I_a(q_a||p)=I_b(q_b||p)=0$, and (ii) $I(\alpha,p)\leq I_a(q_b||p)<\infty$ and $I(\alpha,p)\leq I_b(q_a||p)<\infty$ by definition of $I(\alpha,p)$. Together with the continuity of $I_c(q||p)$ w.r.t. $q\in\mathcal{Q}$, this implies that there exists a $\lambda\in(0,1)$ such that

$$I_a(\lambda q_a+(1-\lambda)q_b||p)=I_b(\lambda q_a+(1-\lambda)q_b||p).$$

This establishes the existence.

Next we prove the second claim of the lemma. Recall that

$$\mathbb{E}_Q[L|\sigma(V^*)]=\sum_{\text{all sample paths }\chi}\mathbb{P}_Q[\chi|\sigma(V^*)]\ln\frac{\mathbb{P}_Q[\chi|\sigma(V^*)]}{\mathbb{P}_P[\chi]}$$

is a KL-divergence. As a consequence, $\mathbb{E}_Q[L]=0$ if and only if

$$\mathbb{P}_Q[\chi|\sigma(V^*)]=\prod_{t=1}^{T}Q_{x_{t-1},x_t}=\prod_{t=1}^{T}P_{x_{t-1},x_t}=\mathbb{P}_P[\chi]\quad\text{for all sample paths }\chi.$$

Equivalently $\mathbb{E}_Q[L|\sigma(V^*)]=0$ if and only if $Q_{x,y}=P_{x,y}$ for all $x,y\in\mathcal{V}$, which can be seen by considering the set of paths that disagree only on the last jump. Since

$$I_{\sigma(V^*)}(q||p)=\lim_{n\to\infty}\frac{n}{T}\Big(\mathbb{E}_Q[L|\sigma(V^*)]+o(1)\Big),$$

we obtain that $I_{\sigma(V^*)}(q||p)=0$ if and only if $q=q_{\sigma(V^*)}$. Since $I_a(q_a||p)=0$ and $I_b(q_a||p)\geq I(\alpha,p)>0$ by definition of $I(\alpha,p)$, it must be that for any $\bar{q}$ such that $I_a(\bar{q}||p)=I_b(\bar{q}||p)$, it holds that $\bar{q}\neq q_a,q_b$. This completes the proof. □

Note that we do not upper bound $I_a(\bar{q}||p)$ or $I_b(\bar{q}||p)$ by $I(\alpha, p)$. While we believe that such a strong statement may indeed hold, proving this would require the use of additional monotonicity-like properties. For example, if it can be established that $I_c(q||p)$ is quasiconvex in q, one could try to leverage this fact. Establishing such a property however proved elusive to us, and we were able to establish that $I_c(q||p)$ is not convex in q.

We now prove that the change of measure Ψ satisfies condition (9).

LEMMA 2. *For any two clusters $a \neq b$, if Q is constructed using any $\bar{q} \in \mathcal{Q}(a,b)$, recall its definition in* (7)*, then there exists an absolute constant δ such that $\mathbb{P}_\Psi[V^* \in \mathcal{E}] \geq \delta > 0$.*

PROOF. Let $a \neq b$ denote any two distinct clusters, and construct Q using some $\bar{q} \in \mathcal{Q}(a,b)$. By (i) the law of total probability,

$$\mathbb{P}_\Psi[V^* \in \mathcal{E}] = 1 - \mathbb{P}_\Psi[V^* \notin \mathcal{E}] \overset{\text{(i)}}{=} 1 - \sum_{c=a,b} \frac{\alpha_c}{\alpha_a + \alpha_b} \mathbb{P}_Q[V^* \in \hat{\mathcal{V}}_{\gamma(c)} | \sigma(V^*) = c].$$

Since $q_a, q_b \notin \mathcal{Q}(a,b)$, the state V^* behaves differently than any state in either cluster a or b. We must therefore have $0 \leq \mathbb{P}_Q[V^* \in \hat{\mathcal{V}}_{\gamma(c)} | \sigma(V^*) = c] < 1$ for $c = a, b$. When $x, y \in [0,1]$, $\alpha_a x + \alpha_b y = \alpha_a + \alpha_b$ if and only if $x, y = 1$. This implies that there exists a $\delta > 0$ such that $\mathbb{P}_\Psi[V^* \in \mathcal{E}] \geq \delta > 0$. This completes the proof. □

It is important to note that given $a \neq b$ and $\bar{q} \in \mathcal{Q}(a,b)$ with which Q is constructed, it need *not* be that e.g. $\mathbb{P}_Q[V^* \in \hat{\mathcal{V}}_a | \sigma(V^*) = a]$ equals $\mathbb{P}_Q[V^* \in \hat{\mathcal{V}}_a | \sigma(V^*) = b]$ for any clustering algorithm when n is finite. This is because the rows of the transition matrix are normalized. In particular, depending on which cluster V^* originated from, the transition probabilities from/to the originating cluster differ on the order of $O(1/n^2)$.

4.5.2. *Deconditioning.* We now revisit Proposition 3 to account for the specific choices made in the change-of-measure.

LEMMA 3. *If $T = \omega(n)$, then for any two clusters $a \neq b$, there exists a strictly positive constant $c > 0$ independent of n such that*

$$\frac{\mathbb{E}_P[|\mathcal{E}|]}{n} \geq c \exp\Big(-\frac{T}{n} I_{a,b}(\bar{q}||p) + o\Big(\frac{T}{n}\Big)\Big). \tag{37}$$

Here,

$$I_{a,b}(\bar{q}||p) = \frac{\alpha_a}{\alpha_a + \alpha_b} I_a(\bar{q}||p) + \frac{\alpha_b}{\alpha_a + \alpha_b} I_b(\bar{q}||p)$$

for any point $\bar{q}$ selected from the set $\mathcal{Q}(a,b)$.

PROOF. Let $a \neq b$ be any two distinct clusters. Choose $q = \bar{q} \in Q(a,b)$ such that $I_a(\bar{q}||p) = I_b(\bar{q}||p) \triangleq I_{a,b}(\bar{q}||p) \in (0, \infty)$. This is possible by Lemma 1. Select V^* uniformly at random in $\mathcal{V}_a \cup \mathcal{V}_b$. Then by (i) the law of total probability and (ii) Proposition 5

$$\mathbb{E}_\Psi[L] \overset{\text{(i)}}{=} \frac{\alpha_a}{\alpha_a + \alpha_b} \mathbb{E}_Q[L | \sigma(V^*) = a] + \frac{\alpha_b}{\alpha_a + \alpha_b} \mathbb{E}_Q[L | \sigma(V^*) = b] \overset{\text{(ii)}}{=} \frac{T}{n} I_{a,b}(p) + o\Big(\frac{T}{n}\Big).$$

By (iii) the law of total variance $\mathrm{Var}[X] = \mathbb{E}_Y[\mathrm{Var}[X|Y]] + \mathrm{Var}_Y[\mathbb{E}[X|Y]]$, we have

$$\begin{aligned}\mathrm{Var}_\Psi[L] &\overset{\text{(iii)}}{=} \frac{\alpha_a}{\alpha_a + \alpha_b} \mathrm{Var}_Q[L | \sigma(V^*) = a] + \frac{\alpha_b}{\alpha_a + \alpha_b} \mathrm{Var}_Q[L | \sigma(V^*) = b] \\ &+ \frac{\alpha_a}{\alpha_a + \alpha_b} \Big(\mathbb{E}_Q[L | \sigma(V^*) = a] - \mathbb{E}_\Psi[L]\Big)^2 + \frac{\alpha_b}{\alpha_a + \alpha_b} \Big(\mathbb{E}_Q[L | \sigma(V^*) = b] - \mathbb{E}_\Psi[L]\Big)^2 \overset{\text{(iv)}}{=} o\Big(\frac{T^2}{n^2}\Big),\end{aligned}$$

where for (iv) we have used Proposition 6 for the first two terms, and the fact that $\bar{q} \in Q(a,b)$ guarantees that $I_a(\bar{q}||p) = I_b(\bar{q}||p)$ for the last two terms. □

4.5.3. *Bound optimization.* We finally optimize the bound in (37). This is straightforward: build the change of measure using the parameters

$$(k^{\mathrm{opt}}, l^{\mathrm{opt}}, q^{\mathrm{opt}}) \in \arg\min_{k\neq l} \min_{q\in\mathcal{Q}(k,l)} \Big\{ \frac{\alpha_k}{\alpha_k+\alpha_l} I_k(q||p) + \frac{\alpha_l}{\alpha_k+\alpha_l} I_l(q||p) \Big\}.$$

Then by construction

$$\mathbb{E}_\Psi[L] = \frac{T}{n} J(\alpha, p) + o\Big(\frac{T}{n}\Big),$$

and since $q^{\mathrm{opt}} \in Q(k^{\mathrm{opt}}, l^{\mathrm{opt}})$, we have $0 < J(\alpha, p) < \infty$. This completes the proof of Theorem 1.

5. The Spectral Clustering Algorithm. The Spectral Clustering Algorithm, whose pseudo-code is presented in Algorithm 1, aims at providing good first estimates of the clusters, and works as follows. First we observe a finite trajectory $X_0, X_1, \ldots, X_T$ of the Markov chain, where $T \in \mathbb{N}_0$. We then calculate the empirical kernels $\hat{P} \in \mathbb{\Delta}^{(n-1)\times n}$, an approximation of the transition matrix P, element-wise as

$$\hat{P}_{x,y} \triangleq \frac{\sum_{t=0}^{T-1} \mathbb{1}[X_t = x, X_{t+1} = y]}{\sum_{t=0}^{T-1} \mathbb{1}[X_t = x]} \quad \text{for} \quad x, y \in \mathcal{V}.$$

Note that because $\hat{P}$ is constructed from a finite trajectory of a Markov chain, its elements are not independent. We also calculate the matrix

$$\hat{N}_{x,y} \triangleq \sum_{s=0}^{T-1} \mathbb{1}[X_{s-1} = x, X_s = y] \quad \text{for} \quad x, y \in \mathcal{V},$$

and write their expectations as $N_{x,y} = \mathbb{E}[\hat{N}_{x,y}]$ for $x, y \in \mathcal{V}$.

Next we construct a rank-K approximation $\hat{R}$ of $\hat{P}$, or of $\hat{N}$ (one may choose), from its singular value decomposition $\hat{P} = U\Sigma V^{\mathrm{T}}$. Specifically, we define

$$\hat{R} = \sum_{k=1}^{K} \sigma_k U_{\cdot,k} {V_{\cdot,k}}^{\mathrm{T}},$$

where the values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ denote the singular values of $\hat{P}$ in decreasing order.

We apply a clustering algorithm to these basis vectors to determine the clusters. While in practice you may choose to use a different algorithm, for the analysis we use the following: first we calculate the *neighborhoods*

$$\mathcal{N}_x \triangleq \{y \in \mathcal{V} | \|\hat{R}_{x,\cdot} - \hat{R}_{y,\cdot}\|_2 \leq h(n,T)\} \quad \text{for} \quad x \in \mathcal{V}.$$

Then we initialize $\hat{\mathcal{V}}_k \leftarrow \emptyset$ for $k = 0, 1, \ldots, K$ and select K *centers* $z_1^*, \ldots, z_K^* \in \mathcal{V}$ with which we construct approximate clusters. Specifically, we iterate

$$\hat{\mathcal{V}}_k \leftarrow \mathcal{N}_{z_k^*} \backslash \{\cup_{l=0}^{k-1} \hat{\mathcal{V}}_l\} \quad \text{where} \quad z_k^* \triangleq \arg\max_{x\in\mathcal{V}} |\mathcal{N}_x \backslash \{\cup_{l=0}^{k-1} \hat{\mathcal{V}}_l\}| \tag{38}$$

for $k = 1, \ldots K$. Any remaining state is finally associated to the center closest to it, i.e., we iterate for $y \in \{\cup_{k=1}^{K} \hat{\mathcal{V}}_k\}^{\mathrm{c}}$

$$\hat{\mathcal{V}}_{k_y^*} \leftarrow \hat{\mathcal{V}}_{k_y^*} \cup \{y\} \quad \text{with} \quad k_y^* \triangleq \arg\min_{k=1,\ldots,K} \|\hat{R}_{z_k^*,\cdot} - \hat{R}_{y,\cdot}\|_2. \tag{39}$$

Finally, we output $\hat{\mathcal{V}}_k^{[0]} = \hat{\mathcal{V}}_k$ for $k = 1, \ldots, K$.

Theorem 2 gives an upper bound on the number of misclassified states after the Spectral Clustering Algorithm has been applied. Given a desired level of accuracy, Theorem 2 can provide sufficient conditions on α, p, n, T that guarantee the Spectral Clustering Algorithm achieves a desired level of accuracy.

THEOREM 2. *If $\|\hat{P} - P\| = o_{\mathbb{P}}(g(n,T))$ for some $g(n,T) = o(1)$ and $h(n,T)$ is chosen such that $\omega(g^2/n) = h^2 = o(D_P(\alpha,p)/n)$, then*

$$|\mathcal{E}| = O_{\mathbb{P}}\Big(\frac{n}{D_P(\alpha,p)} \cdot \big(g(n,T)\big)^2\Big) \quad \textit{with} \quad D_P(\alpha,p) = \min_{a\neq b} \sum_{k=1}^{K} \frac{1}{\alpha_k}(p_{a,k} - p_{b,k})^2.$$

Similarly if $\|\hat{N} - N\| = o_{\mathbb{P}}(f(n,T))$ for some $f(n,T) = o(T/n)$ and $h(n,T)$ is chosen such that $\omega((n/T)^2 \cdot (f^2/n)) = h^2 = o(D_P(\alpha,p)/n)$, then

$$|\mathcal{E}| = O_{\mathbb{P}}\Big(\frac{n}{D_N(\alpha,p)} \cdot \Big(\frac{n}{T} f(n,T)\Big)^2\Big) \quad \textit{with} \quad D_N(\alpha,p) = \min_{a\neq b} \sum_{k=1}^{K} \frac{1}{\alpha_k}\Big(\frac{\pi_a p_{a,k}}{\alpha_a} - \frac{\pi_b p_{b,k}}{\alpha_b}\Big)^2. \tag{40}$$

Here, π denotes the solution to $\pi^{\mathrm{T}} p = \pi^{\mathrm{T}}$.

5.1. *Proof of Theorem 2.* We will prove Theorem 2 in the case that the Spectral Clustering Algorithm is applied to $\hat{P}$. By repeating the arguments, the analogous result for $\hat{N}$ can be established: we only indicate the different separability argument from which (40) follows. The proof of Theorem 2 consists of four steps.

Step 1. We show that the transition matrix P satisfies a *separability property*: i.e., if two states $x, y \in \mathcal{V}$ do not belong to the same cluster, the l_2-distance between their respective rows $P_{x,\cdot}$, $P_{y,\cdot}$ is at least $\Omega(\sqrt{D(\alpha,p)/n})$.

Step 2. We upper bound the error $\|\hat{R} - P\|_{\mathrm{F}}$ using $\|\hat{P} - P\|$.

Step 3. We prove that $\hat{R}$ also satisfies the separability property if $\|\hat{P} - P\| \to 0$, as suggested by Step 1 and Step 2.

Step 4. Because of $\hat{R}$'s separability property, we must conclude that the number of misclassified states satisfies Theorem 2. Otherwise the separability property of Step 3 would contradict with Step 2.

Step 1: P satisfies a separability property.

LEMMA 4. *For any $x, y \in \mathcal{V}$ for which $\sigma(x) \neq \sigma(y)$,*

$$\|P_{x,\cdot} - P_{y,\cdot}\|_2 = \Omega\Big(\sqrt{\frac{D_P(\alpha,p)}{n}}\Big).$$

PROOF. By (i) definition of the elements $P_{x,y}$, see (1),

$$\|P_{x,\cdot} - P_{y,\cdot}\|_2^2 = \sum_{z\in\mathcal{V}} |P_{x,z} - P_{y,z}|^2 \overset{\text{(i)}}{=} \sum_{k=1}^{K} \sum_{z\in\mathcal{V}_k} \Big| \frac{p_{\sigma(x),k}}{|\mathcal{V}_k| - \mathbb{1}[\sigma(x) = k]} - \frac{p_{\sigma(y),k}}{|\mathcal{V}_k| - \mathbb{1}[\sigma(y) = k]} \Big|^2.$$

We therefore have that

$$\|P_{x,\cdot} - P_{y,\cdot}\|_2^2 \sim \sum_{k=1}^{K} \frac{(p_{\sigma(x),k} - p_{\sigma(y),k})^2}{n\alpha_k} \geq \frac{1}{n} D_P(\alpha,p)$$

asymptotically. This completes the proof. □

Step 2: The error $\|\hat{R} - P\|_{\mathrm{F}}$ is asymptotically bounded by $\|\hat{P} - P\|$.

LEMMA 5. $\|\hat{R} - P\|_{\mathrm{F}} \leq \sqrt{K}(1 + \sqrt{2})\|\hat{P} - P\|$.

PROOF. Recall that for the Frobenius norm it holds for any matrix $A \in \mathbb{R}^{n\times n}$ that $\|A\|_{\mathrm{F}}^2 = \sum_{i=1}^n \sigma_i^2(A)$, and that for the spectral norm $\|A\| = \max_{i=1,\dots,n}\{\sigma_i(A)\}$. Because both $\hat{R}$ and P are of rank K, the matrix $\hat{R}-P$ is also of rank K, and therefore

$$\|\hat{R}-P\|_{\mathrm{F}}^2 \le K\|\hat{R}-P\|^2.$$

By the triangle inequality it then follows that

$$\|\hat{R}-P\|_{\mathrm{F}} \le \sqrt{K}(\|\hat{R}-\hat{P}\| + \|\hat{P}-P\|). \tag{41}$$

Since K is independent of n, all that remains is to bound $\|\hat{R}-\hat{P}\|$ using $\|\hat{P}-P\|$.

