Clustering in Block Markov Chains

Jaron Sanders[1,2] Alexandre Proutière[1] Se Young Yun[3]

[1]KTH Royal Institute of Technology, Sweden

[2]Delft University of Technology, The Netherlands

[3]Korea Advanced Institute of Science and Technology, South Korea

Korteweg–de Vries Institute for Mathematics
General Mathematics Colloquium 2019

Part I

Our idea and the motivation

Our idea: Can we do clustering in Markov Chains (MCs)?

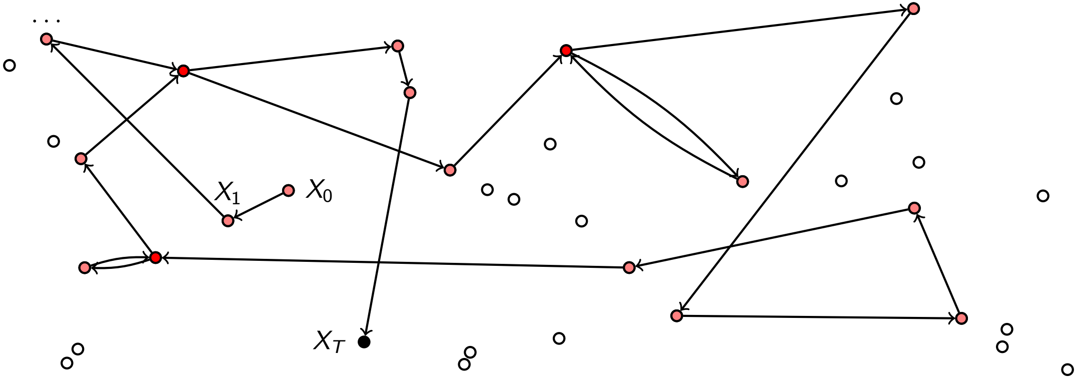


Figure: The goal of this paper is to infer the hidden cluster structure underlying a Markov chain $\{X_t\}_{t\geq 0}$, from one observation of a sample path $X_0, X_1, \ldots, X_T$ of length T.

The motivation

Clustering in MCs is motivated by *Reinforcement Learning (RL)* on large state spaces.

RL has recently received substantial attention due to its wide spectrum of applications (robotics, games, medicine, finance, etc), or more popularly said, *artificial intelligence*.

The motivation

Clustering in MCs is motivated by *Reinforcement Learning (RL)* on large state spaces.

RL has recently received substantial attention due to its wide spectrum of applications (robotics, games, medicine, finance, etc), or more popularly said, *artificial intelligence*.

In RL, the objective is to quickly identify an optimal control policy by *observing a trajectory of a Markov chain*.

Unfortunately, the time to learn the best policies using e.g. Q-learning *increases dramatically* with the number of states.

The motivation

Clustering in MCs is motivated by *Reinforcement Learning (RL)* on large state spaces.

RL has recently received substantial attention due to its wide spectrum of applications (robotics, games, medicine, finance, etc), or more popularly said, *artificial intelligence*.

In RL, the objective is to quickly identify an optimal control policy by *observing a trajectory of a Markov chain*.

Unfortunately, the time to learn the best policies using e.g. Q-learning *increases dramatically* with the number of states.

In practical problems however, different states may yield *similar* reward and exhibit *similar* transition probabilities. **In other words, states could maybe be clustered.**

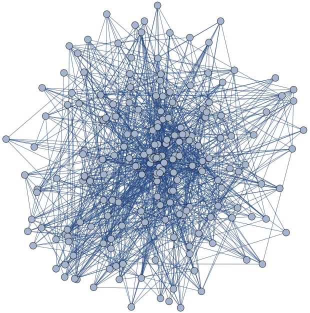
Part II

The literature and our model

Clustering in Stochastic Block Models (SBMs)

SBMs generate random graphs with groups of similar vertices.

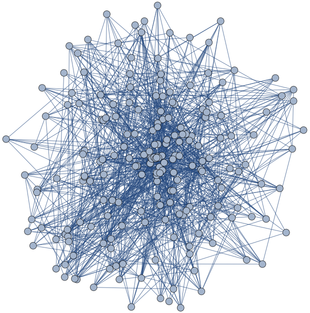
E.g. Suppose $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$. An edge is drawn between $x, y \in \mathcal{V}$ w.p. $p \in (0,1)$ if they belong to the same group, and w.p. $q \in (0,1), p \neq q$ otherwise.



Clustering in Stochastic Block Models (SBMs)

SBMs generate random graphs with groups of similar vertices.

E.g. Suppose $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$. An edge is drawn between $x, y \in \mathcal{V}$ w.p. $p \in (0,1)$ if they belong to the same group, and w.p. $q \in (0,1), p \neq q$ otherwise.



The goal is to infer the clusters from such an observed random graph.

Fundamental limits for clustering in SBMs in literature

Much literature exists on **when** and **how** we can cluster in SBMs.

[1] *"Community detection and SBMs: recent developments"*, Emmanuel Abbe, 2017 gives overview.

Fundamental limits for clustering in SBMs in literature

Much literature exists on **when** and **how** we can cluster in SBMs.