We (i) use [21, Thm. 9.1] to bound $\|\hat{R}-\hat{P}\|$. When choosing $\Omega = I$ and $Q = U_{(:,[1,\dots,K])}$, this theorem gives the upper bound

$$\|\hat{P}-\hat{R}\| = \|\hat{P} - U_{(:,[1,\dots,K])}{U_{(:,[1,\dots,K])}}^{\mathrm{T}}\hat{P}\| = \|(I-QQ^{\mathrm{T}})\hat{P}\| \overset{\text{(i)}}{\le} \sqrt{2}\sigma_{K+1}(\hat{P}). \tag{42}$$

By (ii) applying the triangle inequality and (iii) since P is of rank K, it then follows that

$$\begin{aligned}\sigma_{K+1}(\hat{P}) &= \|U_{(:,[K+1,\dots,n])}{U_{(:,[K+1,\dots,n])}}^{\mathrm{T}}\hat{P}\| = \|U_{(:,[K+1,\dots,n])}{U_{(:,[K+1,\dots,n])}}^{\mathrm{T}}(\hat{P}-P+P)\| \\ &\overset{\text{(ii)}}{\le} \|U_{(:,[K+1,\dots,n])}{U_{(:,[K+1,\dots,n])}}^{\mathrm{T}}(\hat{P}-P)\| + \|U_{(:,[K+1,\dots,n])}{U_{(:,[K+1,\dots,n])}}^{\mathrm{T}}P\| \overset{\text{(iii)}}{\le} \|\hat{P}-P\|.\end{aligned} \tag{43}$$

The proof is completed after bounding (41) by (42) and (43). □

Step 3: $\hat{R}$ also satisfies a separability property.

LEMMA 6. *If $\|\hat{P}-P\| = o_{\mathbb{P}}(g(n,T))$ for some $g(n,T) = o(1)$ and $h(n,T)$ is such that $\omega\big((g(n,T))^2/n\big) = \big(h(n,T)\big)^2 = o(D_P(\alpha,p)/n)$, then*

$$\|\hat{R}_{x,\cdot} - P_{x,\cdot}\|_2 = \Omega_{\mathbb{P}}\Big(\sqrt{\frac{D_P(\alpha,p)}{n}}\Big) \quad \textit{for any misclassified vertex} \quad x \in \mathcal{E}.$$

PROOF. Define $\bar{P}_k \triangleq \langle P_{z,\cdot}\rangle_{z\in\mathcal{V}_k}$ for $k = 1,\dots,K$. Let $0 < a < 1/2$, $1+a < b < \infty$ be two constants. Define the set of *cores*:

$$\mathcal{C}_k \triangleq \{x \in \mathcal{V}_k \,|\, \|\hat{R}_{x,\cdot} - \bar{P}_k\|_2 \le a h(n,T)\} \quad \text{for} \quad k = 1,\dots,K,$$

i.e., states from cluster k for which $\hat{R}_{x,\cdot}$ is correctly close to cluster k's center. Define also the set of *outliers*:

$$\mathcal{O} \triangleq \{x \in \mathcal{V} \,|\, \|\hat{R}_{x,\cdot} - \bar{P}_k\|_2 \ge b h(n,T) \text{ for all } k = 1,\dots,K\},$$

so states for which $\hat{R}_{x,\cdot}$ is far from any cluster's center.

Let $x \in \mathcal{O}$, and then choose any cluster $k \in \{1,\dots,K\}$ and any of its core states $y \in \mathcal{C}_k$. By (i) first centering and then applying the reverse triangle inequality, we find

$$\|\hat{R}_{x,\cdot} - \hat{R}_{y,\cdot}\|_2 \overset{\text{(i)}}{\ge} \big|\|\hat{R}_{x,\cdot} - \bar{P}_k\|_2 - \|\hat{R}_{y,\cdot} - \bar{P}_k\|_2\big|$$

Since $x \in \mathcal{O}$ and $y \in \mathcal{C}_k$, it follows that $\|\hat{R}_{x,\cdot} - \hat{R}_{y,\cdot}\|_2 \ge (b-a)h(n,T)$. Furthermore $b - a > 1$, implying that $y \notin \mathcal{N}_x$. We have shown that $\mathcal{N}_x \cap (\cup_{k=1}^K \mathcal{C}_k) = \emptyset$ for all $x \in \mathcal{O}$.

By (ii) Lemma 5

$$\begin{aligned}K(1+\sqrt{2})^2\|\hat{P}-P\|^2 &\overset{\text{(ii)}}{\ge} \|\hat{R}-P\|_{\mathrm{F}}^2 = \sum_{x\in\mathcal{V}} \|\hat{R}_{x,\cdot} - \bar{P}_{\sigma(x)}\|_2^2 \\ &\ge |(\cup_{k=1}^K \mathcal{C}_k)^{\mathrm{c}}| \min_{x \in (\cup_{k=1}^K \mathcal{C}_k)^{\mathrm{c}}} \{\|\hat{R}_{x,\cdot} - \bar{P}_{\sigma(x)}\|_2^2\} \ge |(\cup_{k=1}^K \mathcal{C}_k)^{\mathrm{c}}|(a h(n,T))^2.\end{aligned}$$

Rearrange to conclude that $|(\cup_{k=1}^K \mathcal{C}_k)^{\mathrm{c}}| = O_{\mathbb{P}}((g(n,T)/h(n,T))^2) = o_{\mathbb{P}}(n)$ by our assumptions on $h(n,T)$. Similarly for any $k \in \{1,\dots,K\}$, $K(1+\sqrt{2})^2\|\hat{P}-P\|^2 \geq |\mathcal{C}_k^{\mathrm{c}} \cap \mathcal{V}_k|(ah(n,T))^2$ such that $|\mathcal{C}_k| = |\mathcal{V}_k| - |\mathcal{C}_k^{\mathrm{c}} \cap \mathcal{V}_k| \geq n\alpha_k - O_{\mathbb{P}}((g(n,T)/h(n,T))^2) = n\alpha_k(1-o_{\mathbb{P}}(1))$ by the assumptions on $h(n,T)$.

For any $x \in \mathcal{O}$, $|\mathcal{N}_x| \leq |(\cup_{k=1}^K \mathcal{C}_k)^{\mathrm{c}}| = O_{\mathbb{P}}((g(n,T)/h(n,T))^2)$, since $\mathcal{N}_x \cap (\cup_{k=1}^K \mathcal{C}_k) = \emptyset$ and therefore $\mathcal{N}_x \subseteq (\cup_{k=1}^K \mathcal{C}_k)^{\mathrm{c}}$. For any $y \in \cup_{k=1}^K \mathcal{C}_k$, $\mathcal{C}_{\sigma(y)} \subseteq \mathcal{N}_y$ since $a < 1/2$ and therefore $|\mathcal{N}_y| \geq |\mathcal{C}_{\sigma(y)}| = n\alpha_{\sigma(y)} - O_{\mathbb{P}}((g(n,T)/h(n,T))^2)$. Note furthermore that because $h(n,T) = o(\sqrt{D_P(\alpha,p)/n})$, $\mathcal{C}_k \cap \mathcal{C}_l = \emptyset$ for $k \neq l$ and sufficiently large n,T. By (38), it is then impossible that the centers $z_1^*,\dots,z_K^*$ are outliers if n,T are sufficiently large (we have shown the existence of at least K disjoint sets that would be selected through maximization before any outlier would be considered for promotion to center). Specifically we have that for $k = 1,\dots,K$

$$\exists z \in (\cup_{l=1}^K \mathcal{C}_l) \setminus \cup_{l=0}^{k-1} S_l : |\mathcal{N}_z| \geq |\mathcal{C}^{(k)}| \geq n\alpha_k(1-o_{\mathbb{P}}(1)),$$

where the $|\mathcal{C}^{(1)}| \geq \ldots \geq |\mathcal{C}^{(K)}|$ denote the order statistic for the core cardinalities and $\alpha^{(1)} \geq \ldots \geq \alpha^{(K)}$ denote the order statistic for the cluster concentrations. Thus for sufficiently large n,T there exists a permutation γ such that

$$\|\hat{R}_{z_k^*,\cdot} - \bar{P}_{\gamma(k)}\|_2 < ah(n,T) \quad \text{for} \quad k = 1,\dots,K. \tag{44}$$

Finally, let $x \in \mathcal{E}$ be any misclassified state. Necessarily $x \notin \mathcal{N}_{z^*_{\sigma(x)}}$, for otherwise x would not be misclassified. If $x \in \mathcal{N}_{z_c^*}$ for some $c \neq \sigma(x)$, we have $\|\hat{R}_{x,\cdot} - \bar{P}_c\|_2 \leq (1+a)h(n,T)$ by (44) and thus

$$\|\hat{R}_{x,\cdot} - \bar{P}_{\sigma(x)}\|_2 \overset{\text{(i)}}{\geq} |\|\hat{R}_{x,\cdot} - \bar{P}_c\|_2 - \|\bar{P}_c - \bar{P}_{\sigma(x)}\|_2| \geq \sqrt{\frac{D_P(\alpha,p)}{n}} - (1+a)h(n,T).$$

Since $h(n,T) = o(\sqrt{D_P(\alpha,p)/n})$, the result in Lemma 6 follows. If $x \in (\cup_{k=1}^K \mathcal{N}_{z_k^*})^{\mathrm{c}}$, the algorithm has associated x to the closest (but incorrect) center via (39), i.e., to some cluster $c \neq \sigma(x)$ satisfying $\|\hat{R}_{z_c^*,\cdot} - \hat{R}_{x,\cdot}\|_2 \leq \|\hat{R}_{z^*_{\sigma(x)},\cdot} - \hat{R}_{x,\cdot}\|_2$. Since each center z_k^* is $ah(n,T)$ close to its truth $\bar{P}_k$, which themselves are at least $\Omega(\sqrt{D_P(\alpha,p)/n})$ apart, it must be that $\|\hat{R}_{x,\cdot} - \bar{P}_{\sigma(x)}\|_2 = \Omega(\sqrt{D_P(\alpha,p)/n})$. This completes the proof. □

Step 4: Separability in $\hat{R}$ implies an upper bound on $|\mathcal{E}|$.

Proof. The final step to prove Theorem 2 is almost immediate. Since (i) by Lemma 5 and (ii) strict positivity of the summands and Lemma 6, if $\|\hat{P}-P\| = o_{\mathbb{P}}(g)$ for some $g = o(1)$ and $\omega(g^2/n) = h = o(1/n)$, then

$$\begin{aligned} K(1+\sqrt{2})^2\|\hat{P}-P\|^2 &\overset{\text{(i)}}{\geq} \|\hat{R}-P\|_{\mathrm{F}}^2 = \sum_{x\in\mathcal{V}} \|\hat{R}_{x,\cdot} - P_{x,\cdot}\|_2^2 \\ &= \sum_{x\in\mathcal{E}} \|\hat{R}_{x,\cdot} - P_{x,\cdot}\|_2^2 + \sum_{x\in\mathcal{V}\setminus\mathcal{E}} \|R_{x,\cdot} - P_{x,\cdot}\|_2^2 \overset{\text{(ii)}}{=} \Omega_{\mathbb{P}}\Big(|\mathcal{E}|\frac{D_P(\alpha,p)}{n}\Big). \end{aligned} \tag{45}$$

It must therefore be that Theorem 2 holds for otherwise (45) would be contradictory. □

When $\hat{N}$ is used instead of $\hat{P}$. To establish the analogous result for when $\hat{N}$ is used instead of $\hat{P}$, we first establish that N also satisfies a separability result. All remaining arguments then follow analogously.

Lemma 7. *For any $x,y \in \mathcal{V}$ for which $\sigma(x) \neq \sigma(y)$,*

$$\|N_{x,\cdot} - N_{y,\cdot}\|_2 = \Omega\Big(\sqrt{\frac{T^2 D_N(\alpha,p)}{n^3}}\Big).$$

PROOF. By (i) definition $N_{x,y} = T\Pi_x P_{x,y}$, and (ii) by the definition of $P_{x,y}$, recall (1),

$$\|N_{x,\cdot} - N_{y,\cdot}\|_2^2 = \sum_{z\in\mathcal{V}} |N_{x,z} - N_{y,z}|^2 \overset{\text{(i)}}{=} \sum_{z\in\mathcal{V}} |T\Pi_x P_{x,z} - T\Pi_y P_{y,z}|^2$$
$$\overset{\text{(ii)}}{=} T^2 \sum_{k=1}^{K} \sum_{z\in\mathcal{V}_k} \Big|\bar{\Pi}_{\sigma(x)} \frac{p_{\sigma(x),k}}{|\mathcal{V}_k| - \mathbb{1}[\sigma(x)=k]} - \bar{\Pi}_{\sigma(y)} \frac{p_{\sigma(y),k}}{|\mathcal{V}_k| - \mathbb{1}[\sigma(y)=k]}\Big|^2$$

we have that

$$\|N_{x,\cdot} - N_{y,\cdot}\|_2^2 \sim \frac{T^2}{n^3} \sum_{k=1}^{K} \frac{1}{\alpha_k} \Big(\frac{\pi_{\sigma(x)} p_{\sigma(x),k}}{\alpha_{\sigma(x)}} - \frac{\pi_{\sigma(y)} p_{\sigma(y),k}}{\alpha_{\sigma(y)}}\Big)^2 \geq \frac{T^2}{n^3} D_N(\alpha, p)$$

asymptotically. This completes the proof. □

6. The Cluster Improvement Algorithm. The Cluster Improvement Algorithm, whose pseudo-code is presented in Algorithm 2, aims at sequentially improving the cluster assignment identified by the Spectral Clustering Algorithm. In each iteration, it works as follows. Given a cluster assignment $\{\hat{\mathcal{V}}_k^{[t]}\}_{k=1,\dots,K}$ obtained after the t-th iteration, it first calculates the estimates

$$\hat{p}_{a,b} = \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[a=b]}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \sum_{x\in\hat{\mathcal{V}}_a^{[t]}} \sum_{y\in\hat{\mathcal{V}}_b^{[t]}} \hat{P}_{x,y} = \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[a=b]}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \hat{P}_{\hat{\mathcal{V}}_a^{[t]},\hat{\mathcal{V}}_b^{[t]}} \quad \text{for} \quad a,b = 1,\dots,K,$$

$$\hat{\pi}_k = \frac{1}{T} \sum_{x\in\hat{\mathcal{V}}_k^{[t]}} \sum_{y\in\mathcal{V}} \hat{N}_{x,y} = \frac{1}{T} \hat{N}_{\hat{\mathcal{V}}_k^{[t]},\mathcal{V}} \quad \text{and} \quad \hat{\alpha}_k = \frac{|\hat{\mathcal{V}}_k^{[t]}|}{n} \quad \text{for} \quad k = 1,\dots,K. \tag{46}$$

It then initializes $\hat{\mathcal{V}}_k^{[t+1]} = \emptyset$ for $k = 1,\dots,K$, and assigns each vertex $x = 1,\dots,n$ to $\mathcal{V}_{c_x^{\text{opt}}}^{[t+1]} \leftarrow \mathcal{V}_{c_x^{\text{opt}}}^{[t+1]} \cup \{x\}$, where

$$c_x^{\text{opt}} \triangleq \arg\max_{c=1,\dots,K} u_x^{[t]}(c), \quad \text{and} \quad u_x^{[t]}(c) \triangleq \Big\{\sum_{k=1}^{K} (\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} \ln \hat{p}_{c,k} + \hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} \ln \frac{\hat{p}_{k,c}}{\hat{\alpha}_c}) - \frac{T}{n} \cdot \frac{\hat{\pi}_c}{\hat{\alpha}_c}\Big\}. \tag{47}$$

This results in a new cluster assignment $\{\hat{\mathcal{V}}_k^{[t+1]}\}_{k=1,\dots,K}$.

The algorithm works by placing each state in the cluster it most likely belongs to, based on the known structure and the sample path. This can be seen by noting that the objective function in (47) is the difference between two log-likelihood functions, as we discuss in Appendix G. If the initial cluster assignment $\{\hat{\mathcal{V}}_k^{[0]}\}_{k=1,\dots,K}$ is sufficiently close to the ground truth we can expect that $|\mathcal{E}^{[t]}| \to 0$ as $t \to \infty$. It turns out that the Spectral Clustering Algorithm provides such sufficiently close initial cluster assignment. Theorem 3 formally states our result, and provides sufficient conditions under which the number of misclassifications decreases with each iteration.

THEOREM 3. *If $I(\alpha,p) > 0$, $\exists_{0<\eta\neq 1}$: $\max_{a,b,c=1,\dots,K}\{p_{b,a}/p_{c,a}, p_{a,b}/p_{a,c}\} \leq \eta$, $T = \omega(n)$, $|\mathcal{E}^{[t]}| = O_{\mathbb{P}}(e_n^{[t]})$ for some $0 < e_n^{[t]} = o(n)$, $\|\hat{N} - N\| = O_{\mathbb{P}}(f(n,T))$ for some $f(n,T) = o(T/n)$, $\|\hat{P} - P\| = O_{\mathbb{P}}(g(n,T))$ for some $g(n,T) = o(1)$, and $|\mathcal{E}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, then*

$$e_n^{[t+1]} = O\Big(e_n^{[t]} \Big(\frac{n}{T} f(n,T)\Big)^2\Big) = o(e_n^{[t]}). \tag{48}$$

As a consequence of the above theorem, if we initialize the Cluster Improvement Algorithm with a cluster assignment for which $|\mathcal{E}^{[0]}| = o_{\mathbb{P}}(n)$, we find by iterating (48) that after $\tau \in \mathbb{N}_+$ steps

$$|\mathcal{E}^{[\tau]}| = O_{\mathbb{P}}\Big(n\Big(\frac{n}{T} f(n,T)\Big)^{2\tau}\Big).$$

In Section 8, we conjecture that a $f(n,T) = o(T/n)$ exists such that $\|\hat{N} - N\| = O_{\mathbb{P}}(f(n,T))$ whenever $T = \omega(n)$. The improvement step then improves the cluster assignment after each iteration when n, T are sufficiently large. Furthermore, Section 8 contains numerical results that suggest that $\|\hat{N} - N\| \approx O_{\mathbb{P}}(\sqrt{T/n})$, i.e., that a function $f(n,T) \approx \sqrt{T/n}$ would suffice. If this is indeed the correct asymptotic scaling, it would for example follow for the case $T = n\ln(n/s)$ that $|\mathcal{E}^{[\tau]}| \approx O_{\mathbb{P}}(n(1/\ln(n/s))^{\tau})$. That would imply that $|\mathcal{E}^{[\tau^*]}| \approx o_{\mathbb{P}}(s)$ after as few as $\tau^* \approx \ln(n/s)/(\ln\ln(n/s)) = O(\ln n/s)$ iterations.

6.1. *Proof of Theorem 3.* The proof of Theorem 3 consists in first observing that after the $(t+1)$-st iteration, for any misclassified state x, its true cluster $\sigma(x)$ does not maximize the objective function $u_x^{[t]}(c)$. Hence summing over all misclassified states, we get

$$E \triangleq \sum_{x \in \mathcal{E}^{[t+1]}} \big(u_x^{[t]}(\sigma^{[t+1]}(x)) - u_x^{[t]}(\sigma(x))\big) \geq 0.$$

Next the proof proceed in two steps.

Step 1. We show through concentration arguments that asymptotically $E \approx -(T/n)I(\alpha,p)|\mathcal{E}^{[t+1]}| + \|\hat{N} - N\|\sqrt{|\mathcal{E}^{[t+1]}||\mathcal{E}^{[t]}|}$.

Step 2. For sufficiently large n, T, putting the result of Step 1 together with the aforementioned suboptimality $E \geq 0$ yields Theorem 3.

Step 1. Substituting $u_x^{[t]}$'s definition from (47), we obtain after simplifying

$$E = \sum_{x \in \mathcal{E}^{[t+1]}} \Big[\sum_{k=1}^{K} \Big(\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} \ln \frac{\hat{p}_{\sigma^{[t+1]}(x),k}}{\hat{p}_{\sigma(x),k}} + \hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} \ln \frac{\hat{p}_{k,\sigma^{[t+1]}(x)}}{\hat{p}_{k,\sigma(x)}}\Big) + \Big(\frac{\hat{N}_{\hat{\mathcal{V}}_{\sigma(x)}^{[t]},\mathcal{V}}}{|\hat{\mathcal{V}}_{\sigma(x)}^{[t]}|} - \frac{\hat{N}_{\hat{\mathcal{V}}_{\sigma^{[t+1]}(x)}^{[t]},\mathcal{V}}}{|\hat{\mathcal{V}}_{\sigma^{[t+1]}(x)}^{[t]}|}\Big)\Big].$$

Next, we split $E = E_1 + E_2 + E_3 + E_4$ into different terms, each centered around a different object that is expected to concentrate. Specifically, we define $E_1 = E_1^{\text{out}} + E_1^{\text{in}} + E_1^{\text{cross}}$ with

$$E_1^{\text{out}} = \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \hat{N}_{x,\mathcal{V}_k} \ln \frac{p_{\sigma^{[t+1]}(x),k}}{p_{\sigma(x),k}}, \quad E_1^{\text{in}} = \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \hat{N}_{\mathcal{V}_k,x} \ln \frac{p_{k,\sigma^{[t+1]}(x)}}{p_{k,\sigma(x)}},$$

$$E_1^{\text{cross}} = \sum_{x \in \mathcal{E}^{[t+1]}} \Big(\frac{\hat{N}_{\mathcal{V}_{\sigma(x)},\mathcal{V}}}{|\mathcal{V}_{\sigma(x)}|} - \frac{\hat{N}_{\mathcal{V}_{\sigma^{[t+1]}(x)},\mathcal{V}}}{|\mathcal{V}_{\sigma^{[t+1]}(x)}|}\Big)$$

as well as $E_2 = E_2^{\text{out}} + E_2^{\text{in}}$ with

$$E_2^{\text{out}} = \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} (\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} - \hat{N}_{x,\mathcal{V}_k}) \ln \frac{p_{\sigma^{[t+1]}(x),k}}{p_{\sigma(x),k}},$$

$$E_2^{\text{in}} = \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} (\hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} - \hat{N}_{\mathcal{V}_k,x}) \ln \frac{p_{k,\sigma^{[t+1]}(x)}}{p_{k,\sigma(x)}},$$

and $E_3 = E_3^{\text{out}} + E_3^{\text{in}}$ with

$$E_3^{\text{out}} = \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} \Big(\ln \frac{\hat{p}_{\sigma^{[t+1]}(x),k}}{\hat{p}_{\sigma(x),k}} - \ln \frac{p_{\sigma^{[t+1]}(x),k}}{p_{\sigma(x),k}}\Big),$$

$$E_3^{\text{in}} = \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} \Big(\ln \frac{\hat{p}_{k,\sigma^{[t+1]}(x)}}{\hat{p}_{k,\sigma(x)}} - \ln \frac{p_{k,\sigma^{[t+1]}(x)}}{p_{k,\sigma(x)}}\Big),$$

and finally

$$E_4 = \sum_{x\in\mathcal{E}^{[t+1]}} \Big(\frac{\hat{N}_{\hat{\mathcal{V}}^{[t]}_{\sigma(x)},\mathcal{V}}}{|\hat{\mathcal{V}}^{[t]}_{\sigma(x)}|} - \frac{\hat{N}_{\mathcal{V}_{\sigma(x)},\mathcal{V}}}{|\mathcal{V}_{\sigma(x)}|} \Big) - \sum_{x\in\mathcal{E}^{[t+1]}} \Big(\frac{\hat{N}_{\hat{\mathcal{V}}^{[t]}_{\sigma^{[t+1]}(x)},\mathcal{V}}}{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[t+1]}(x)}|} - \frac{\hat{N}_{\mathcal{V}_{\sigma^{[t+1]}(x)},\mathcal{V}}}{|\mathcal{V}_{\sigma^{[t+1]}(x)}|} \Big). \tag{49}$$

We proceed by bounding the terms E_1, E_2, E_3 and E_4.

LEMMA 8. *If $|\mathcal{E}^{[t]}| = O_{\mathbb{P}}(e_n^{[t]})$, $\|\hat{N} - N\| = O_{\mathbb{P}}(f(n,T))$ for some $f(n,T) = o(T/n)$, and $|\mathcal{E}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, then*

$$|E_2| = O_{\mathbb{P}}\Big(\frac{T}{n}\frac{e_n^{[t]}}{n} e_n^{[t+1]} + f(n,T)\sqrt{e_n^{[t]} e_n^{[t+1]}} \Big).$$

PROOF. By assumption there exists a constant $\eta > 0$ so that $p_{b,a}/p_{c,a} \leq \eta$ for all $a, b, c = 1, \ldots, K$. With this constant it holds moreover that for all $a, b, c = 1, \ldots, K$, $p_{c,a}/p_{b,a} \geq 1/\eta$. By the triangle inequality $|E_2| \leq |E_2^{\text{out}}| + |E_2^{\text{in}}|$, and

$$|E_2^{\text{out}}| \leq \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} |\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} - \hat{N}_{x,\mathcal{V}_k}| \Big| \ln \frac{p_{\sigma^{[t+1]}(x),k}}{p_{\sigma(x),k}} \Big| \leq |\ln\eta| \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} |\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} - \hat{N}_{x,\mathcal{V}_k}|.$$

Similarly

$$|E_2^{\text{in}}| \leq \Big|\ln\frac{1}{\eta}\Big| \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} |\hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} - \hat{N}_{\mathcal{V}_k,x}|.$$

Recall that $|\ln\eta| = |\ln(1/\eta)|$. Thus

$$|E_2| \leq |\ln\eta| \Big(\sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} |\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} - \hat{N}_{x,\mathcal{V}_k}| + \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} |\hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} - \hat{N}_{\mathcal{V}_k,x}| \Big). \tag{50}$$

Next we deal with the summations within the brackets. By (i) the definition of $\hat{N}_{\mathcal{A},\mathcal{B}}$, and (ii) the triangle inequality and strict positivity of the entries $\hat{N}_{x,y}$

$$\sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} |\hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} - \hat{N}_{x,\mathcal{V}_k}| \overset{\text{(i)}}{=} \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \Big| \sum_{y\in\hat{\mathcal{V}}_k^{[t]}} \hat{N}_{x,y} - \sum_{y\in\mathcal{V}_k} \hat{N}_{x,y} \Big|$$

$$= \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \Big| \sum_{y\in\hat{\mathcal{V}}_k^{[t]}\setminus\mathcal{V}_k} \hat{N}_{x,y} - \sum_{y\in\mathcal{V}_k\setminus\hat{\mathcal{V}}_k^{[t]}} \hat{N}_{x,y} \Big| \overset{\text{(ii)}}{\leq} \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \sum_{y\in\hat{\mathcal{V}}_k^{[t]}\Delta\mathcal{V}_k} \hat{N}_{x,y} = 2 \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{y\in\mathcal{E}^{[t]}} \hat{N}_{x,y}.$$

Aside from swapping the indices, the conclusion holds similarly for the second summation in (50). We thus conclude that

$$|E_2| \leq 2|\ln\eta| \Big(\sum_{x\in\mathcal{E}^{[t+1]}} \sum_{y\in\mathcal{E}^{[t]}} \hat{N}_{x,y} + \sum_{x\in\mathcal{E}^{[t+1]}} \sum_{y\in\mathcal{E}^{[t]}} \hat{N}_{y,x} \Big) = 2|\ln\eta| (\hat{N}_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} + \hat{N}_{\mathcal{E}^{[t]},\mathcal{E}^{[t+1]}}).$$

We next center both terms around their means. Since the Markov chain is in equilibrium by assumption, it holds for the first term that

$$\begin{aligned} \hat{N}_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} &= N_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} + \hat{N}_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} - N_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} \\ &\leq \max_{x,y}\{T\Pi_x P_{x,y}\} |\mathcal{E}^{[t]}||\mathcal{E}^{[t+1]}| + \hat{N}_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} - N_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}}. \end{aligned}$$

Then after applying Lemma 17 (iii), see Appendix H, we find that

$$\hat{N}_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} - N_{\mathcal{E}^{[t+1]},\mathcal{E}^{[t]}} = 1_{\mathcal{E}^{[t+1]}}{}^{\mathrm{T}}(\hat{N}-N)1_{\mathcal{E}^{[t]}} \overset{\text{(iii)}}{\leq} \|\hat{N}-N\|\sqrt{|\mathcal{E}^{[t]}||\mathcal{E}^{[t+1]}|}. \tag{51}$$

The same conclusion holds for $\hat{N}_{\mathcal{E}^{[t]},\mathcal{E}^{[t+1]}} - N_{\mathcal{E}^{[t]},\mathcal{E}^{[t+1]}}$.

Summarizing, we have so far shown that

$$|E_2| \leq 4|\ln\eta|\Big(\max_{x,y}\{T\Pi_x P_{x,y}\}|\mathcal{E}^{[t]}||\mathcal{E}^{[t+1]}| + \|\hat{N}-N\|\sqrt{|\mathcal{E}^{[t]}||\mathcal{E}^{[t+1]}|}\Big).$$

By recalling that $\Pi_x P_{x,y} = O(1/n^2)$ and applying Lemma 18, see Appendix I, the result is finally proven. □

LEMMA 9. *There exists an absolute constant c (independent of n) such that $\mathrm{Var}_\Phi[\hat{N}_{x,\mathcal{V}_k}] \leq c(T\ln T)/n$ for every $k = 1,\ldots,K$ and all $x \in \mathcal{V}_k$.*

PROOF. Proving the statement uses the same techniques as those applied in Proposition 6. To see this, let $z \in \mathcal{V}$ and write

$$\hat{N}_{z,\mathcal{V}_k} = \sum_{y\in\mathcal{V}_k}\hat{N}_{x,y} = \sum_{s=1}^{T}\sum_{y\in\mathcal{V}_k}\mathbb{1}[X_{s-1}=x, X_s=y] = \sum_{s=1}^{T}\mathbb{1}[X_{s-1}=x, X_s\in\mathcal{V}_k].$$

Then note that

$$\mathrm{Var}_\Phi[\hat{N}_{z,\mathcal{V}_k}] = \sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{Cov}_\Phi[\mathbb{1}[X_{t-1}=z, X_t\in\mathcal{V}_k], \mathbb{1}[X_{s-1}=z, X_s\in\mathcal{V}_k]]$$

has the same form as (25). The result now follows after applying the exact same steps. In particular, one concludes that when $|t-s|\geq 2$,

$$\begin{aligned}&\mathrm{Cov}_\Phi[\mathbb{1}[X_{t-1}=x, X_t\in\mathcal{V}_k], \mathbb{1}[X_{s-1}=x, X_s\in\mathcal{V}_k]]\\ &= \sum_{x,y,u,v}\Pi_x P_{x,y}(P_{y,u}^{|t-s|-1} - \Pi_u)P_{u,v}\mathbb{1}[x=z, y\in\mathcal{V}_k, u=z, v\in\mathcal{V}_k]\end{aligned}$$

and subsequently

$$\begin{aligned}&\Big|\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{1}[|t-s|\geq 2]\mathrm{Cov}_\Phi[\mathbb{1}[X_{t-1}=x, X_t\in\mathcal{V}_k], \mathbb{1}[X_{s-1}=x, X_s\in\mathcal{V}_k]]\Big|\\ &\leq \frac{c}{n^3}\sum_{t=1}^{T}\sum_{s=t+2}^{T}d(|t-s|-1)\sum_{x,y,u,v}\mathbb{1}[x=z, y\in\mathcal{V}_k, u=z, v\in\mathcal{V}_k]\end{aligned}$$

for some absolute constant c. Then using the filter effect $\sum_{x,y,u,v}\mathbb{1}[x=z, y\in\mathcal{V}_k, u=z, v\in\mathcal{V}_k] = O(n^2)$ and continuing the same arguments proves the statement. □

LEMMA 10. *If $T=\omega(n)$, $\|\hat{N}-N\| = O_{\mathbb{P}}(f(n,T))$ for some $f(n,T) = o(T/n)$, and $|\mathcal{E}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, then*

$$-E_1 = \Omega_{\mathbb{P}}\Big(I(\alpha,p)\frac{T}{n}e_n^{[t+1]}\Big).$$

PROOF. We (i) center and use the facts that $N_{x,\mathcal{V}_k} = (T/n)((\pi_{\sigma(x)}p_{\sigma(x),k})/\alpha_{\sigma(x)})$, $N_{\mathcal{V}_k,x} = (T/n)((\pi_k p_{k,\sigma(x)})/\alpha_{\sigma(x)})$ to write

$$
\begin{aligned}
&- E_1 \\
&\overset{\text{(i)}}{=} \frac{T}{n}\sum_{x\in\mathcal{E}^{[t+1]}}\Big[\sum_{k=1}^{K}\Big(\frac{\pi_{\sigma(x)}p_{\sigma(x),k}}{\alpha_{\sigma(x)}}\ln\frac{p_{\sigma(x),k}}{p_{\sigma^{[t+1]}(x),k}} + \frac{\pi_k p_{k,\sigma(x)}}{\alpha_{\sigma(x)}}\ln\frac{p_{k,\sigma(x)}}{p_{k,\sigma^{[t+1]}(x)}}\Big) + \Big(\frac{\pi_{\sigma(x)}}{\alpha_{\sigma(x)}} - \frac{\pi_{\sigma^{[t+1]}(x)}}{\alpha_{\sigma^{[t+1]}(x)}}\Big)\Big] \\
&\quad + \sum_{x\in\mathcal{E}^{[t+1]}}\Big\{\sum_{k=1}^{K}\Big((\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k})\ln\frac{p_{\sigma(x),k}}{p_{\sigma^{[t+1]}(x),k}} + (\hat{N}_{\mathcal{V}_k,x} - N_{\mathcal{V}_k,x})\ln\frac{p_{k,\sigma(x)}}{p_{k,\sigma^{[t+1]}(x)}}\Big)\Big\} \\
&\quad - \sum_{x\in\mathcal{E}^{[t+1]}}\Big(\frac{\hat{N}_{\mathcal{V}_{\sigma(x)},\mathcal{V}} - N_{\mathcal{V}_{\sigma(x)},\mathcal{V}}}{|\mathcal{V}_{\sigma(x)}|} + \frac{N_{\mathcal{V}_{\sigma^{[t+1]}(x)},\mathcal{V}} - \hat{N}_{\mathcal{V}_{\sigma^{[t+1]}(x)},\mathcal{V}}}{|\mathcal{V}_{\sigma^{[t+1]}(x)}|}\Big).
\end{aligned}
$$

We first deal with the sum with square brackets. Note that for each $x \in \mathcal{E}^{[t+1]}$, the summand is lower bounded by $I(\alpha, p)$, recall definition (5). Furthermore $I(\alpha, p) > 0$ by assumption. This implies that

$$
\frac{T}{n}\sum_{x\in\mathcal{E}^{[t+1]}}\Big[\cdots\Big] \geq \frac{T}{n}|\mathcal{E}^{[t+1]}|I(\alpha,p). \tag{52}
$$

Together with $|\mathcal{E}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$ conclude that $(T/n)\sum_{x\in\mathcal{E}^{[t+1]}}[\cdots] = \Omega_{\mathbb{P}}(I(\alpha,p)(T/n)e_n^{[t+1]})$.

We proceed by bounding the sum involving the curly brackets. By (ii) using the same steps that proved (50), conclude that

$$
\Big|\sum_{x\in\mathcal{E}^{[t+1]}}\sum_{k=1}^{K}(\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k})\ln\frac{p_{\sigma(x),k}}{p_{\sigma^{[t+1]}(x),k}}\Big| \overset{\text{(ii)}}{\leq} |\ln\eta|\sum_{k=1}^{K}\sum_{x\in\mathcal{E}^{[t+1]}}|\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}|.
$$

Now we prepare to apply Lemma 20, see Appendix I. Let $k \in \{1,\dots,K\}$ and identify $X_{m,n} \equiv |\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}|$ and $Y_{m,n} \equiv |\mathcal{E}^{[t+1]}|$. We have by (iv) strict positivity and (v) Jensen's inequality that

$$
\mathbb{E}[|\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}|] \overset{\text{(iv)}}{=} \sqrt{\mathbb{E}[|\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}|]^2} \overset{\text{(v)}}{\leq} \sqrt{\mathbb{E}[|\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}|^2]} = \sqrt{\mathrm{Var}[\hat{N}_{x,\mathcal{V}_k}]}
$$

for all $k = 1,\dots,K$ and $x \in \mathcal{V}_k$. Therefore by Lemma 9 there exists an absolute constant c (independent of n) such that $\mathbb{E}[|\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}|] \leq c\sqrt{(T\ln T)/n}$ for all $k = 1,\dots,K$ and $x \in \mathcal{V}_k$. By assumption $|\mathcal{E}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, so $|\mathcal{E}^{[t+1]}| = O_{\mathbb{P}}(e_n^{[t+1]})$. The prerequisites of Lemma 20 are thus met, and hence

$$
\sum_{k=1}^{K}\sum_{x\in\mathcal{E}^{[t+1]}}|\hat{N}_{x,\mathcal{V}_k} - N_{x,\mathcal{V}_k}| = O_{\mathbb{P}}\Big(e_n^{[t+1]}\sqrt{\frac{T\ln T}{n}}\Big).
$$

The terms involving $\hat{N}_{\mathcal{V}_k,x} - N_{\mathcal{V}_k,x}$ are dealt with similarly, and we conclude that $|\sum_{x\in\mathcal{E}^{[t+1]}}\{\cdots\}| = O_{\mathbb{P}}(e_n^{[t+1]}\sqrt{(T\ln T)/n})$. By assumption $T = \omega(n)$, implying in particular that $\sqrt{T\ln T/n} = o(T/n)$. Hence this sum is asymptotically negligible compared to (52).