To start, many papers laid foundation for the discovery of the fundamental limits:[1]

Including: Holland, Laskey, Leinhardt 1983; Bui, Chaudhuri, Leighton, Sipser 1984; Boppana 1987; Dyer, Frieze 1989; Snijders, Nowicki 1997; Jerrum, Sorkin 1998; Condon, Karp 1999; Carson, Impagliazzo 2001; McSherry 2001; Bickel, Chen 2009; Rohe, Chatterjee, Yi 2011, and more.

[1] *"Community detection and SBMs: recent developments"*, Emmanuel Abbe, 2017 gives overview.

Fundamental limits for clustering in SBMs in literature

Much literature exists on **when** and **how** we can cluster in SBMs.

To start, many papers laid foundation for the discovery of the fundamental limits:[1]

Including: Holland, Laskey, Leinhardt 1983; Bui, Chaudhuri, Leighton, Sipser 1984; Boppana 1987; Dyer, Frieze 1989; Snijders, Nowicki 1997; Jerrum, Sorkin 1998; Condon, Karp 1999; Carson, Impagliazzo 2001; McSherry 2001; Bickel, Chen 2009; Rohe, Chatterjee, Yi 2011, and more.

Theorem (Decelle, Krzakala, Moore, Zdeborova 2011; Massoulié 2014; Mossel, Neeman, Sly 2015)

If $p = a/n$, $q = b/n$, and $|\mathcal{V}_1| = |\mathcal{V}_2|$, then $a - b \geq \sqrt{2(a+b)}$ is a necessary and sufficient condition for the existence of algorithms that can detect the clusters.

Theorem (Abbe, Bandeira, Hall, 2014; Mossel, Neeman, Sly 2014)

If $p = a \ln n/n$, $q = b \ln n/n$, then $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ allows for exact recovery.

[1] *"Community detection and SBMs: recent developments"*, Emmanuel Abbe, 2017 gives overview.

Fundamental limits for clustering in SBMs in literature

Much literature exists on **when** and **how** we can cluster in SBMs.

To start, many papers laid foundation for the discovery of the fundamental limits:[1]

Including: Holland, Laskey, Leinhardt 1983; Bui, Chaudhuri, Leighton, Sipser 1984; Boppana 1987; Dyer, Frieze 1989; Snijders, Nowicki 1997; Jerrum, Sorkin 1998; Condon, Karp 1999; Carson, Impagliazzo 2001; McSherry 2001; Bickel, Chen 2009; Rohe, Chatterjee, Yi 2011, and more.

Theorem (Decelle, Krzakala, Moore, Zdeborova 2011; Massoulié 2014; Mossel, Neeman, Sly 2015)

If $p = a/n$, $q = b/n$, and $|\mathcal{V}_1| = |\mathcal{V}_2|$, then $a - b \geq \sqrt{2(a+b)}$ is a necessary and sufficient condition for the existence of algorithms that can <u>detect</u> the clusters.

Theorem (Abbe, Bandeira, Hall, 2014; Mossel, Neeman, Sly 2014)

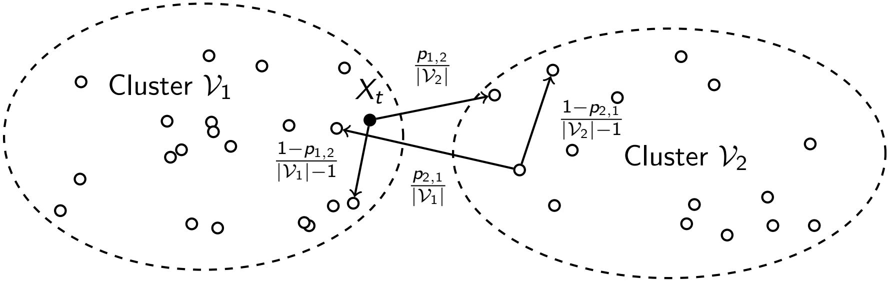
If $p = a \ln n/n$, $q = b \ln n/n$, then $|\sqrt{a} - \sqrt{b}| > \sqrt{2}$ allows for <u>exact</u> recovery.

In both cases, **efficient algorithms** were also developed that achieve the thresholds!

[1] *"Community detection and SBMs: recent developments"*, Emmanuel Abbe, 2017 gives overview.

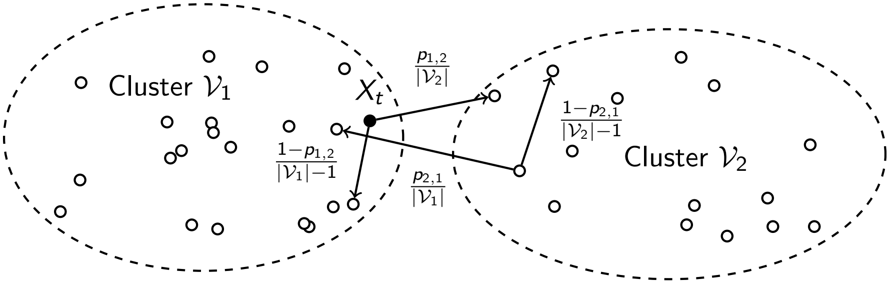
Clustering in Block Markov Chains (BMCs)

Our work also investigates **when** and **how** we can cluster, **but then in BMCs**!