What remains is to deal with the sum with round brackets. Similar to the derivation of (51), we have that $|\hat{N}_{\mathcal{V}_{\sigma(x)},\mathcal{V}} - N_{\mathcal{V}_{\sigma(x)},\mathcal{V}}|/|\mathcal{V}_{\sigma(x)}| \leq (n/|\mathcal{V}_{\sigma(x)}|)^{1/2}\|\hat{N} - N\|$ for each $x \in \mathcal{E}^{[t+1]}$ and $|N_{\mathcal{V}_{\sigma^{[t+1]}(x)},\mathcal{V}} - \hat{N}_{\mathcal{V}_{\sigma^{[t+1]}(x)},\mathcal{V}}|/|\mathcal{V}_{\sigma^{[t+1]}(x)}| \leq (n/|\mathcal{V}_{\sigma^{[t+1]}(x)}|)^{1/2}\|\hat{N} - N\|$ for each $x \in \mathcal{E}^{[t+1]}$. Since $|\mathcal{V}_k| \sim n\alpha_k$, we conclude that

$$
\Big|\sum_{x\in\mathcal{E}^{[t+1]}}\Big(\cdots\Big)\Big| \leq \frac{2}{\min_{k=1,\dots,K}\sqrt{\alpha_k}}|\mathcal{E}^{[t+1]}|\|\hat{N} - N\|.
$$

Applying Lemma 18 gives $|\sum_{x\in\mathcal{E}^{[t+1]}}(\cdots)| = O_{\mathbb{P}}(e_n^{[t+1]}f(n,T))$. By assumption $f(n,T) = o(T/n)$, so this sum is asymptotically negligible compared to (52). This completes the proof. □

LEMMA 11. *If* $|\mathcal{E}^{[t]}| = O_{\mathbb{P}}(e_n^{[t]})$, $\|\hat{P} - P\| = O_{\mathbb{P}}(g(n,T))$ *for some* $g(n,T) = o(1)$, *and* $|\mathcal{E}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, *then*

$$|E_3| = O_{\mathbb{P}}\Big(\frac{T}{n} g(n,T) e_n^{[t+1]} + \frac{T}{n}\frac{e_n^{[t]}}{n} e_n^{[t+1]}\Big).$$

PROOF. By the triangle inequality, we have $E_3 \leq |E_3| \leq |E_3^{\text{in}}| + |E_3^{\text{out}}|$ with

$$|E_3^{\text{in}}| \leq \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \hat{N}_{x,\hat{\mathcal{V}}_k^{[t]}} \Big(\Big| \ln \frac{\hat{p}_{\sigma^{[t+1]}(x),k}}{p_{\sigma^{[t+1]}(x),k}} \Big| + \Big| \ln \frac{\hat{p}_{\sigma(x),k}}{p_{\sigma(x),k}} \Big| \Big),$$

$$|E_3^{\text{out}}| \leq \sum_{x \in \mathcal{E}^{[t+1]}} \sum_{k=1}^{K} \hat{N}_{\hat{\mathcal{V}}_k^{[t]},x} \Big(\Big| \ln \frac{\hat{p}_{k,\sigma^{[t+1]}(x)}}{p_{k,\sigma^{[t+1]}(x)}} \Big| + \Big| \ln \frac{\hat{p}_{k,\sigma(x)}}{p_{k,\sigma(x)}} \Big| \Big). \tag{53}$$

We now bound the summands. From the inequalities $x/(1+x) \leq \ln(1+x) \leq x$ for $x > -1$, it follows that for $a, b = 1, \ldots, K$,

$$\Big| \ln \frac{\hat{p}_{a,b}}{p_{a,b}} \Big| = \Big| \ln \Big(1 + \frac{\hat{p}_{a,b} - p_{a,b}}{p_{a,b}} \Big) \Big| \leq \max \Big\{ \Big| \frac{\hat{p}_{a,b} - p_{a,b}}{\hat{p}_{a,b}} \Big|, \Big| \frac{\hat{p}_{a,b} - p_{a,b}}{p_{a,b}} \Big| \Big\}. \tag{54}$$

We proceed by bounding the numerator $|\hat{p}_{a,b} - p_{a,b}|$. By (i) definition of $\hat{p}_{a,b}$ and the block structure of $P_{x,y}$, recall (46) and (1), respectively, and (ii) $|\mathcal{V}_b| - \mathbb{1}[b=a] \leq |\mathcal{V}_b|$ and centering,

$$\begin{aligned}
|\hat{p}_{a,b} - p_{a,b}| &\overset{\text{(i)}}{=} \Big| \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[b=a]}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \sum_{x \in \hat{\mathcal{V}}_a^{[t]}} \sum_{y \in \hat{\mathcal{V}}_b^{[t]}} \hat{P}_{x,y} - \frac{|\mathcal{V}_b| - \mathbb{1}[b=a]}{|\mathcal{V}_a||\mathcal{V}_b|} \sum_{x \in \mathcal{V}_a} \sum_{y \in \mathcal{V}_b} P_{x,y} \Big| \\
&= \frac{|\mathcal{V}_b| - \mathbb{1}[b=a]}{|\mathcal{V}_a||\mathcal{V}_b|} \Big| \frac{|\mathcal{V}_a||\mathcal{V}_b|}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[b=a]}{|\mathcal{V}_b| - \mathbb{1}[b=a]} \sum_{x \in \hat{\mathcal{V}}_a^{[t]}} \sum_{y \in \hat{\mathcal{V}}_b^{[t]}} \hat{P}_{x,y} - \sum_{x \in \mathcal{V}_a} \sum_{y \in \mathcal{V}_b} P_{x,y} \Big| \\
&\overset{\text{(ii)}}{\leq} \frac{1}{|\mathcal{V}_a|} \Big| \sum_{x \in \hat{\mathcal{V}}_a^{[t]}} \sum_{y \in \hat{\mathcal{V}}_b^{[t]}} \hat{P}_{x,y} - \sum_{x \in \mathcal{V}_a} \sum_{y \in \mathcal{V}_b} P_{x,y} \Big| + \frac{\hat{P}_{\hat{\mathcal{V}}_a^{[t]},\hat{\mathcal{V}}_b^{[t]}}}{|\mathcal{V}_a|} \Big| \frac{|\mathcal{V}_a||\mathcal{V}_b|}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[b=a]}{|\mathcal{V}_b| - \mathbb{1}[b=a]} - 1 \Big|.
\end{aligned} \tag{55}$$

Note that

$$\Big| \frac{|\mathcal{V}_a||\mathcal{V}_b|}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \cdot \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[b=a]}{|\mathcal{V}_b| - \mathbb{1}[b=a]} - 1 \Big| = \begin{cases} \Big| \frac{|\mathcal{V}_a|}{|\hat{\mathcal{V}}_a^{[t]}|} - 1 \Big| & \text{if } b \neq a, \\ \Big| \frac{|\mathcal{V}_a| - |\mathcal{V}_a|/|\hat{\mathcal{V}}_a^{[t]}|}{|\hat{\mathcal{V}}_a^{[t]}| - |\hat{\mathcal{V}}_a^{[t]}|/|\mathcal{V}_a|} - 1 \Big| & \text{otherwise.} \end{cases}$$

For the case $b \neq a$,

$$\Big| \frac{|\mathcal{V}_a|}{|\hat{\mathcal{V}}_a^{[t]}|} - 1 \Big| = \frac{||\mathcal{V}_a| - |\hat{\mathcal{V}}_a^{[t]}||}{|\hat{\mathcal{V}}_a^{[t]}|} \leq |\mathcal{E}^{[t]}|.$$

For the case $b = a$, temporarily adopt the notation $\hat{v} = |\hat{\mathcal{V}}_a^{[t]}|$ and $v = |\mathcal{V}_a|$, note that there exists a constant c_1 so that

$$\Big| \frac{v - v/\hat{v}}{\hat{v} - \hat{v}/v} - 1 \Big| = \Big| \frac{v - \hat{v} + \hat{v}/v - v/\hat{v}}{\hat{v} - \hat{v}/v} \Big| \leq |v - \hat{v}| \Big(\Big| \frac{1}{\hat{v} - \hat{v}/v} \Big| + \Big| \frac{\hat{v} + v}{\hat{v}^2 v - \hat{v}^2} \Big| \Big) \leq c_1 |\mathcal{E}^{[t]}|.$$

Finally note that $|\mathcal{V}_a| \sim n\alpha_a$ and $\hat{P}_{\hat{\mathcal{V}}_a^{[t]},\hat{\mathcal{V}}_b^{[t]}} \leq 1$ for $a, b = 1, \ldots, K$. Therefore there exists a constant c_2 so that

$$\frac{\hat{P}_{\hat{\mathcal{V}}_a^{[t]},\hat{\mathcal{V}}_b^{[t]}}}{|\mathcal{V}_a|} \Big| \frac{|\mathcal{V}_a||\mathcal{V}_b|}{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|} \frac{|\hat{\mathcal{V}}_b^{[t]}| - \mathbb{1}[b=a]}{|\mathcal{V}_b| - \mathbb{1}[b=a]} - 1 \Big| \leq c_2 \frac{|\mathcal{E}^{[t]}|}{n}.$$

Continuing again from (55), we find using (iii) the triangle inequality and (iv) Lemma 17, see Appendix H, that

$$\begin{aligned}|\hat{p}_{a,b}-p_{a,b}| &\overset{\text{(iii)}}{\leq} \frac{1}{|\mathcal{V}_a|}\Big(\Big|\sum_{x\in\hat{\mathcal{V}}_a^{[t]}}\sum_{y\in\hat{\mathcal{V}}_b^{[t]}}(\hat{P}_{x,y}-P_{x,y})\Big|+\Big|\sum_{x\in\hat{\mathcal{V}}_a^{[t]}}\sum_{y\in\hat{\mathcal{V}}_b^{[t]}}P_{x,y}-\sum_{x\in\mathcal{V}_a}\sum_{y\in\mathcal{V}_b}P_{x,y}\Big|\Big)+c_2\frac{|\mathcal{E}^{[t]}|}{n}\\ &\overset{\text{(iv)}}{\leq} \frac{1}{|\mathcal{V}_a|}\Big(\|\hat{P}-P\|\sqrt{|\hat{\mathcal{V}}_a^{[t]}||\hat{\mathcal{V}}_b^{[t]}|}+\sum_{(x,y)\in(\hat{\mathcal{V}}_a^{[t]}\times\hat{\mathcal{V}}_b^{[t]})\Delta(\mathcal{V}_a\times\mathcal{V}_b)}|P_{x,y}|\Big)+c_2\frac{|\mathcal{E}^{[t]}|}{n}\\ &\leq \frac{1}{\alpha_a}\|\hat{P}-P\|+\frac{1}{n\alpha_a}\sum_{(x,y)\in(\hat{\mathcal{V}}_a^{[t]}\times\hat{\mathcal{V}}_b^{[t]})\Delta(\mathcal{V}_a\times\mathcal{V}_b)}|P_{x,y}|+c_2\frac{|\mathcal{E}^{[t]}|}{n}.\end{aligned}$$

We next bound the cardinality:

$$\begin{aligned}|(\hat{\mathcal{V}}_a^{[t]}\times\hat{\mathcal{V}}_b^{[t]})\Delta(\mathcal{V}_a\times\mathcal{V}_b)| &= |\hat{\mathcal{V}}_a^{[t]}\Delta\mathcal{V}_a||\hat{\mathcal{V}}_b^{[t]}\cup\mathcal{V}_b|+|\hat{\mathcal{V}}_a^{[t]}\cap\mathcal{V}_a||\hat{\mathcal{V}}_b^{[t]}\Delta\mathcal{V}_b|\\ &= |\hat{\mathcal{V}}_a^{[t]}\Delta\mathcal{V}_a|(|\hat{\mathcal{V}}_b^{[t]}\Delta\mathcal{V}_b|+|\hat{\mathcal{V}}_b^{[t]}\cap\mathcal{V}_b|)+|\hat{\mathcal{V}}_a^{[t]}\cap\mathcal{V}_a||\hat{\mathcal{V}}_b^{[t]}\Delta\mathcal{V}_b|\\ &\leq 4|\mathcal{E}_a^{[t]}||\mathcal{E}_b^{[t]}|+2|\mathcal{E}_a^{[t]}||\mathcal{V}_b|+2|\mathcal{V}_a||\mathcal{E}_b^{[t]}|\leq 8n|\mathcal{E}^{[t]}|.\end{aligned}\tag{56}$$

Summarizing, since $P_{x,y}=O(1/n)$ there exists a constant c_1 so that

$$|\hat{p}_{a,b}-p_{a,b}| = c_1\|\hat{P}-P\|+c_2\frac{|\mathcal{E}^{[t]}|}{n}.\tag{57}$$

From Lemma 9 it follows that for all $k=1,\dots,K$ and $x\in\mathcal{V}$, $\hat{N}_{x,\mathcal{V}_k}=O_{\mathbb{P}}(T/n)$ and $\hat{N}_{\mathcal{V}_k,x}=O_{\mathbb{P}}(T/n)$. By using this fact, bounding (53) using (57) via (54), and then applying Lemma 18, the argument is finished. □

Lemma 12. *If $|\mathcal{E}^{[t]}|=O_{\mathbb{P}}(e_n^{[t]})$, $\|\hat{N}-N\|=O_{\mathbb{P}}(f(n,T))$ for some $f(n,T)=o(T/n)$, and $|\mathcal{E}^{[t+1]}|\asymp_{\mathbb{P}} e_n^{[t+1]}$, then*

$$|E_4| = O_{\mathbb{P}}\Big(\frac{T}{n}\frac{e_n^{[t]}}{n}e_n^{[t+1]}+f(n,T)\sqrt{\frac{e_n^{[t]}}{n}}e_n^{[t+1]}\Big).$$

Proof. Let $k\in\{1,\dots,K\}$ to examine any one of the summands in E_4. We (i) center and use the triangle inequality to bound all summands as

$$\Big|\frac{\hat{N}_{\hat{\mathcal{V}}_k^{[t]},\mathcal{V}}}{|\hat{\mathcal{V}}_k^{[t]}|}-\frac{\hat{N}_{\mathcal{V}_k,\mathcal{V}}}{|\mathcal{V}_k|}\Big| \overset{\text{(i)}}{\leq} \frac{1}{|\mathcal{V}_k|}|\hat{N}_{\hat{\mathcal{V}}_k^{[t]},\mathcal{V}}-\hat{N}_{\mathcal{V}_k,\mathcal{V}}|+\frac{\hat{N}_{\hat{\mathcal{V}}_k^{[t]},\mathcal{V}}}{|\mathcal{V}_k|}\Big|\frac{|\mathcal{V}_k|}{|\hat{\mathcal{V}}_k^{[t]}|}-1\Big|.\tag{58}$$

The left term in (58) can be bounded (ii) using the arguments of (56) and (iii) Lemma 17

$$\begin{aligned}|\hat{N}_{\hat{\mathcal{V}}_k^{[t]},\mathcal{V}}-\hat{N}_{\mathcal{V}_k,\mathcal{V}}| &\leq \sum_{(x,y)\in(\hat{\mathcal{V}}_k^{[t]}\times\mathcal{V})\Delta(\mathcal{V}_k\times\mathcal{V})}(N_{x,y}+\hat{N}_{x,y}-N_{x,y})\\ &\overset{\text{(ii,iii)}}{\leq} 2\max_{x,y}\{T\Pi_x P_{x,y}\}n|\mathcal{E}^{[t]}|+\sum_{x\in\hat{\mathcal{V}}_k^{[t]}\Delta\mathcal{V}_k}\sum_{y\in\mathcal{V}}(\hat{N}_{x,y}-N_{x,y})\\ &\overset{\text{(iv)}}{\leq} c_1\frac{T}{n}|\mathcal{E}^{[t]}|+\|\hat{N}-N\|\sqrt{2|\mathcal{E}^{[t]}|n},\end{aligned}$$

where (iv) we have used that $T\Pi_x P_{x,y}=O(T/n^2)$. The right term in (58) can be bounded using $\hat{N}_{\hat{\mathcal{V}}_k^{[t]},\mathcal{V}}\leq T$, and $||\mathcal{V}_k|/|\hat{\mathcal{V}}_k^{[t]}|-1|\leq|\mathcal{E}^{[t]}|/|\hat{\mathcal{V}}_k^{[t]}|\leq c_2|\mathcal{E}^{[t]}|/n$ for some constant c_2. Using these bounds together with (58), (49), and $|\mathcal{V}_k|\sim n\alpha_k$ shows that there exist constants c_3,c_4 such that

$$|E_4|\leq c_3\frac{T}{n}\frac{|\mathcal{E}^{[t]}|}{n}|\mathcal{E}^{[t+1]}|+c_4\|\hat{N}-N\|\sqrt{\frac{|\mathcal{E}^{[t]}|}{n}}|\mathcal{E}^{[t+1]}|.$$