Clustering in Block Markov Chains (BMCs)

Our work also investigates **when** and **how** we can cluster, **but then in BMCs**!



Let $\{X_t\}_{t\geq 0}$ be a BMC with parameters (n, α, p). Its transition matrix is given by

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

Its equilibrium distribution will be denoted by Π_x for $x \in \mathcal{V}$.

Structure of the transition matrix

Here's an example transition matrix for $K = 3$ clusters:

$$P = \left(\begin{array}{cc:ccc:ccccc}
0 & p_{1,1} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} \\
p_{1,1} & 0 & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,2}}{3} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} & \frac{p_{1,3}}{5} \\
\hdashline
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & 0 & \frac{p_{2,2}}{2} & \frac{p_{2,2}}{2} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & \frac{p_{2,2}}{2} & 0 & \frac{p_{2,2}}{2} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} \\
\frac{p_{2,1}}{2} & \frac{p_{2,1}}{2} & \frac{p_{2,2}}{2} & \frac{p_{2,2}}{2} & 0 & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} & \frac{p_{2,3}}{5} \\
\hdashline
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & 0 & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & 0 & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & 0 & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & 0 & \frac{p_{3,3}}{4} \\
\frac{p_{3,1}}{2} & \frac{p_{3,1}}{2} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,2}}{3} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & \frac{p_{3,3}}{4} & 0
\end{array}\right)$$

Note the **block structure**, and that p must be a **stochastic matrix**.

Equilibrium behavior of the inner chain

The block structure motivates us to define

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

Proposition

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 $\Omega = \{1, \dots, K\}$.

Example ($K = 2$ clusters)

After solving the balance equations that the limiting equilibrium behavior is given by $\pi_1 = p_{21}/(p_{12} + p_{21})$ and $\pi_2 = p_{12}/(p_{12} + p_{21})$.

Mixing time

Analyzing and bounding the **mixing time** of a BMC is crucial.

Without mixing within T time steps, we would not expect to be able to cluster.

We define $d(t) \triangleq \sup_{x \in \mathcal{V}} \{ d_{\mathrm{TV}}(P^t_{x,\cdot}, \Pi) \}$ and $t_{\mathrm{mix}}(\varepsilon) \triangleq \min\{ t \geq 0 : d(t) \leq \varepsilon \}$, where

$$d_{\mathrm{TV}}(\mu, \nu) \triangleq \frac{1}{2} \sum_{x \in \mathcal{V}} |\mu_x - \nu_x|.$$

Mixing time

Analyzing and bounding the **mixing time** of a BMC is crucial.

Without mixing within T time steps, we would not expect to be able to cluster.

We define $d(t) \triangleq \sup_{x\in\mathcal{V}}\{d_{\mathrm{TV}}(P^t_{x,\cdot},\Pi)\}$ and $t_{\mathrm{mix}}(\varepsilon) \triangleq \min\{t \geq 0 : d(t) \leq \varepsilon\}$, where

$$d_{\mathrm{TV}}(\mu,\nu) \triangleq \frac{1}{2}\sum_{x\in\mathcal{V}} |\mu_x - \nu_x|.$$

Proposition

There exists a strictly positive absolute constant c_{mix} *such that* $t_{\mathrm{mix}}(\varepsilon) \leq -c_{\mathrm{mix}} \ln \varepsilon$, *for every BMC of finite size* $n \geq K$.

In other words, the mixing times are **very short** in light of our system size n.

Part III

Our main results

Main results

We obtain quantitative statements for

$$\mathcal{E} \triangleq \bigcup_{k=1}^{K} \hat{\mathcal{V}}_{\gamma^{\mathrm{opt}}(k)} \backslash \mathcal{V}_k \quad \text{where} \quad \gamma^{\mathrm{opt}} \in \arg \min_{\gamma \in \mathrm{Perm}(K)} \Big| \bigcup_{k=1}^{K} \hat{\mathcal{V}}_{\gamma(k)} \backslash \mathcal{V}_k \Big|.$$

Here, the sets $\hat{\mathcal{V}}_1, \ldots, \hat{\mathcal{V}}_K$ will always denote an approximate cluster assignment obtained from some clustering algorithm.

Main results

We obtain quantitative statements for

$$\mathcal{E} \triangleq \bigcup_{k=1}^{K} \hat{\mathcal{V}}_{\gamma^{\mathsf{opt}}(k)} \backslash \mathcal{V}_k \quad \text{where} \quad \gamma^{\mathsf{opt}} \in \arg \min_{\gamma \in \mathrm{Perm}(K)} \Big| \bigcup_{k=1}^{K} \hat{\mathcal{V}}_{\gamma(k)} \backslash \mathcal{V}_k \Big|.$$

Here, the sets $\hat{\mathcal{V}}_1, \dots, \hat{\mathcal{V}}_K$ will always denote an approximate cluster assignment obtained from some clustering algorithm.

Remark

Throughout, we assume that K, α, p are fixed, and we study the asymptotic regime $n \to \infty$. Our clustering procedure will assume that K is known, and α, p unknown.