Using Lemma 18 then completes the proof. □

Step 2. Under the assumptions of Theorem 3, Lemmas 8–12 imply

$$-\frac{n}{T}E_1 = \Omega_{\mathbb{P}}\Big(I(\alpha,p)e_n^{[t+1]}\Big)$$

and

$$\begin{aligned}&-\frac{n}{T}\Big(|E_2|+|E_3|+|E_4|\Big)\\&= O_{\mathbb{P}}\Big(\frac{e_n^{[t]}}{n}e_n^{[t+1]}+\frac{n}{T}f(n,T)\sqrt{e_n^{[t]}e_n^{[t+1]}}+g(n,T)e_n^{[t+1]}+\frac{n}{T}f(n,T)\sqrt{\frac{e_n^{[t]}}{n}}e_n^{[t+1]}\Big).\end{aligned}$$

Additionally, recall that $E = E_1+E_2+E_3+E_4 \geq 0$, i.e., $-E_1 \leq E_2+E_3+E_4 \leq |E_2|+|E_3|+|E_4|$ almost surely. The prerequisites of Lemma 21, see Appendix I, are therefore met, so necessarily

$$I(\alpha,p)e_n^{[t+1]} = O\Big(\frac{e_n^{[t]}}{n}e_n^{[t+1]}+\frac{n}{T}f(n,T)\sqrt{e_n^{[t]}e_n^{[t+1]}}+g(n,T)e_n^{[t+1]}+\frac{n}{T}f(n,T)\sqrt{\frac{e_n^{[t]}}{n}}e_n^{[t+1]}\Big).$$

Note that equality holds when $e_n^{[t+1]} = 0$. When $e_n^{[t+1]} > 0$, and recall that $e_n^{[t]} > 0$ by assumption, we can divide by $(e_n^{[t]}e_n^{[t+1]})^{1/2}$ to obtain

$$I(\alpha,p)\sqrt{\frac{e_n^{[t+1]}}{e_n^{[t]}}} = O\Big(\frac{n}{T}f(n,T)+\sqrt{\frac{e_n^{[t+1]}}{e_n^{[t]}}}\Big[\frac{e_n^{[t]}}{n}+g(n,T)+\frac{n}{T}f(n,T)\sqrt{\frac{e_n^{[t]}}{n}}\Big]\Big).$$

Since $e_n^{[t]} = o(n)$, $f(n,T) = o(T/n)$, and $g(n,T) = o(1)$, conclude that

$$\sqrt{\frac{e_n^{[t+1]}}{e_n^{[t]}}} = O\Big(\frac{n}{T}f(n,T)\Big).$$

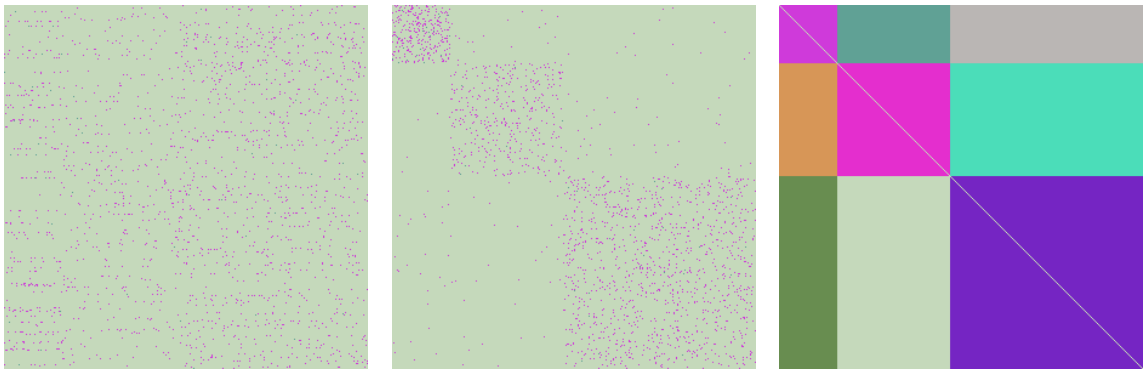
This completes the proof of Theorem 3.

7. Numerical experiments. In this section, we numerically assess the performance of our algorithms. We first investigate a simple illustrative example. Then we study the sensitivity of the error rate of the Spectral Clustering Algorithm w.r.t. the number of states and the length of the observed trajectory. Finally we show the performance of the Cluster Improvement Algorithm depending on the number of its iterations.

7.1. *An example.* Consider $n = 300$ states grouped into three clusters of respective relative sizes $\alpha = (0.15, 0.35, 0.5)$, i.e., the cluster sizes are cluster sizes $|\mathcal{V}_1| = 48$, $|\mathcal{V}_2| = 93$ and $|\mathcal{V}_3| = 159$. The transition rates between these clusters are defined by:

$$p = \begin{pmatrix} 0.92 & 0.045 & 0.035 \\ 0.0125 & 0.8975 & 0.09 \\ 0.0175 & 0.02 & 0.9625 \end{pmatrix}.$$

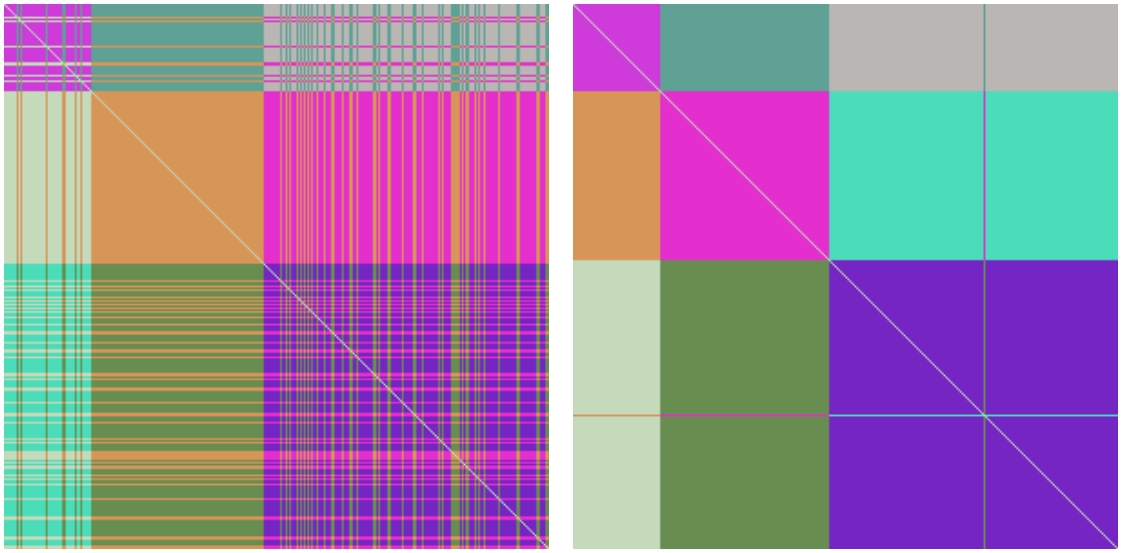
We generate a sample path of the Markov chain of length $T = n^{1.025}\ln n \approx 1973$ and calculate $\hat{N}$. A density plot of a typical sample of $\hat{N}$ is shown in Figure 3a. The same density plot is presented in Figure 3b where the states have been sorted so as states in the same cluster are neighbors. It is important to note that the algorithms are of course not aware of the structure initially – sorting states constitutes their objective. Next in Figure 3c, we show a color representation of the kernel P with sorted rows and columns, in which we can clearly see the groups. Note that the specific colors have no meaning, except for the fact that within the same image two entries with the same color have the same numerical value.



(a) $\hat{N}$, unsorted (b) $\hat{N}$, sorted (c) P, sorted

Fig 3: A sample path of length $T = n^{1.025} \ln n \approx 1973$ was generated, from which $\hat{N}$ is calculated. If we sort the vertices according to the clusters they belong to, we can see that vertices within the same cluster share similar dynamics.

Next we apply the Spectral Clustering Algorithm. This generates an initial approximate clustering $\hat{\mathcal{V}}_1^{[0]}, \hat{\mathcal{V}}_2^{[0]}, \hat{\mathcal{V}}_3^{[0]}$ of the vertices. We generate a visual representation of this clustering by constructing a BMC kernel $\hat{P}^{[0]}$ from the approximate cluster structure and the estimate $\hat{p}^{[0]}$. This represents the belief that the algorithm has at this point of the true BMC kernel P. A color representation of this kernel is shown in Figure 4a. We finally execute the Cluster Improvement Algorithm. After 3 iterations, it has settled on a final clustering. We generate a color representation of the clustering similar to before, resulting in Figure 4b. The algorithms achieved a 99.7% accuracy: all but one state have been accurately clustered.



(a) Initial clustering. (b) Final clustering.

Fig 4: (a) Result after applying the Spectral Clustering Algorithm to the approximation $\hat{N}$. (b) Result after applying 3 iterations of the Cluster Improvement Algorithm. 99.7% of all states were accurately clustered.

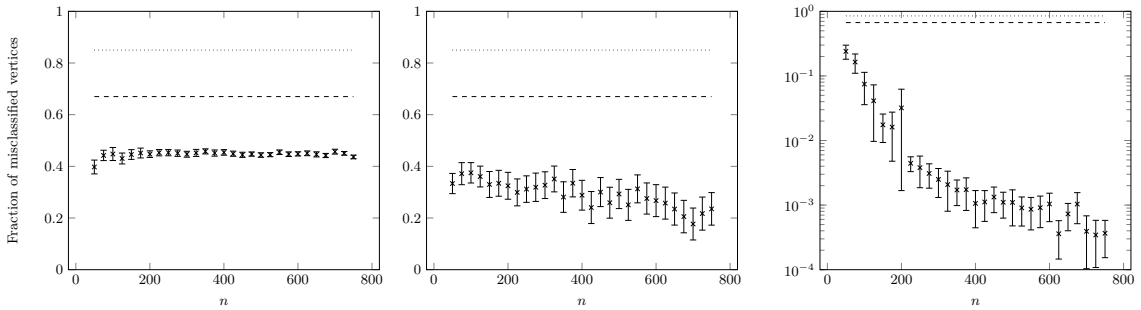
7.2. *Performance sensitivity of the Spectral Clustering Algorithm.* In this section, we examine the dependency of the number of misclassified states on the size of the kernel n, when we only

apply the Spectral Clustering Algorithm. We choose $\alpha = (0.15, 0.35, 0.5)$, and set

$$p = \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.1 & 0.7 & 0.2 \\ 0.35 & 0.05 & 0.6 \end{pmatrix}.$$

These parameters imply that $I(\alpha, p) \approx 0.88 > 0$. This value for $I(\alpha, p)$ is lower than in the case examined in Section 7.1, so we expect clustering to be more difficult. We have selected a more challenging model so that the initial number of misclassified states will be large and the asymptotics clear.

Figure 5 displays the error rate of the Spectral Clustering algorithm as a function of n, for different trajectory lengths T. As benchmarks, we include a dashed line that indicates the error rate obtained by assigning states to clusters uniformly at random, i.e., $\mathbb{P}[v \notin \mathcal{V}_{\sigma(v)}] = \sum_{k=1}^{K} \mathbb{P}[v \notin \mathcal{V}_k | \sigma(v) = k]\alpha_k = 1 - 1/K$, as well as a dotted line that indicates the error rate when assigning all states to the smallest cluster, i.e., $1 - \min_k\{\alpha_k\}$. For the K-means step of the algorithms, we use Mathematica's default implementation for convenience. Observe that when $T = n \ln n$, the fraction of misclassified states hardly decrease as a function of n. This is in line with our lower bound. When T gets larger, the error converges to zero faster. Note that the Spectral Clustering Algorithm recovers the clusters exactly when the sample path is sufficiently long.



(a) $T = n \ln n$ (b) $T = n(\ln n)^{3/2}$ (c) $T = n(\ln n)^2$

Fig 5: The error rate of the Spectral Clustering Algorithm as function of n, for different scalings of T. Every point is the average result of 40 simulations, and the bars indicate a 95%-confidence interval.

7.3. *Performance sensitivity of the Spectral Clustering Improvement algorithm.* We now examine the number of misclassified states as a function of T, when we apply the Spectral Clustering Algorithm and a certain number of iterations of the Cluster Improvement Algorithm. We choose $\alpha = (1/3, 1/3, 1/3)$, and set

$$p = \begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.1 & 0.2 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}.$$

Different from the previous experiments, the clusters are now of equal size and the off-diagonal entries of p are dominant. These parameters imply that $I(\alpha, p) \approx 0.27 > 0$, so the cluster algorithms should work, but the situation is again more challenging than in Section 7.1 and Section 7.2.

Figure 6 depicts the error after applying the Spectral Clustering Algorithm and subsequently the Cluster Improvement Algorithm up to two times, as a function of T. We have chosen both n, T relatively small so that the inputs are significantly noisy. For short sample paths, $T \lesssim 15000$, the data is so noisy that the Cluster Improvement Algorithm does not provide any improvement

over the Spectral Clustering Algorithm. For $T \gtrsim 15000$, the Spectral Clustering Algorithm provides a sufficiently accurate initial clustering for the Cluster Improvement Algorithm to work. Because marks 1 and 2 overlap in almost all cases, we can conclude that there is (on average, and in the present situation) no benefit in running the Clustering Improvement Algorithm more than once. There is no mark 2 at $T = 30000$ in this logarithmic plot, because the Cluster Improvement Algorithm achieved 100% accurate detection after 2 iterations in *all* 200 simulations.

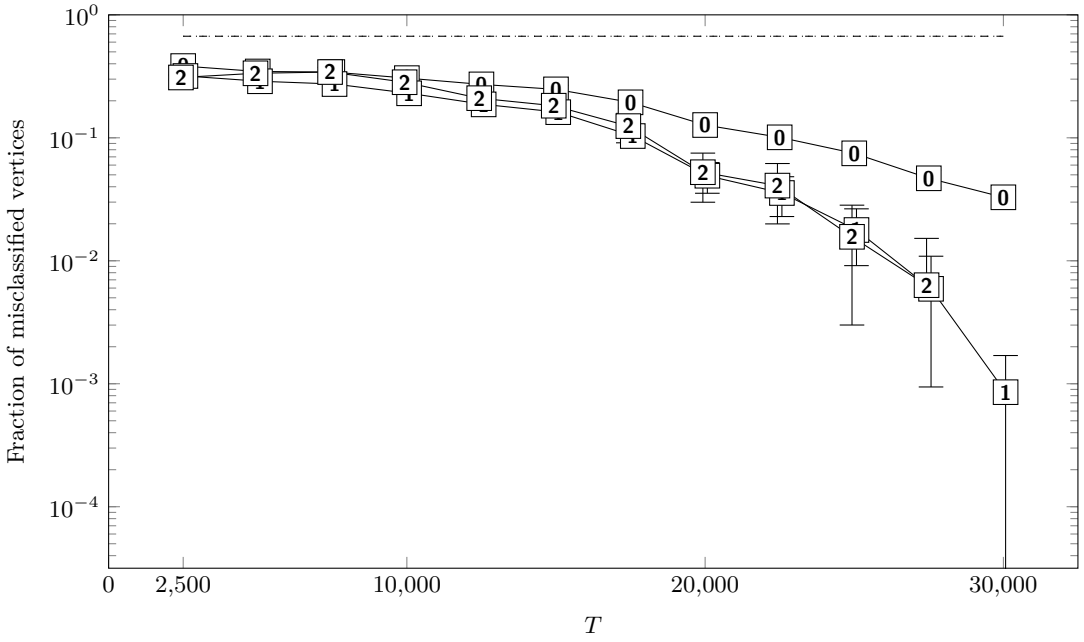


Fig 6: The error after applying the Spectral Clustering Algorithm (mark 0), and subsequently the Cluster Improvement Algorithm (marks $1, 2$) several times, as a function of T. Each number represents the number of improvement steps. Here, $n = 240$. Every point is the average result of 200 simulations, and the bars indicate a 95%-confidence interval. We have minorly offset marks $1, 2$ to the right and left for readability, respectively. At $T = 30000$, the Cluster Improvement Algorithm achieved 100% accurate detection after 2 iterations in *all* 200 instances.

8. Concentration of the spectral norm of the noise matrix. Recall that the Spectral Clustering Algorithm and Cluster Improvement Algorithm rely on calculating the matrices $\hat{N} \in \mathbb{N}_0^{n\times n}$ and $\hat{P} \in \mathbb{\Delta}^{(n-1)\times n}$ element-wise as

$$\hat{N}_{x,y} = \sum_{t=0}^{T-1} \mathbb{1}[X_t = x, X_{t+1} = y], \quad \text{and} \quad \hat{P}_{x,y} = \frac{\sum_{t=0}^{T-1} \mathbb{1}[X_t = x, X_{t+1} = y]}{\sum_{t=0}^{T-1} \mathbb{1}[X_t = x]} \quad \text{for} \quad x, y \in \mathcal{V}.$$

In the convergence proofs of the algorithms, we encountered the spectral norms of the noise matrices $\hat{N} - N$ and $\hat{P} - P$. Specifically, we require that $\|\hat{N} - N\|$ is atleast $o_{\mathbb{P}}(T/n)$ and $\|\hat{P} - P\|$ atleast $o_{\mathbb{P}}(1)$. The primary difficulty in proving these statements though is that $\hat{N}$ constitutes a random matrix with stochastic *dependent* entries. While concentration of the eigenvalues of a random matrix has been actively investigated when the entries are independent or satisfy a weak condition of dependence [22–27], or when the transition matrix of the Markov chain itself is random [28, 29], we were unable to find work relating to the case when the entries are dictated by a Markov chain with a fixed transition matrix with a block structure.

We therefore examined proof strategies to obtain bounds on the concentration of $\|\hat{N} - N\|$ and $\|\hat{P} - P\|$, each with varying degrees of success. These included an attempt to reduce the problem to a case of independent entries [30], applying an epsilon-net argument [23], Wigner's trace method [22, 31], Stein's method for concentration inequalities [32], generalized Hoeffding

inequalities for Markov chains [33–35], matrix perturbation techniques [36–39], and a method of exchangeable pairs [40]. We give this (nonexhaustive) list of approaches to facilitate future research, because while we have so far been unable to prove the following conjecture, we believe it holds and warrants further investigation.