Information theoretical lower bound

Definition

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

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

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

Information theoretical lower bound

Definition

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

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

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

Theorem

An algorithm is (ε, c)-locally good at (α, p) if it satisfies $\mathbb{E}_P[|\mathcal{E}|] \leq \varepsilon$ for all BMC models constructed from the given p and partitions satisfying $||\mathcal{V}_k| - \alpha_k n| \leq c$ for all k. Assume that $T = \omega(n)$. Then there exists a strictly positive and finite constant C independent of n such that: there exists no $(\varepsilon, 1)$-locally good clustering algorithm at (α, p) when

$$\varepsilon < Cn \exp\Big(- I(\alpha, p) \frac{T}{n} (1 + o(1)) \Big).$$

Asymptotically accurate / exact detection

Conditions for asymptotically accurate detection

In view of our lower bound,

$$\mathbb{E}_P\Big[\frac{|\mathcal{E}|}{n}\Big] \geq C \exp\Big(- I(\alpha, p) \frac{T}{n} (1 + o(1)) \Big),$$

there may exist asymptotically *accurate* $(\varepsilon, 1)$-locally good algorithms at (α, p) only if $I(\alpha, p) > 0$ and $T = \omega(n)$.

Asymptotically accurate / exact detection

Conditions for asymptotically accurate detection
In view of our lower bound,

$$\mathbb{E}_P\Big[\frac{|\mathcal{E}|}{n}\Big] \geq C \exp\Big(- I(\alpha,p)\frac{T}{n}(1+o(1))\Big),$$

there may exist asymptotically *accurate* $(\varepsilon,1)$-locally good algorithms at (α,p) only if $I(\alpha,p) > 0$ and $T = \omega(n)$.

Conditions for asymptotically exact detection
Similarly,

$$\mathbb{E}_P[|\mathcal{E}|] \geq C \exp\Big(\ln n - I(\alpha,p)\frac{T}{n}(1+o(1))\Big),$$

so necessary conditions for the existence of an asymptotically *exact* $(\varepsilon,1)$-locally good algorithm at (α,p) are $I(\alpha,p) > 0$ and $T - \frac{n\ln(n)}{I(\alpha,p)} = \omega(1)$. In particular, T must scale atleast as $n \ln n$.

Information quantity $I(\alpha, p)$ for $K = 2$ clusters

These systems have three parameters: $\alpha_2, p_{1,2}, p_{2,1} \in (0, 1)$

Question! Consider a BMC with $\alpha_2 = \frac{1}{2}$ and $p_{1,2} = 1 - p_{2,1} \neq \frac{1}{2}$ and $p_{1,2} > p_{2,1}$ w.l.o.g. In this scenario, $P_{x,z} = P_{y,z}$ for all $x, y, z \in \mathcal{V}$, that is, every row of the kernel is identical to any other row. *Intuitively, do you expect that we are able to cluster?*

Information quantity $I(\alpha, p)$ for $K = 2$ clusters

These systems have three parameters: $\alpha_2, p_{1,2}, p_{2,1} \in (0, 1)$

Question! Consider a BMC with $\alpha_2 = \frac{1}{2}$ and $p_{1,2} = 1 - p_{2,1} \neq \frac{1}{2}$ and $p_{1,2} > p_{2,1}$ w.l.o.g. In this scenario, $P_{x,z} = P_{y,z}$ for all $x, y, z \in \mathcal{V}$, that is, every row of the kernel is identical to any other row. *Intuitively, do you expect that we are able to cluster?*

Answer. In spite of the transition matrix' rows all being identical, we *can* still cluster. Here $\pi_2 > \pi_1$, and we could cluster based on the equilibrium distribution as $T \to \infty$.

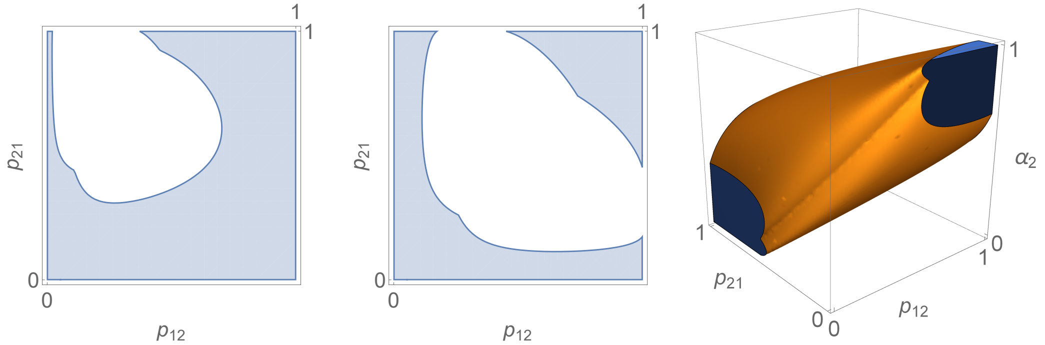
More precisely,

$$I(\alpha, p) = 0 \quad \text{if and only if} \quad \alpha_2 = p_{1,2} = 1 - p_{2,1}$$

Asymptotically **accurate** recovery thus seems possible as soon as $T = \omega(n)$, and asymptotically **exact** recovery as soon as $T = \omega(n \ln n)$.

Clustering in the critical regime

There is a **phase transition** in the *critical regime* $T = n \ln n$



$\alpha_2 = 1/4$ $\alpha_2 = 1/2$ Region where $I(\alpha, p) < 1$.

Figure: (left, middle) The parameters $(p_{1,2}, p_{2,1})$ in blue for which asymptotic exact recovery should be possible in the critical regime $T = n \ln n$ for $K = 2$ clusters. (right) The parameters $(\alpha_2, p_{1,2}, p_{2,1})$ for which asymptotic exact recovery is likely not possible, i.e., $I(\alpha, p) < 1$.

Procedure for cluster recovery

We have now established **necessary conditions** for asymptotically accurate and exact recovery, and identified **performance limits** satisfied by any $(\varepsilon, 1)$-locally good clustering algorithms at (α, p).

Procedure for cluster recovery

We have now established **necessary conditions** for asymptotically accurate and exact recovery, and identified **performance limits** satisfied by any $(\varepsilon, 1)$-locally good clustering algorithms at (α, p).

Next, we devised an $(\varepsilon, 1)$-locally good clustering procedure at (α, p) that **reaches** these limits order-wise. Our procedure takes as input $X_0, X_1, \ldots, X_T$, calculates

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

and then proceeds in two steps called:

- the *Spectral Clustering Algorithm (SCA)*, and
- the *Cluster Improvement Algorithm (CIA)*

Spectral Clustering Algorithm (SCA)

Input: n, K, and 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 matrix $\hat{N}$

begin

    for $x \leftarrow 1$ **to** n **do**

        for $y \leftarrow 1$ **to** n **do**

            $\hat{N}_{x,y} \leftarrow \sum_{t=0}^{T-1} \mathbb{1}[X_t = x, X_{t+1} = y]$;

        end

    end

    Calculate the trimmed matrices $\hat{N}_\Gamma$;

    Calculate the Singular Value Decomposition (SVD) $U \Sigma V^{\mathrm{T}}$ of $\hat{N}_\Gamma$;

    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 $[\hat{R}, \hat{R}^\top]$ to determine $\hat{\mathcal{V}}_1^{[0]}, \dots, \hat{\mathcal{V}}_K^{[0]}$;

end

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

Performance of the SCA

Theorem

Assume that $T = \omega(n)$ *and* $I(\alpha, p) > 0$. *Then the proportion of misclassified states after the Spectral Clustering Algorithm satisfies:*

$$\frac{|\mathcal{E}|}{n} = O_{\mathbb{P}}\Big(\frac{n}{T} \ln \frac{T}{n} \Big) = o_{\mathbb{P}}(1).$$

Thus the SCA achieves asymptotically accurate detection whenever this is possible.

Question! But there's a huge problem. What does the SCA fail at?

Performance of the SCA

Theorem

Assume that $T = \omega(n)$ and $I(\alpha, p) > 0$. Then the proportion of misclassified states after the Spectral Clustering Algorithm satisfies:

$$\frac{|\mathcal{E}|}{n} = O_{\mathbb{P}}\Big(\frac{n}{T} \ln \frac{T}{n} \Big) = o_{\mathbb{P}}(1).$$

Thus the SCA achieves asymptotically accurate detection whenever this is possible.

Question! But there's a huge problem. What does the SCA fail at?

Answer. The bound fails to guarantee asymptotic exact recovery, even in the case $T = \omega(n \ln(n))$. We cannot guarantee that its recovery rate approaches Theorem 5's fundamental limit!

Cluster Improvement Algorithm (CIA)

Input: An approximate assignment $\hat{\mathcal{V}}_1^{[t]}, \ldots, \hat{\mathcal{V}}_K^{[t]}$, and matrix $\hat{N}$

Output: A revised assignment $\hat{\mathcal{V}}_1^{[t+1]}, \ldots, \hat{\mathcal{V}}_K^{[t+1]}$

begin

  $n \leftarrow \dim(\hat{N})$, $\mathcal{V} \leftarrow \{1, \ldots, n\}$, $T \leftarrow \sum_{x \in \mathcal{V}} \sum_{y \in \mathcal{V}} \hat{N}_{x,y}$;

  for $a \leftarrow 1$ **to** K **do**

    $\hat{\pi}_a \leftarrow \hat{N}_{\hat{\mathcal{V}}_a^{[t]}, \mathcal{V}} / T$, $\hat{\alpha}_a \leftarrow |\hat{\mathcal{V}}_a^{[t]}| / n$, $\hat{\mathcal{V}}_a^{[t+1]} \leftarrow \emptyset$;

    for $b \leftarrow 1$ **to** K **do**

      $\hat{p}_{a,b} \leftarrow \hat{N}_{\hat{\mathcal{V}}_a^{[t]}, \hat{\mathcal{V}}_b^{[t]}} / \hat{N}_{\hat{\mathcal{V}}_a^{[t]}, \mathcal{V}}$;

    end

  end

  for $x \leftarrow 1$ **to** n **do**

    $c_x^{\text{opt}} \leftarrow \arg\max_{c=1,\ldots,K} \left\{ \sum_{k=1}^{K} \left(\hat{N}_{x, \hat{\mathcal{V}}_k^{[t]}} \ln \hat{p}_{c,k} + \hat{N}_{\hat{\mathcal{V}}_k^{[t]}, x} \ln \frac{\hat{p}_{k,c}}{\hat{\alpha}_c} \right) - \frac{T}{n} \cdot \frac{\hat{\pi}_c}{\hat{\alpha}_c} \right\}$;