CONJECTURE. *If the Markov chain* $\{X_t\}_{t\geq 0}$ *is a BMC and* $T = \omega(n)$*, then there exist functions* $f(n,T) = o(T/n)$ *and* $g(n,T) = o(1)$ *such that* $\|\hat{N} - N\| = O_{\mathbb{P}}(f(n,T))$ *and* $\|\hat{P} - P\| = O_{\mathbb{P}}(g(n,T))$*, respectively.*

We verified numerically that the above conjecture holds. In Figure 7, we plot $\|\hat{N} - N\|$ as a function of the number n of states. The results reported there suggest that the conjecture holds for $f(n,T) \approx \sqrt{T/n}$.

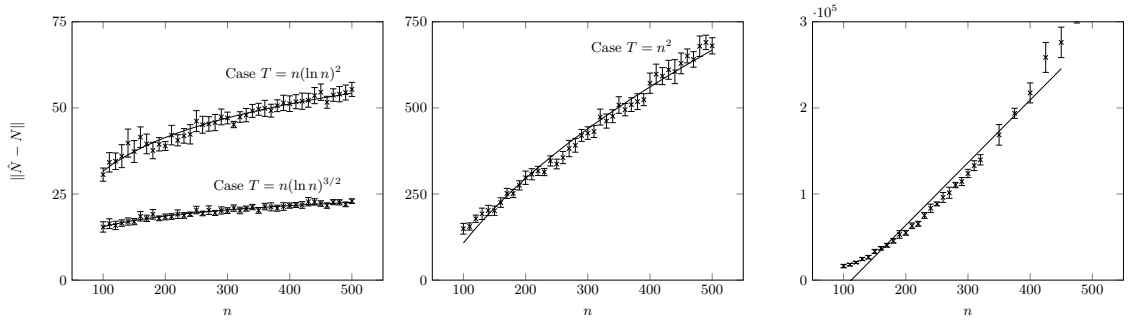


Fig 7: Simulations of the asymptotic behavior of $\|\hat{N} - N\|$, together with best fits of the form $c_1 + c_2\sqrt{T/n}$ for different scalings of T. The simulations suggest that $f(n,T) \approx \sqrt{T/n} = o(T/n)$.

Bibliography.

[1] P. W. Holland, K. B. Laskey, and S. Leinhardt. "Stochastic blockmodels: First steps". In: *Social networks* 5.2 (1983), pp. 109–137.

[2] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. en. American Mathematical Soc., 2009.

[3] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998.

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

[5] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. "Inference and phase transitions in the detection of modules in sparse networks". In: *Physical Review Letters* 107.6 (2011), p. 065701.

[6] L. Massoulié. "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM. 2014, pp. 694–703.

[7] E. Mossel, J. Neeman, and A. Sly. "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162.3-4 (2015), pp. 431–461.

[8] S.-Y. Yun and A. Proutiere. "Community Detection via Random and Adaptive Sampling." In: *COLT*. 2014, pp. 138–175.

[9] S.-Y. Yun and A. Proutiere. "Optimal cluster recovery in the labeled stochastic block model". In: *Advances in Neural Information Processing Systems*. 2016, pp. 965–973.

[10] S.-Y. Yun and A. Proutiere. "Accurate community detection in the stochastic block model via spectral algorithms". In: *arXiv preprint arXiv:1412.7335* (2014).

[11] E. Abbe and C. Sandon. "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery". In: *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE. 2015, pp. 670–688.

[12] E. Abbe and C. Sandon. "Recovering communities in the general stochastic block model without knowing the parameters". In: *Advances in neural information processing systems*. 2015, pp. 676–684.

[13] V. Jog and P.-L. Loh. "Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence". In: *arXiv preprint arXiv:1509.06418* (2015).

[14] E. Mossel, J. Neeman, and A. Sly. "Consistency thresholds for the planted bisection model". In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM. 2015, pp. 69–75.

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

[16] B. Hajek, Y. Wu, and J. Xu. "Achieving exact cluster recovery threshold via semidefinite programming". In: *IEEE Transactions on Information Theory* 62.5 (2016), pp. 2788–2797.

[17] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[18] M. Mihail. "Conductance and Convergence of Markov Chains-a Combinatorial Treatment of Expanders". In: *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*. SFCS '89. Washington, DC, USA: IEEE Computer Society, 1989, pp. 526–531. DOI: 10.1109/SFCS.1989.63529. URL: https://doi.org/10.1109/SFCS.1989.63529.

[19] T. Lai and H. Robbins. "Asymptotically efficient adaptive allocation rules". In: *Advances in Applied Mathematics* 6.1 (1985), pp. 4 –22. DOI: https://doi.org/10.1016/0196-8858(85)90002-8. URL: http://www.sciencedirect.com/science/article/pii/0196885885900028.

[20] S. Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[21] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Review* 53.2 (2011), pp. 217–288. DOI: 10.1137/090771806. URL: http://dx.doi.org/10.1137/090771806.

[22] E. P. Wigner. "On the distribution of the roots of certain symmetric matrices". In: *Annals of Mathematics* (1958), pp. 325–327.

[23] T. Tao. *Topics in random matrix theory*. Vol. 132. American Mathematical Society Providence, RI, 2012.

[24] J. A. Tropp et al. "An introduction to matrix concentration inequalities". In: *Foundations and Trends® in Machine Learning* 8.1-2 (2015), pp. 1–230.

[25] W. Hochstättler, W. Kirsch, and S. Warzel. "Semicircle law for a matrix ensemble with dependent entries". In: *Journal of Theoretical Probability* 29.3 (2016), pp. 1047–1068.

[26] W. Kirsch and T. Kriecherbauer. "Sixty years of moments for random matrices". In: *arXiv preprint arXiv:1612.06725* (2016).

[27] W. Kirsch and T. Kriecherbauer. "Semicircle law for generalized Curie-Weiss matrix ensembles at subcritical temperature". In: *arXiv preprint arXiv:1703.05183* (2017).

[28] C. Bordenave, P. Caputo, and D. Chafai. "Spectrum of large random reversible Markov chains: two examples". In: *ALEA: Latin American Journal of Probability and Mathematical Statistics* 7 (2010), pp. 41–64.

[29] C. Bordenave, P. Caputo, D. Chafai, et al. "Spectrum of large random reversible Markov chains: heavy-tailed weights on the complete graph". In: *The Annals of Probability* 39.4 (2011), pp. 1544–1590.

[30] A. Coja-oghlan. "Graph Partitioning via Adaptive Spectral Techniques". In: *Comb. Probab. Comput.* 19.2 (Mar. 2010), pp. 227–284. DOI: 10.1017/S0963548309990514. URL: http://dx.doi.org/10.1017/S0963548309990514.

[31] V. Vu. "Spectral norm of random matrices". In: *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. ACM. 2005, pp. 423–430.

[32] S. Chatterjee. "Stein's method for concentration inequalities". In: *Probability Theory and Related Fields* 138.1 (2007), pp. 305–321. DOI: 10.1007/s00440-006-0029-y. URL: https://doi.org/10.1007/s00440-006-0029-y.

[33] P. W. Glynn and D. Ormoneit. "Hoeffding's inequality for uniformly ergodic Markov chains". In: *Statistics & probability letters* 56.2 (2002), pp. 143–146.

[34] T. R. Boucher. "A Hoeffding inequality for Markov chains using a generalized inverse". In: *Statistics & Probability Letters* 79.8 (2009), pp. 1105–1107.

[35] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher. "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified". In: *arXiv preprint arXiv:1201.0559* (2012).

[36] F. L. Bauer and C. T. Fike. "Norms and exclusion theorems". In: *Numerische Mathematik* 2.1 (1960), pp. 137–141.

[37] K.-w. E. Chu. "Generalization of the Bauer-Fike theorem". In: *Numerische Mathematik* 49.6 (1986), pp. 685–691.

[38] M. Haviv, Y. Ritov, and U. G. Rothblum. "Taylor expansions of eigenvalues of perturbed matrices with applications to spectral radii of nonnegative matrices". In: *Linear algebra and its applications* 168 (1992), pp. 159–188.

[39] C. D. Meyer. "Sensitivity of the stationary distribution of a Markov chain". In: *SIAM Journal on Matrix Analysis and Applications* 15.3 (1994), pp. 715–728.

[40] L. Mackey, M. I. Jordan, R. Y. Chen, B. Farrell, J. A. Tropp, et al. "Matrix concentration inequalities via the method of exchangeable pairs". In: *The Annals of Probability* 42.3 (2014), pp. 906–945.

APPENDIX A: PROPERTIES OF UNIFORM VERTEX SELECTION

LEMMA 13. *If a state V^* is selected uniformly at random from two specific clusters $a, b \in \{1, \dots, K\}$, $a \neq b$, and a state V is selected uniformly at random from all states,*

$$\mathbb{P}_\Phi[V^* \in \mathcal{E}] = \mathbb{P}_\Phi[V \in \mathcal{E} | V \in \mathcal{V}_a \cup \mathcal{V}_b].$$

PROOF. We have:

$$\mathbb{P}_\Phi[V^* \in \mathcal{E}] = \sum_{v \in \mathcal{V}_a \cup \mathcal{V}_b} \mathbb{P}_\Phi[V^* \in \mathcal{E} | V^* = v] \mathbb{P}_\Phi[V^* = v] = \frac{1}{|\mathcal{V}_a| + |\mathcal{V}_b|} \sum_{v \in \mathcal{V}_a \cup \mathcal{V}_b} \mathbb{P}_\Phi[v \in \mathcal{E}],$$

and

$$\begin{aligned}\mathbb{P}_\Phi[V \in \mathcal{E} | V \in \mathcal{V}_a \cup \mathcal{V}_b] &= \frac{\sum_{v \in \mathcal{V}} \mathbb{P}_\Phi[V \in \mathcal{E}, V \in \mathcal{V}_a \cup \mathcal{V}_b | V = v] \mathbb{P}_\Phi[V = v]}{\mathbb{P}_\Phi[V \in \mathcal{V}_a \cup \mathcal{V}_b]} \\ &= \frac{\sum_{v \in \mathcal{V}_a \cup \mathcal{V}_b} \mathbb{P}_\Phi[v \in \mathcal{E}]/|\mathcal{V}|}{(|\mathcal{V}_a| + |\mathcal{V}_b|)/|\mathcal{V}|} = \frac{1}{|\mathcal{V}_a| + |\mathcal{V}_b|} \sum_{v \in \mathcal{V}_a \cup \mathcal{V}_b} \mathbb{P}_\Phi[v \in \mathcal{E}].\end{aligned}$$

The lemma follows. □

LEMMA 14. *If a state V is selected uniformly at random from all states,*

$$\mathbb{E}_\Phi[|\mathcal{E}|] = n \mathbb{P}_\Phi[V \in \mathcal{E}].$$

PROOF. We have:

$$\mathbb{E}_\Phi[|\mathcal{E}|] = \mathbb{E}_\Phi[\sum_{v \in \mathcal{V}} \mathbb{1}[v \in \mathcal{E}]] = \sum_{v \in \mathcal{V}} \mathbb{E}_\Phi[\mathbb{1}[v \in \mathcal{E}]] = \sum_{v \in \mathcal{V}} \mathbb{P}_\Phi[v \in \mathcal{E}],$$

and

$$n \mathbb{P}_\Phi[V \in \mathcal{E}] = n \sum_{v \in \mathcal{V}} \mathbb{P}_\Phi[V \in \mathcal{E} | V = v] \mathbb{P}_\Phi[V = v] = n \sum_{v \in \mathcal{V}} \mathbb{P}_\Phi[v \in \mathcal{E}] \frac{1}{|\mathcal{V}|} = \sum_{v \in \mathcal{V}} \mathbb{P}_\Phi[v \in \mathcal{E}],$$

which completes the proof. □

APPENDIX B: INEQUALITY FOR DISTRIBUTIONAL L_P NORM

LEMMA 15. *For any $p \in [1, \infty)$, there exists a constant c_p independent of n such that $d_1(\mu, \nu) \leq c_p d_p(\mu, \nu)$.*

PROOF. Let $1/q = 1 - 1/p$. We apply Hölder's inequality (i), and bound

$$\begin{aligned}d_1(\mu, \nu) = \|\mu - \nu\|_1 &\overset{\text{(i)}}{\leq} n^{1/q} \|\mu - \nu\|_p = n^{1/q} \Big(\sum_{x \in \mathcal{V}} |\mu_x - \nu_x|^p \Big)^{1/p} \\ &\leq n^{1/q} \Pi_{\max}^{1 - \frac{1}{p}} \Big(\sum_{x \in \mathcal{V}} \Big| \frac{\mu_x}{\Pi_x} - \frac{\nu_x}{\Pi_x} \Big|^p \Pi_x \Big)^{1/p} \overset{\text{(ii)}}{\leq} c_p d_p(\mu, \nu).\end{aligned}$$

Here, we have used (ii), i.e., the fact that the leading order behavior of $\Pi_{\max}$ is $1/n$. □

APPENDIX C: PROOF OF PROPOSITION 2

PROOF. Let $x \in \mathcal{V}$. We use the Markov chain concentration result in Theorem 3.2 of [18]. It is shown there that for any transition matrix P (not necessarily reversible) and $t \in \mathbb{N}_+$

$$d_2(P^t_{x,\cdot}, \Pi) \leq (1 - \tfrac{1}{2}(\Phi^*_P)^2)^{\frac{1}{2}t} d_2(P^0_{x,\cdot}, \Pi).$$

Here, Φ^*_P denotes the *merging conductance* of P defined as

$$\Phi^*_P \triangleq \min_{\{\mathcal{A}\subseteq\mathcal{V}|\sum_{z\in\mathcal{A}}\Pi_z\leq\frac{1}{2}\}} \{\Phi^*_P(\mathcal{A})\}, \quad \text{with} \quad \Phi^*_P(\mathcal{A}) \triangleq \frac{\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{V}\setminus\mathcal{A}}\sum_{z\in\mathcal{V}}\frac{\Pi_x P_{x,z}\Pi_y P_{y,z}}{\Pi_z}}{\sum_{z\in\mathcal{A}}\Pi_z}.$$

The result follows if we prove that there exists a constant $\alpha \in (0,1)$ independent of n such that $\Phi^*_P(\mathcal{A}) \geq \alpha$ for each $\mathcal{A} \subseteq \mathcal{V}$ that satisfies $\sum_{z\in\mathcal{A}}\Pi_z \leq \frac{1}{2}$. Indeed, this would imply that $\Phi^*_P \geq \alpha$ and therefore $(1 - \frac{1}{2}(\Phi^*_P)^2)^{t/2} \leq (1 - \frac{1}{2}\alpha^2)^{t/2}$, and as a consequence

$$d_2(P^t_{x,\cdot}, \Pi) \leq \varepsilon \quad \text{whenever} \quad t \geq \frac{2\ln\varepsilon}{\ln(1-\frac{1}{2}\alpha^2)}.$$

We next prove this assertion.

First, let $\mathcal{A} \subseteq \mathcal{V}$ be an arbitrary set such that $\sum_{z\in\mathcal{A}}\Pi_z \leq \frac{1}{2}$. Then

$$\begin{aligned}\Phi^*_P(\mathcal{A}) &\geq \frac{1}{\Pi_{\max}|\mathcal{A}|}\Big(\frac{(\Pi_{\min})^2}{\Pi_{\max}}|\mathcal{A}|(n-|\mathcal{A}|)n\min_{x,y,z}\{P_{x,z}P_{y,z}\}\Big)\\ &= \Big(\frac{\Pi_{\min}}{\Pi_{\max}}\Big)^2(n-|\mathcal{A}|)n\min_{x,y,z}\{P_{x,z}P_{y,z}\}.\end{aligned}$$

Since the entries $P_{x,y}$ as well as $\Pi_{\min}$, $\Pi_{\max}$ are of order $1/n$, we now need to show that $\lim_{n\to\infty}|\mathcal{A}|/n < 1$.