    $\hat{\mathcal{V}}_{c_x^{\text{opt}}}^{[t+1]} \leftarrow \hat{\mathcal{V}}_{c_x^{\text{opt}}}^{[t+1]} \cup \{x\}$;

  end

end

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

Performance of the CIA

Theorem

Assume that $T = \omega(n)$ and $I(\alpha, p) > 0$. Then for any $t \geq 1$, after t iterations of the Clustering Improvement Algorithm, initially applied to the output of the Spectral Clustering Algorithm, we have:

$$\frac{|\mathcal{E}^{[t]}|}{n} = O_{\mathbb{P}}\Big(\mathrm{e}^{-t\left(\ln \frac{T}{n} - \ln \ln \frac{T}{n}\right)} + \mathrm{e}^{-\frac{\alpha_{\min}^2}{720\eta^3\alpha_{\max}^2}\frac{T}{n} I(\alpha,p)}\Big).$$

Performance of the CIA

Theorem

Assume that $T = \omega(n)$ and $I(\alpha, p) > 0$. Then for any $t \geq 1$, after t iterations of the Clustering Improvement Algorithm, initially applied to the output of the Spectral Clustering Algorithm, we have:

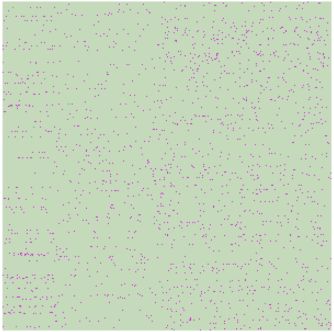
$$\frac{|\mathcal{E}^{[t]}|}{n} = O_{\mathbb{P}}\Big(\mathrm{e}^{-t\left(\ln \frac{T}{n} - \ln \ln \frac{T}{n}\right)} + \mathrm{e}^{-\frac{\alpha_{\min}^2}{720\eta^3\alpha_{\max}^2} \frac{T}{n} I(\alpha,p)}\Big).$$

Observe that for $t = \ln(n)$, the number of misclassified vertices after t applications of the CIA is at most of the order $n\mathrm{e}^{-C\frac{T}{n}I(\alpha,p)}$. Up to the constant $C \triangleq \alpha_{\min}^2/(720\eta^3\alpha_{\max}^2)$, this corresponds to Theorem 5's fundamental recovery rate limit.

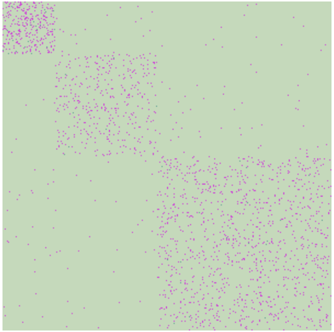
Plus, we have **asymptotically <u>exact</u> detection** when $T = \omega(n \ln n)$ and $I(\alpha, p) > 0$!

Let's start with an example – The observation and truth

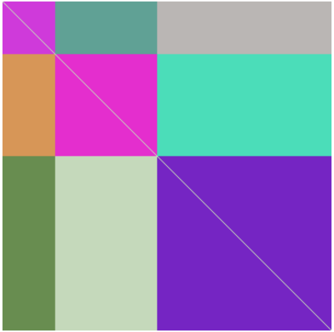
Consider $n = 300$ states grouped into three clusters of respective relative sizes $\alpha = (0.15, 0.35, 0.5)$. The transition rates between these clusters are defined by: $p = (0.9200, 0.0450, 0.0350;\ 0.0125, 0.8975, 0.0900;\ 0.0175, 0.0200, 0.9625)$.



(a) $\hat{N}$, unsorted

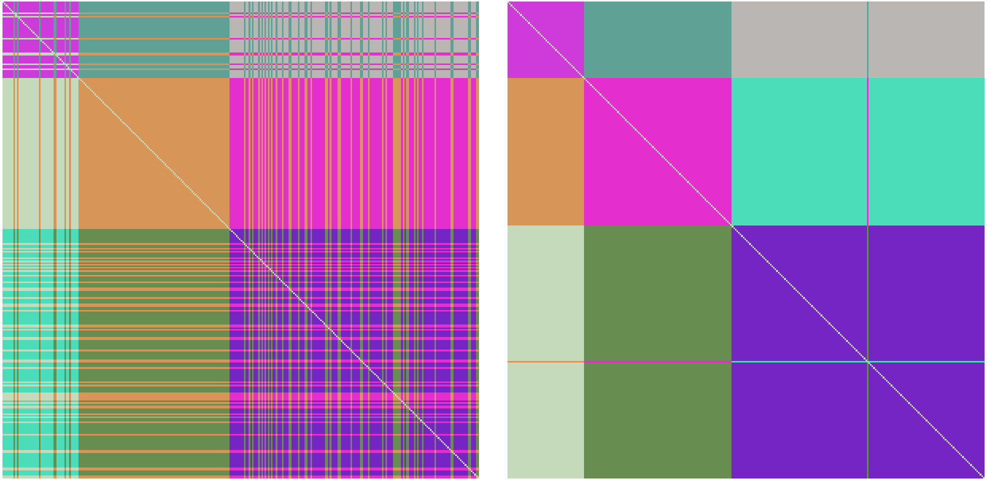


(b) $\hat{N}$, sorted



(c) P, sorted

Let's start with an example – The procedure's 99.7% recovery

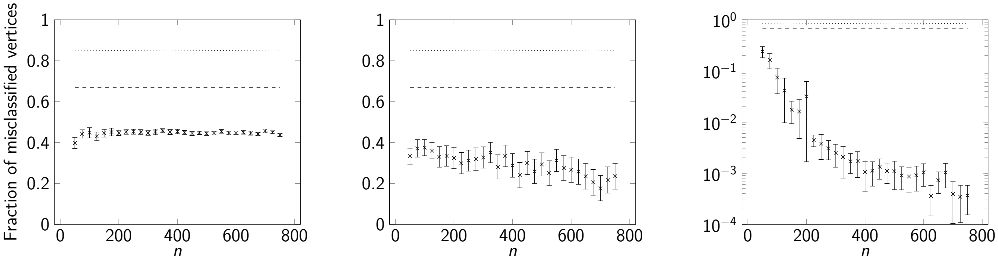


(a) Initial clustering.

(b) Final clustering.

Performance sensitivity of the SCA

Here $\alpha = (0.15, 0.35, 0.5)$, and $p = (0.50, 0.20, 0.30;\ 0.10, 0.70, 0.20;\ 0.35, 0.05, 0.60)$.
Now $I(\alpha, p) \approx 0.88 > 0$, so lower than before, meaning that clustering is more difficult.



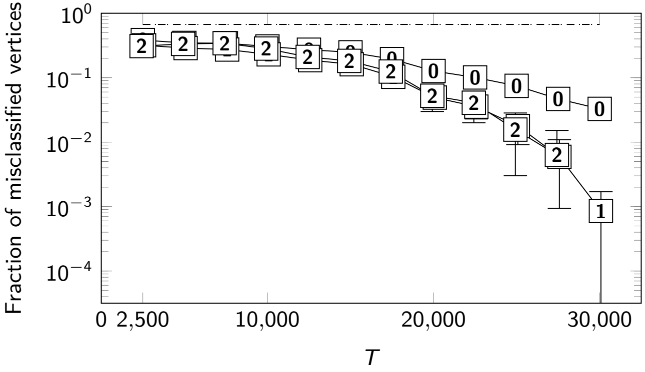
(a) $T = n \ln n$ (b) $T = n(\ln n)^{3/2}$ (c) $T = n(\ln n)^{2}$

Figure: The error rate of the Spectral Clustering Algorithm (without trimming) 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.

Performance sensitivity of the CIA

Here $\alpha = (1/3, 1/3, 1/3)$, and $p = (0.1, 0.4, 0.5;\ 0.7, 0.1, 0.2;\ 0.6, 0.3, 0.1)$.

Different from before, the clusters are now of equal size and the **off-diagonal entries** of p are dominant. Here, $I(\alpha, p) \approx 0.27 > 0$, so clustering is again more challenging.

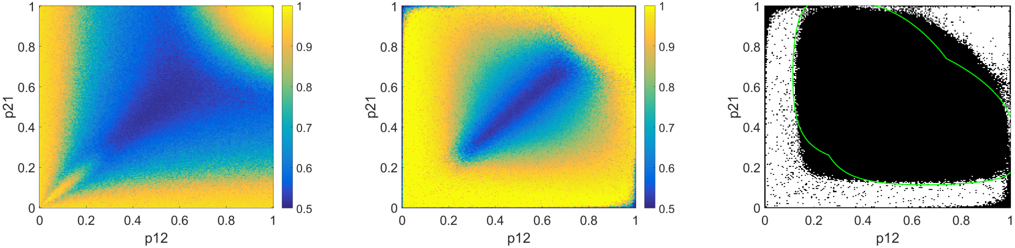


The error after applying the SCA (0), and the CIA (1, 2) twice, as function of T.

At $T = 30000$, the CIA achieved 100% accurate detection after 2 iterations in **all** 200 instances.

Our procedure in the critical regime

Consider $K = 2$, $\alpha_2 = \frac{1}{2}$, and $T = n \ln n$. Pascal Lagerweij (a MSc student) helped us numerically evaluate $\hat{\mathcal{F}}_1(\varepsilon) = \left\{ (p_{1,2}, p_{2,1}) \in (0,1)^2 \middle| \mathbb{E}_P \left[\frac{|\mathcal{E}^{[t]}|}{n} \right] \geq 1 - \varepsilon \right\}$.



After the SCA. After the CIA. $\hat{\mathcal{F}}_1(\varepsilon = 0.027)$

Figure: The average proportion of well-classified states for each rasterpoint $(p_{1,2}, p_{2,1}) \in (0,1)^2$, and numerical feasibility region of our clustering procedure (right), all in the critical regime $T = n \ln n$. The green line outlines the theoretical region $I(\alpha, p) \leq 1$ within which no algorithm exists able to asymptotically recover the clusters exactly.