We give a proof by contradiction. Recall that there exists a constant $c_{\max} \geq 1$ independent of n such that $\lim_{n\to\infty}\Pi_{\max}n = c_{\max}$. In other words,

$$\forall_{\delta>0}\,\exists_{N_1\in\mathbb{N}_+} : |\Pi_{\max}n - c_{\max}| \leq \delta\,\forall_{n>N_1}.$$

Now assume that there exists a way to construct $\mathcal{A}$ such that

$$\forall_{\varepsilon\in(0,1)}\,\exists_{N_2\in\mathbb{N}_+} : \Big(\Big|\frac{|\mathcal{A}|}{n} - 1\Big| \leq \varepsilon, \tfrac{1}{2} \geq \sum_{z\in\mathcal{A}}\Pi_z\Big)\,\forall_{n>N_2}. \tag{59}$$

Since

$$\sum_{z\in\mathcal{A}}\Pi_z = 1 - \sum_{z\in\mathcal{A}^c}\Pi_z \geq 1 - \Pi_{\max}n\Big(1 - \frac{|\mathcal{A}|}{n}\Big),$$

the assumption in (59) would imply that

$$\forall_{\delta>0,\varepsilon\in(0,1)}\,\exists_{N_3=\max\{N_1,N_2\}\in\mathbb{N}_+} : \Big(\Big|\frac{|\mathcal{A}|}{n} - 1\Big| \leq \varepsilon, \tfrac{1}{2} \geq \sum_{z\in\mathcal{A}}\Pi_z \geq 1 - (c_{\max}+\delta)\varepsilon)\Big)\,\forall_{n>N_3}.$$

Specifying e.g. $\delta = c_{\max}$ (this can be chosen arbitrarily) gives

$$\forall_{\varepsilon\in(0,1)}\,\exists_{N_3\in\mathbb{N}_+} : \Big(\Big|\frac{|\mathcal{A}|}{n} - 1\Big| \leq \varepsilon, \tfrac{1}{2} \geq \sum_{z\in\mathcal{A}}\Pi_z \geq 1 - 2c_{\max}\varepsilon)\Big)\,\forall_{n>N_3}.$$

This gives the contradiction as $\varepsilon \downarrow 0$ once $\varepsilon < 1/(4c_{\max})$, and completes the proof. □

APPENDIX D: Q IS A STOCHASTIC MATRIX

To see this, observe that for $x \in \mathcal{V}\backslash\{V^*\}$,

$$\sum_{y\in\mathcal{V}} Q_{x,y} = \frac{q_{\omega(x),0}}{n} + \sum_{y\in\mathcal{W}_{\omega(x)}\backslash\{x\}} \frac{q_{\omega(x),\omega(x)}}{|\mathcal{W}_{\omega(x)}|-1} + \sum_{k=1}^{K} \mathbb{1}[k\neq\omega(x)] \sum_{y\in\mathcal{W}_k} \frac{q_{\omega(x),k}}{|\mathcal{W}_k|}$$
$$= \frac{q_{\omega(x),0}}{n} + \sum_{k=1}^{K} q_{\omega(x),k} \overset{(17)}{=} \frac{q_{\omega(x),0}}{n} + \sum_{k=1}^{K}\Big(p_{\omega(x),k} - \frac{q_{\omega(x),0}}{Kn}\Big) = 1.$$

Similarly for $x = V^*$

$$\sum_{y\in\mathcal{V}} Q_{V^*,y} = \sum_{k=1}^{K}\sum_{y\in\mathcal{W}_k} \frac{q_{0,k}}{|\mathcal{W}_k|} = \sum_{k=1}^{K} q_{0,k} \overset{(7)}{=} 1.$$

APPENDIX E: PROOF OF PROPOSITION 4

PROOF. We first show that $(\gamma_1^{[0]},\dots,\gamma_K^{[0]},\gamma_0^{[0]})$ is a probability distribution. Since (i) $\Pi^{(Q)}$ is a probability distribution

$$\sum_{k=1}^{K}\gamma_k + \gamma_0 = \lim_{n\to\infty}\Big(\sum_{k=1}^{K}|\mathcal{W}_k|\bar{\Pi}_k^{(Q)} + \Pi_{V^*}^{(Q)}\Big) = \lim_{n\to\infty}\Big(\sum_{k=1}^{K}\sum_{x\in\mathcal{W}_k}\Pi_x^{(Q)} + \Pi_{V^*}^{(Q)}\Big) = \lim_{n\to\infty}\sum_{x\in\mathcal{V}}\Pi_x^{(Q)} \overset{\text{(i)}}{=} 1.$$

Next we show that (ii) by the global balance equations for $\Pi^{(Q)}$

(60)
$$\gamma_0^{[0]} = \lim_{n\to\infty}\Pi_{V^*}^{(Q)} \overset{\text{(ii)}}{=} \lim_{n\to\infty}\sum_{x\in\mathcal{V}}\Pi_x^{(Q)}Q_{x,V^*} \overset{(18)}{=} \lim_{n\to\infty}\sum_{k=1}^{K}\sum_{x\in\mathcal{W}_k}\bar{\Pi}_k^{(Q)}\frac{q_{k,0}}{n} = \lim_{n\to\infty}\sum_{k=1}^{K}\gamma_k^{[0]}\frac{q_{k,0}}{n} = 0.$$

Now we establish that the vector $(\gamma_1^{[0]},\dots,\gamma_K^{[0]})^{\mathrm{T}}$ satisfies the balance equations $(\gamma_1^{[0]},\dots,\gamma_K^{[0]})p = (\gamma_1^{[0]},\dots,\gamma_K^{[0]})$. For $l = 1,\dots,K$

$$\gamma_l^{[0]} = \lim_{n\to\infty}|\mathcal{W}_l|\bar{\Pi}_l^{(Q)} = \lim_{n\to\infty}\sum_{y\in\mathcal{W}_l}\Pi_y^{(Q)} \overset{\text{(ii)}}{=} \lim_{n\to\infty}\sum_{y\in\mathcal{W}_l}\sum_{x\in\mathcal{V}}\Pi_x^{(Q)}Q_{x,y}$$
$$\overset{(18)}{=} \lim_{n\to\infty}\sum_{y\in\mathcal{W}_l}\Big(\sum_{k=1}^{K}\sum_{x\in\mathcal{W}_k\backslash\{y\}}\bar{\Pi}_k^{(Q)}\frac{q_{k,l}}{|\mathcal{W}_l|-\mathbb{1}[k=l]} + \Pi_{V^*}^{(Q)}\frac{q_{0,l}}{|\mathcal{W}_l|}\Big)$$
$$= \lim_{n\to\infty}\Big(\sum_{k=1}^{K}(|\mathcal{W}_k|-\mathbb{1}[k=l])\bar{\Pi}_k^{(Q)}\frac{|\mathcal{W}_l|}{|\mathcal{W}_l|-\mathbb{1}[k=l]}q_{k,l} + \Pi_{V^*}^{(Q)}q_{0,l}\Big) \overset{(17)}{=} \sum_{k=1}^{K}\gamma_k^{[0]}p_{k,l}, \tag{61}$$

where we also recalled (19). This proves the first two assertions.

The third assertion follows from (60) when multiplying the intermediate steps by n, i.e.,

$$\gamma_0^{[1]} \overset{(60)}{=} \lim_{n\to\infty} n\sum_{k=1}^{K}\gamma_k^{[0]}q_{k,0}\frac{1}{n} = \sum_{k=1}^{K}\gamma_k^{[0]}q_{k,0}. \tag{62}$$

Together with the first assertion, this completes the proof. □

APPENDIX F: ASYMPTOTIC COMPARISONS BETWEEN P AND Q'S ENTRIES

Recall that $R_{x,y} = Q_{x,y}/P_{x,y}$ for $x,y\in\mathcal{V}$.

LEMMA 16. *The following properties hold:*

(i) $R_{x,y} = 1 + n^{-1}(\mathbb{1}[\sigma(y)=\sigma(V^*)]/\alpha_{\sigma(y)} - q_{\sigma(x),0}/(p_{\sigma(x),\sigma(y)}K)) + O(n^{-2})$ *for* $x,y\neq V^*$,

(ii) $R_{x,V^*} = q_{\omega(x),0}\alpha_{\sigma(V^*)}/p_{\omega(x),\sigma(V^*)} + O(n^{-1})$ *for* $x \in \mathcal{V}\backslash\{V^*\}$,
(iii) $R_{V^*,y} = q_{0,\omega(x)}/p_{\sigma(V^*),\omega(x)} + O(n^{-1})$ *for* $y \in \mathcal{V}\backslash\{V^*\}$.

PROOF. Let $x, y \in \mathcal{V}\backslash\{V^*\}$. Using a Taylor expansion (i), we find that:

$$R_{x,y} \overset{(1,18)}{=} \frac{p_{\sigma(x),\sigma(y)} - q_{\sigma(x),0}/(Kn)}{p_{\sigma(x),\sigma(y)}} \cdot \frac{|\mathcal{V}_{\sigma(y)}| - \mathbb{1}[\sigma(x) = \sigma(y)]}{|\mathcal{V}_{\sigma(y)}| - \mathbb{1}[\sigma(y) = \sigma(V^*)] - \mathbb{1}[\sigma(x) = \sigma(y)]}$$
$$\overset{(i)}{=} 1 + \frac{1}{n}\Big(\frac{\mathbb{1}[\sigma(y) = \sigma(V^*)]}{\alpha_{\sigma(y)}} - \frac{q_{\sigma(x),0}}{p_{\sigma(x),\sigma(y)}K}\Big) + O\Big(\frac{1}{n^2}\Big).$$

Similarly for $x \in \mathcal{V}\backslash\{V^*\}$

$$R_{x,V^*} \overset{(1,18)}{=} \frac{q_{\omega(x),0}}{p_{\sigma(x),\sigma(V^*)}} \cdot \frac{|\mathcal{V}_{\sigma(V^*)}| - \mathbb{1}[\sigma(x) = \sigma(V^*)]}{n} \overset{(i)}{=} \frac{q_{\omega(x),0}\alpha_{\sigma(V^*)}}{p_{\sigma(x),\sigma(V^*)}} + O\Big(\frac{1}{n}\Big),$$

and for $y \in \mathcal{V}\backslash\{V^*\}$

$$R_{V^*,y} \overset{(1,18)}{=} \frac{q_{0,\omega(y)}}{p_{\sigma(V^*),\sigma(y)}} \cdot \frac{|\mathcal{V}_{\sigma(y)}| - \mathbb{1}[\sigma(V^*) = \sigma(y)]}{|\mathcal{W}_{\omega(y)}|} \overset{(i)}{=} \frac{q_{0,\omega(y)}}{p_{\sigma(V^*),\sigma(y)}} + O\Big(\frac{1}{n}\Big).$$

This completes the proof. □

Recall that $S_{x,y,u,v} = \ln R_{x,y} \cdot \ln R_{u,v}$ for $x, y, u, v \in \mathcal{V}$.

COROLLARY 1. *The following properties hold:*

(i) $S_{x,y,u,v} = O(n^{-2})$ *if all* $x, y, u, v \neq V^*$,
(ii) $S_{x,y,u,v} = O(n^{-1})$ *if one of* x, y, u, v *is* V^*,
(iii) $S_{x,y,u,v} = O(1)$ *if two of* $x \neq y, u \neq v$ *are* V^*.

PROOF. These properties are all direct consequences of Lemma 16, which can be seen by using the Taylor expansion $\ln(1 + x) = x + O(x^2)$ for $x \approx 0$ and expanding the product. Consider for example the case $x, y, u, v \in \mathcal{V}\backslash\{V^*\}$:

$$S_{x,y,u,v} = \ln R_{x,y} \cdot \ln R_{u,v} = \ln\Big(1 + O\Big(\frac{1}{n}\Big)\Big) \cdot \ln\Big(1 + O\Big(\frac{1}{n}\Big)\Big) = O\Big(\frac{1}{n^2}\Big).$$

The remaining cases follow similarly. □

APPENDIX G: THE OBJECTIVE IS A LOG-LIKELIHOOD FUNCTION

In this section, we show that as for the leading terms are concerned, for any vertex $x \in \mathcal{V}$ and $c \in \{1, \ldots, K\}$, maximizing (47) is equivalent to maximizing

$$u_x(c) = \ln \frac{\mathbb{P}_M[X_0 = x_0, \ldots, X_T = x_T]}{\mathbb{P}_L[X_0 = x_0, \ldots, X_T = x_T]} = \sum_{s=1}^{T} \ln \frac{M_{x_{s-1},x_s}}{L_{x_{s-1},x_s}}.$$

Here, L denotes the transition matrix of a BMC constructed from the cluster assignment $\{\hat{\mathcal{V}}_k^{[t]}\}_{k=1,\ldots,K}$, and M denotes the transition matrix of a modified BMC. Specifically, it is the transition matrix of a BMC in which the state x is moved into cluster c. Note that the conclusions in the paper do not require a formal proof of this statement, which is why we have opted to include only a rough justification here.

By construction of L and M, we have that $M_{x_{s-1},x_s} \neq L_{x_{s-1},x_s}$ only if $\{x_{s-1} = x, x_s \neq x\}$, $\{x_{s-1} \neq x, x_s \in \hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}\}$ or $\{x_{s-1} \neq x, x_s \in \hat{\mathcal{V}}^{[t]}_c\}$. Let $\sigma^{[L]}(x)$ denote the cluster of state x w.r.t.

the cluster structure used to construct L. Hence only the ratios

$$\frac{M_{x,y}}{L_{x,y}} = \frac{\hat{p}_{c,\sigma^{[L]}(y)}}{\hat{p}_{\sigma^{[L]}(x),\sigma^{[L]}(y)}} \cdot \frac{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(y)}| - \mathbb{1}[\sigma^{[L]}(x) = \sigma^{[L]}(y)]}{(|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(y)}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)] + \mathbb{1}[\sigma^{[L]}(y) = c]) - \mathbb{1}[c = \sigma^{[L]}(y)]}$$

$$= \frac{\hat{p}_{c,\sigma^{[L]}(y)}}{\hat{p}_{\sigma^{[L]}(x),\sigma^{[L]}(y)}} \cdot \frac{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(y)}| - \mathbb{1}[\sigma^{[L]}(x) = \sigma^{[L]}(y)]}{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(y)}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]} = \frac{\hat{p}_{c,\sigma^{[L]}(y)}}{\hat{p}_{\sigma^{[L]}(x),\sigma^{[L]}(y)}} \tag{63}$$

for $y \in \mathcal{V}\backslash\{x\}$,

$$\frac{M_{y,x}}{L_{y,x}} = \frac{\hat{p}_{\sigma^{[L]}(y),c}}{\hat{p}_{\sigma^{[L]}(y),\sigma^{[L]}(x)}} \cdot \frac{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]}{(|\hat{\mathcal{V}}^{[t]}_{c}| + 1) - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]}$$

$$= \frac{\hat{p}_{\sigma^{[L]}(y),c}}{\hat{p}_{\sigma^{[L]}(y),\sigma^{[L]}(x)}} \cdot \frac{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]}{|\hat{\mathcal{V}}^{[t]}_{c}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]} \cdot \frac{1}{1 + 1/(|\hat{\mathcal{V}}^{[t]}_{c}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)])}$$

$$\sim \frac{\hat{p}_{\sigma^{[L]}(y),c}}{\hat{p}_{\sigma^{[L]}(y),\sigma^{[L]}(x)}} \cdot \frac{\hat{\alpha}_{\sigma^{[L]}(x)}}{\hat{\alpha}_c} + O\Big(\frac{1}{n}\Big)$$

for $y \in \mathcal{V}\backslash\{x\}$,

$$\frac{M_{y,z}}{L_{y,z}} = \frac{\hat{p}_{\sigma^{[L]}(y),\sigma^{[L]}(x)}}{\hat{p}_{\sigma^{[L]}(y),\sigma^{[L]}(x)}} \cdot \frac{|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]}{(|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}| - 1) - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)]}$$

$$= \frac{1}{1 - 1/(|\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}| - \mathbb{1}[\sigma^{[L]}(y) = \sigma^{[L]}(x)])} \sim 1 + \frac{1}{n\hat{\alpha}_{\sigma^{[L]}(x)}} + O\Big(\frac{1}{n^2}\Big).$$

for $y \in \mathcal{V}\backslash\{x\}$, $z \in \hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}\backslash\{x\}$, and

$$\frac{M_{y,z}}{L_{y,z}} = \frac{\hat{p}_{\sigma^{[L]}(y),c}}{\hat{p}_{\sigma^{[L]}(y),c}} \cdot \frac{|\hat{\mathcal{V}}^{[t]}_{c}| - \mathbb{1}[\sigma^{[L]}(y) = c]}{(|\hat{\mathcal{V}}^{[t]}_{c}| + 1) - \mathbb{1}[\sigma^{[L]}(y) = c]}$$

$$= \frac{1}{1 + 1/(|\hat{\mathcal{V}}^{[t]}_{c}| - \mathbb{1}[\sigma^{[L]}(y) = c])} \sim 1 - \frac{1}{n\hat{\alpha}_c} + O\Big(\frac{1}{n^2}\Big). \tag{64}$$

for $y \in \mathcal{V}\backslash\{x\}$, $z \in \hat{\mathcal{V}}^{[t]}_{c}\backslash\{x\}$ differ from unity.

We now rewrite $u_x(c)$ to identify $\hat{N}$. Specifically, we have

$$u_x(c) = \sum_{s=1}^{T} \big(\mathbb{1}[x_{s-1} = x, x_s \neq x] + \mathbb{1}[x_{s-1} \neq x, x_s \in \hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)} \cup \hat{\mathcal{V}}^{[t]}_{c}]\big) \ln \frac{M_{x_{s-1},x_s}}{L_{x_{s-1},x_s}}$$

and write

$$u_x(c) = \sum_{s=1}^{T} \mathbb{1}[x_{s-1} = x] \underbrace{\mathbb{1}[x_s \neq x] \ln \frac{M_{x,x_s}}{L_{x,x_s}}}_{f(x_s)} + \sum_{s=1}^{T} \sum_{z \in \hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)} \cup \hat{\mathcal{V}}^{[t]}_{c}} \mathbb{1}[x_s = z] \underbrace{\mathbb{1}[x_{s-1} \neq x] \ln \frac{M_{x_{s-1},z}}{L_{x_{s-1},z}}}_{g_z(x_{s-1})}.$$

Then, since the summands of both terms above depend on only one variable (x_s and x_{s-1}, respectively),

$$u_x(c) = \sum_{s=1}^{T} \sum_{y \in \mathcal{V}\backslash\{x\}} \Big(\mathbb{1}[x_{s-1} = x]\mathbb{1}[x_s = y] f(y) + \sum_{z \in \hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)} \cup \hat{\mathcal{V}}^{[t]}_{c}} \mathbb{1}[x_{s-1} = y]\mathbb{1}[x_s = z] g_z(y)\Big)$$

$$= \sum_{y \in \mathcal{V}\backslash\{x\}} \hat{N}_{x,y} f(y) + \sum_{y \in \mathcal{V}\backslash\{x\}} \sum_{z \in \hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)} \cup \hat{\mathcal{V}}^{[t]}_{c}} \hat{N}_{y,z} g_z(y).$$

Substituting f and g_z's definitions, we obtain

$$u_x(c) = \sum_{y\in\mathcal{V}\setminus\{x\}} \Big(\hat{N}_{x,y} \ln \frac{M_{x,y}}{L_{x,y}} + \hat{N}_{y,x} \ln \frac{M_{y,x}}{L_{y,x}}\Big) + \sum_{y\in\mathcal{V}\setminus\{x\}} \sum_{z\in(\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)} \cup \hat{\mathcal{V}}^{[t]}_c)\setminus\{x\}} \hat{N}_{y,z} \ln \frac{M_{y,z}}{L_{y,z}}.$$

By now substituting (63)–(64), we find that by restricting our attention to the leading terms

$$\begin{aligned} u_x(c) \sim \sum_{y\in\mathcal{V}\setminus\{x\}} \Big(\hat{N}_{x,y} \ln \frac{\hat{p}_{c,\sigma^{[L]}(y)}}{\hat{p}_{\sigma^{[L]}(x),\sigma^{[L]}(y)}} + \hat{N}_{y,x} \ln \Big(\frac{\hat{p}_{\sigma^{[L]}(y),c}}{\hat{p}_{\sigma^{[L]}(y),\sigma^{[L]}(x)}} \cdot \frac{\hat{\alpha}_{\sigma^{[L]}(x)}}{\hat{\alpha}_c}\Big)\Big) \\ + \frac{T}{n} \cdot \frac{(1/T)\sum_{y\in\mathcal{V}\setminus\{x\}} \sum_{z\in\hat{\mathcal{V}}^{[t]}_{\sigma^{[L]}(x)}\setminus\{x\}} \hat{N}_{y,z}}{\hat{\alpha}_{\sigma^{[L]}(x)}} - \frac{T}{n} \cdot \frac{(1/T)\sum_{y\in\mathcal{V}\setminus\{x\}} \sum_{z\in\hat{\mathcal{V}}^{[t]}_c\setminus\{x\}} \hat{N}_{y,z}}{\hat{\alpha}_c}. \end{aligned}$$

In particular, recognize that for any $k = 1, \dots, K$, asymptotically

$$\frac{1}{T} \sum_{y\in\mathcal{V}\setminus\{x\}} \sum_{z\in\hat{\mathcal{V}}^{[t]}_k\setminus\{x\}} \hat{N}_{y,z} \sim \frac{1}{T} \sum_{y\in\mathcal{V}} \sum_{z\in\hat{\mathcal{V}}^{[t]}_k} \hat{N}_{y,z} \overset{\text{(i)}}{\sim} \hat{\pi}_k$$

where for (i) we have used global balance. Finally expand the logarithms and separate out all terms that do not depend on c. Then conclude that when maximizing over c, this is equivalent to maximizing the reduced objective function

$$\begin{aligned} u^{\text{red}}_x(c) &= \sum_{y\in\mathcal{V}\setminus\{x\}} \big(\hat{N}_{x,y} \ln \hat{p}_{c,\sigma^{[t]}(y)} + \hat{N}_{y,x} \ln \hat{p}_{\sigma^{[t]}(y),c}\big) - \frac{T}{n} \cdot \frac{\hat{\pi}_c}{\hat{\alpha}_c} \\ &= \sum_{k=1}^{K} \big(\hat{N}_{x,\hat{\mathcal{V}}^{[t]}_k} \ln \hat{p}_{c,k} + \hat{N}_{\hat{\mathcal{V}}^{[t]}_k,x} \ln \frac{\hat{p}_{k,c}}{\hat{\alpha}_c}\big) - \frac{T}{n} \cdot \frac{\hat{\pi}_c}{\hat{\alpha}_c} \end{aligned}$$

over c. This concludes the proof.

APPENDIX H: SPECTRAL NORM BOUND FOR SUMS OF ELEMENTS OF MATRICES

LEMMA 17. *For any matrix $B \in \mathbb{R}^{n\times n}$ and subsets $\mathcal{A}, \mathcal{C} \subseteq \{1, \dots, n\}$, we have*

$$\sum_{r\in\mathcal{A}} \sum_{c\in\mathcal{C}} B_{rc} = {1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}}.$$

Furthermore, ${1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}} \leq \|B\| \sqrt{|\mathcal{A}||\mathcal{C}|}$.

PROOF. We have:

$$\begin{aligned} {1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}} &= {1_{\mathcal{A}}}^{\mathrm{T}} \Big(\sum_{r=1}^{n} \Big(\sum_{c=1}^{n} B_{rc} \mathbb{1}[c \in \mathcal{C}] e_{n,r}\Big)\Big) = \sum_{c'=1}^{n} \mathbb{1}[c' \in \mathcal{A}] {e_{n,c'}}^{\mathrm{T}} \Big(\sum_{r=1}^{n} \Big(\sum_{c\in\mathcal{C}} B_{rc} e_{n,r}\Big)\Big) \\ &= \sum_{c'\in\mathcal{A}} \sum_{r=1}^{n} \sum_{c\in\mathcal{C}} B_{rc} {e_{n,c'}}^{\mathrm{T}} e_{n,r} = \sum_{c'\in\mathcal{A}} \sum_{r=1}^{n} \sum_{c\in\mathcal{C}} B_{rc} \mathbb{1}[c' = r] = \sum_{r\in\mathcal{A}} \sum_{c\in\mathcal{C}} B_{rc}, \end{aligned}$$

which proves the first statement.

For the second statement, first note that (i) ${1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}} \in \mathbb{R}$ and therefore ${1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}} \leq |{1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}}|$. By (ii) applying the Cauchy–Schwarz inequality twice, and (iii) the consistency of subordinate norms, we obtain

$${1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}} \overset{\text{(i)}}{\leq} |{1_{\mathcal{A}}}^{\mathrm{T}} B 1_{\mathcal{C}}| \overset{\text{(ii)}}{\leq} \|1_{\mathcal{A}}\|_2 \|B 1_{\mathcal{C}}\|_2 \overset{\text{(iii)}}{\leq} \|1_{\mathcal{A}}\|_2 \|B\| \|1_{\mathcal{C}}\|_2.$$

Lastly for any set $\mathcal{A} \subseteq \{1, \dots, n\}$, we have that $1_{\mathcal{A}} \in \{0,1\}^n$, and therefore $\|1_{\mathcal{A}}\|_2 = \sqrt{\|1_{\mathcal{A}}\|_1} = \sqrt{|\mathcal{A}|}$. Applying this bound for the sets $\mathcal{A}, \mathcal{C}$ concludes the proof. □

APPENDIX I: STOCHASTIC BOUNDEDNESS PROPERTIES

Recall that when we write $X_n = O_{\mathbb{P}}(a_n)$ for a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ and some deterministic sequence $\{a_n\}_{n=1}^{\infty}$, this is equivalent to saying

$$\forall_{\varepsilon>0}\exists_{\delta_\varepsilon,N_\varepsilon} : \mathbb{P}\Big[\Big|\frac{X_n}{a_n}\Big| \geq \delta_\varepsilon\Big] \leq \varepsilon \, \forall_{n>N_\varepsilon}.$$

LEMMA 18. *Let $\cup_{n=1}^{\infty}\{X_n\}_{n\geq 0}$, $\cup_{n=1}^{\infty}\{Y_n\}$ denote two families of random variables with the properties that $X_n, Y_n \geq 0$, $X_n = O_{\mathbb{P}}(x_n)$, and $Y_n = O_{\mathbb{P}}(y_n)$, where $\{x_n\}_{n=1}^{\infty}$, $\{y_n\}_{n=1}^{\infty}$ denote two deterministic sequences with $x_n, y_n \in [0,\infty)$. Then $X_nY_n = O_{\mathbb{P}}(x_ny_n)$. Similarly if $X_n = \Omega_{\mathbb{P}}(x_n)$, $Y_n = \Omega_{\mathbb{P}}(y_n)$, then $X_nY_n = \Omega_{\mathbb{P}}(x_ny_n)$.*

PROOF. Let $\varepsilon > 0$. Choose $\delta_\varepsilon^X, N_\varepsilon^X$ and $\delta_\varepsilon^Y, N_\varepsilon^Y$ such that $\mathbb{P}[X_n \geq \delta_\varepsilon^X x_n] \leq \varepsilon/3$ and $\mathbb{P}[Y_n \geq \delta_\varepsilon^Y y_n] \leq \varepsilon/3$. Pick any $\delta_\varepsilon > \delta_\varepsilon^X\delta_\varepsilon^Y$. With these choices,

$$\begin{aligned}
&\mathbb{P}\Big[\Big|\frac{X_nY_n}{x_ny_n}\Big| \geq \delta_\varepsilon\Big] = \mathbb{P}\Big[\Big|\frac{X_nY_n}{x_ny_n}\Big| \geq \delta_\varepsilon, X_n \geq \delta_\varepsilon^X x_n, Y_n \geq \delta_\varepsilon^Y y_n\Big] \\
&+ \mathbb{P}\Big[\Big|\frac{X_nY_n}{x_ny_n}\Big| \geq \delta_\varepsilon, X_n \geq \delta_\varepsilon^X x_n, Y_n < \delta_\varepsilon^Y y_n\Big] + \mathbb{P}\Big[\Big|\frac{X_nY_n}{x_ny_n}\Big| \geq \delta_\varepsilon, X_n < \delta_\varepsilon^X x_n, Y_n \geq \delta_\varepsilon^Y y_n\Big] \\
&+ \mathbb{P}\Big[\Big|\frac{X_nY_n}{x_ny_n}\Big| \geq \delta_\varepsilon, X_n < \delta_\varepsilon^X x_n, Y_n < \delta_\varepsilon^Y y_n\Big] \leq \varepsilon.
\end{aligned}$$

We have shown that

$$\forall_{\varepsilon>0}\exists_{\delta_\varepsilon=\delta_\varepsilon^X\delta_\varepsilon^Y, N_\varepsilon=\max\{N_\varepsilon^X,N_\varepsilon^Y\}} : \mathbb{P}\Big[\Big|\frac{X_nY_n}{x_ny_n}\Big| \geq \delta_\varepsilon\Big] \leq \varepsilon \, \forall_{n>N_\varepsilon}.$$

This completes the proof. □

LEMMA 19. *Let $\{s_n\}_{n=1}^{\infty}$ denote a deterministic sequence with $s_n \in \mathbb{N}_+$. Let $\cup_{n=1}^{\infty}\cup_{m=1}^{s_n}\{X_{m,n}\}$ denote a family of random variables with the properties that $X_{m,n} \geq 0$, and $\exists_{\delta,N} : \mathbb{E}[X_{m,n}] \leq \delta x_n \, \forall_{m=1,\ldots,s_n}\forall_{n>N}$. Then $S_n = \sum_{m=1}^{s_n} X_{m,n} = O_{\mathbb{P}}(s_nx_n)$.*

PROOF. Let $\varepsilon > 0, \delta_\varepsilon^\Sigma > 0$. Since (i) $X_{m,n} > 0$ for all m, n, by (ii) Markov's inequality

$$\mathbb{P}\Big[\Big|\frac{S_n}{s_nx_n}\Big| \geq \delta_\varepsilon^\Sigma\Big] \overset{\text{(i)}}{=} \mathbb{P}\Big[\frac{1}{s_nx_n}\sum_{m=1}^{s_n} X_{m,n} \geq \delta_\varepsilon^\Sigma\Big] \overset{\text{(ii)}}{\leq} \frac{\sum_{m=1}^{s_n}\mathbb{E}[X_{m,n}]}{\delta_\varepsilon^\Sigma s_nx_n}. \tag{65}$$

By assumption $\exists_{\delta,N} : \mathbb{E}[X_{m,n}] \leq \delta x_n \, \forall_{m=1,\ldots,s_n}\forall_{n>N}$. Choose δ, N as such. Specify $\delta_\varepsilon^\Sigma = \delta/\varepsilon$. By (65), we have thus shown that

$$\forall_{\varepsilon>0}\exists_{\delta_\varepsilon^\Sigma=\delta/\varepsilon, N_\varepsilon=N} : \mathbb{P}\Big[\Big|\frac{S_n}{s_nx_n}\Big| \geq \delta_\varepsilon^\Sigma\Big] \leq \varepsilon \, \forall_{n>N_\varepsilon}.$$

Equivalently, $S_n = O_{\mathbb{P}}(s_nx_n)$. This completes the proof. □

LEMMA 20. *Let $\cup_{n=1}^{\infty}\cup_{m=1}^{n}\{X_{m,n}\}$ denote a family of random variables with the properties that $X_{m,n} \geq 0$, and $\exists_{\delta,N} : \mathbb{E}[X_{m,n}] \leq \delta x_n \, \forall_{m=1,\ldots,n}\forall_{n>N}$. If $\{Y_n\}_{n=1}^{\infty}$ is a sequence of random variables with the properties that $Y_n \in \{1,\ldots,n\}$, and $Y_n = O_{\mathbb{P}}(y_n)$ for some deterministic sequence $\{y_n\}_{n=1}^{\infty}$ with $y_n \in \mathbb{N}_+$, then $Z_n = \sum_{m=1}^{Y_n\wedge n} X_{m,n} = O_{\mathbb{P}}((y_n \wedge n)x_n)$.*

PROOF. Let $\varepsilon > 0, \delta_\varepsilon^Z > 0$. Then

$$\begin{aligned}
\mathbb{P}\Big[\Big|\frac{Z_n}{y_nx_n}\Big| \geq \delta_\varepsilon^Z\Big] &= \mathbb{P}\Big[\Big|\frac{Z_n}{y_nx_n}\Big| \geq \delta_\varepsilon^Z, \Big|\frac{Y_n}{y_n}\Big| \geq \delta_\varepsilon^Y\Big] + \mathbb{P}\Big[\Big|\frac{Z_n}{y_nx_n}\Big| \geq \delta_\varepsilon^Z, \Big|\frac{Y_n}{y_n}\Big| < \delta_\varepsilon^Y\Big] \\
&\leq \mathbb{P}\Big[\Big|\frac{Y_n}{y_n}\Big| \geq \delta_\varepsilon^Y\Big] + \mathbb{P}\Big[\Big|\frac{1}{y_nx_n}\sum_{m=1}^{(\delta_\varepsilon^Y y_n)\wedge n} X_{m,n}\Big| \geq \delta_\varepsilon^Z\Big].
\end{aligned}$$

By assumption $Y_n = O_{\mathbb{P}}(y_n)$, so we can choose $\delta_\varepsilon^Y \in \mathbb{N}_+, N_\varepsilon^Y > 0$ such that $\mathbb{P}[|Y_n/y_n| \geq \delta_\varepsilon^Y] \leq \varepsilon/2$ for all $n > N_\varepsilon^Y$. Write $\delta_\varepsilon^Z = \delta_\varepsilon^Y \delta_\varepsilon^\Sigma$, and we will specify $\delta_\varepsilon^\Sigma$ in a moment. Presently, we are at

$$\begin{aligned}\mathbb{P}\Big[\Big|\frac{Z_n}{y_n x_n}\Big| \geq \delta_\varepsilon^Z\Big] &\leq \frac{\varepsilon}{2} + \mathbb{P}\Big[\Big|\frac{1}{(\delta_\varepsilon^Y y_n)x_n} \sum_{m=1}^{(\delta_\varepsilon^Y y_n)\wedge n} X_{m,n}\Big| \geq \delta_\varepsilon^\Sigma\Big] \\ &\leq \frac{\varepsilon}{2} + \mathbb{P}\Big[\Big|\frac{1}{(\delta_\varepsilon^Y y_n \wedge n)x_n} \sum_{m=1}^{\delta_\varepsilon^Y y_n \wedge n} X_{m,n}\Big| \geq \delta_\varepsilon^\Sigma\Big].\end{aligned}$$

The assumptions on the family $\{X_{m,n}\}_{m,n=1}^\infty$ now allow us to apply Lemma 19: specifically, there exist $\delta_\varepsilon^\Sigma, N_\varepsilon^\Sigma$ such that the final term is bounded by $\varepsilon/2$ for all $n > N_\varepsilon^\Sigma$. Summarizing, we have shown that

$$\forall_{\varepsilon>0} \exists_{\delta_\varepsilon^Z = \delta_\varepsilon^Y \delta_\varepsilon^\Sigma, N_\varepsilon^Z = \max\{N_\varepsilon^Y, N_\varepsilon^\Sigma,\}} : \mathbb{P}\Big[\Big|\frac{Z_n}{y_n x_n}\Big| \geq \delta_\varepsilon^\Sigma\Big] \leq \varepsilon \, \forall_{n > N_\varepsilon^Z}.$$

Equivalently, $Z_n = O_{\mathbb{P}}(y_n x_n)$. □

LEMMA 21. *Let $\cup_{n=1}^\infty \{X_n\}_{n\geq 0}$, $\cup_{n=1}^\infty \{Y_n\}$ denote two families of random variables with the properties that $\mathbb{P}[X_n \leq Y_n] = 1$, $X_n = \Omega_{\mathbb{P}}(x_n)$, and $Y_n = O_{\mathbb{P}}(y_n)$, where $\{x_n\}_{n=1}^\infty$, $\{y_n\}_{n=1}^\infty$ denote two deterministic sequences with $x_n, y_n \in \mathbb{R}$. Then, $x_n = O(y_n)$.*

PROOF. We prove the result by contradiction. Recall first that the assumptions imply that for every $\varepsilon^X, \varepsilon^Y > 0$, there exist $\delta_\varepsilon^X, \delta_\varepsilon^Y > 0$ such that

$$\lim_{n\to\infty} \mathbb{P}[X_n \leq \delta_\varepsilon^X x_n] \leq \varepsilon^X, \quad \lim_{n\to\infty} \mathbb{P}[Y_n \geq \delta_\varepsilon^Y y_n] \leq \varepsilon^Y.$$

Also note that by (i) definition of conditional probability, (ii) the De Morgan laws, and (iii) $\mathbb{P}[\{X_n \leq \delta^X x_n\} \cap \{Y_n \geq \delta^Y y_n\}] \geq 0$, it follows that

$$\begin{aligned}0 = \mathbb{P}[X_n > Y_n] &\geq \mathbb{P}[\{X_n > Y_n\} \cap \{X_n > \delta^X x_n\} \cap \{Y_n < \delta^Y y_n\}] \\ &\overset{\text{(i)}}{=} \mathbb{P}[X_n > Y_n | \{X_n > \delta^X x_n\} \cap \{Y_n < \delta^Y y_n\}](1 - \mathbb{P}[(\{X_n > \delta^X x_n\} \cap \{Y_n < \delta^Y y_n\})^{\mathrm{c}}]) \\ &\overset{\text{(ii)}}{=} \mathbb{P}[X_n > Y_n | \{X_n > \delta^X x_n\} \cap \{Y_n < \delta^Y y_n\}](1 - \mathbb{P}[\{X_n \leq \delta^X x_n\} \cup \{Y_n \geq \delta^Y y_n\}]) \\ &\overset{\text{(iii)}}{\geq} \mathbb{P}[X_n > Y_n | \{X_n > \delta^X x_n\} \cap \{Y_n < \delta^Y y_n\}](1 - \mathbb{P}[\{X_n \leq \delta^X x_n\}] - \mathbb{P}[\{Y_n \geq \delta^Y y_n\}]).\end{aligned}$$

Now suppose that $x_n = \omega(y_n)$. By then taking the limit $n \to \infty$ both left and right, we obtain the inequality $0 \geq 1 - \varepsilon^X - \varepsilon^Y$, which is a contradiction. Hence it must be that $x_n = O(y_n)$. □

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ELECTRICAL ENGINEERING
DEPT. OF AUTOMATIC CONTROL
OSQULDASVÄG 10, STOCKHOLM 10044, SWEDEN
E-MAIL: jarons@kth.se
alepro@kth.se