Part IV

In conclusion

Let us summarize

Our paper “Clustering in Block Markov Chains”:

- introduces BMCs, a new interesting model;
- provides an information-theoretical lower bound for the detection error, tight conditions for asymptotically accurate detection and an almost tight condition for exact recovery;

Let us summarize

Our paper “Clustering in Block Markov Chains”:

- introduces BMCs, a new interesting model;
- provides an information-theoretical lower bound for the detection error, tight conditions for asymptotically accurate detection and an almost tight condition for exact recovery;
- proposes an algorithm that almost reaches our information-theoretical lower bound;
- develops a new spectrum concentration bound for random matrices with *dependent* entries.

A preprint “Clustering in Block Markov Chains” is available on **`https://arxiv.org/abs/1712.09232`**.

Part V

Appendix: Our proofs

The information bound

Theorem

If $T = \omega(n)$ and $I(\alpha, p) > 0$, then there exists a strictly positive and finite constant C independent of n such that: for any clustering algorithm

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

where

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

Here

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

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

$$\mathcal{Q}(k,l) \triangleq \left\{ q \in \mathcal{Q} \middle| I_k(q||p) = I_l(q||p) \right\} \neq \emptyset \quad \textit{for all} \quad k \neq l,$$

$$\mathcal{Q} \triangleq \left\{ (q_{k,0}, q_{0,k})_{k=0,\dots,K} \in (0,\infty) \middle| q_{0,0} = 0, \sum_{l=1}^{K} q_{0,l} = 1 \right\}.$$

Our change of measure

In the proof, we suppose that the path $X_0, \ldots, X_T$ is generated by a perturbed stochastic model Ψ, rather than the true model Φ.

Specifically, we randomly choose a vertex $V^* \in \mathcal{V}$ and place it in its **own** cluster with its own **distinct** transition rates. I.e., given V^*, we construct an alternative kernel Q.

Our change of measure

In the proof, we suppose that the path $X_0, \dots, X_T$ is generated by a perturbed stochastic model Ψ, rather than the true model Φ.

Specifically, we randomly choose a vertex $V^* \in \mathcal{V}$ and place it in its **own** cluster with its own **distinct** transition rates. I.e., given V^*, we construct an alternative kernel Q.

Given $X_0, X_1, \dots, X_T \in \mathcal{V}$, the argument then revolves around the log-likelihood ratio

$$L \triangleq \ln \frac{\mathbb{P}_Q[X_0, X_1, \dots, X_T]}{\mathbb{P}_P[X_0, X_1, \dots, X_T]} = \sum_{t=1}^{T} \ln\Big(\frac{Q_{X_{t-1},X_t}}{P_{X_{t-1},X_t}}\Big).$$

Here, $\mathbb{P}_P[X_0, X_1, \dots, X_T] = \prod_{t=1}^{T} P_{X_{t-1},X_t}$. Note that L is a **random variable**.

Intuitively, L measures how likely the path $X_0, \dots, X_T$ is under Q as opposed to P.

The perturbed BMC

$$
Q = \left(\begin{array}{cc|ccc|ccccc}
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} \\
\hline
\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} \\
\hline
\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{array}\right)
- \frac{1}{3 \cdot 10} \left(\begin{array}{cc|ccc|ccccc}
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} \\
\hline
\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} \\
\hline
\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{array}\right).
$$

An intermediate information bound

Using state symmetry, the change of measure's form, and Chebyshev's inequality:

Proposition

Assume that V^* is chosen uniformly at random from two different clusters $\mathcal{V}_a$ and $\mathcal{V}_b$, that Q is constructed from $q \in \mathcal{Q}(a, b)$, and that there exists a $(\varepsilon, 1)$-locally good clustering algorithm at (α, p). Then:
(i) There exists a constant $\delta > 0$ independent of n s.t. $\mathbb{P}_\Psi[V^* \in \mathcal{E}] \geq \delta > 0$.
(ii) There exists a constant $C > 0$ independent of n such that

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

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

Proposition (Leading order behavior of the expectation)

For given $V^* \in \mathcal{V}$ and $q \in \mathcal{Q}$, if $T = \omega(1)$, then

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

Proposition (Variance is negligible due to mixing)

For given $V^* \in \mathcal{V}$ and $q \in \mathcal{Q}$, if $T = \omega(n)$, then

$$\mathrm{Var}_Q[L|\sigma(V^*)] = o(T^2/n^2).$$

The crux is to relate the *covariances between* the T steps of the sample path $X_1, X_2, \dots, X_T$ to the *mixing time* of the underlying Markov chain.

Lemma (Appropriateness)

For any two clusters $a \neq b$ $\exists$ *at least one finite point* $\bar{q} \in \mathcal{Q}$ *s.t.* $I_a(\bar{q}||p) = I_b(\bar{q}||p)$.

Lemma (Deconditioning)

If $T = \omega(n)$*, then for any two clusters* $a \neq b$*, there exists an absolute* $c > 0$ *s.t.*

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

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* $\bar{q} \in \mathcal{Q}(a,b)$.

Bound optimization

You finally optimize the bound: build the change of measure using the parameters

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

By construction $\mathbb{E}_{\Psi}[L] = (T/n) J(\alpha, p) + o(T/n)$, and $0 < J(\alpha, p) < \infty$.

Bound optimization

You finally optimize the bound: build the change of measure using the parameters

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

By construction $\mathbb{E}_\Psi[L] = (T/n)J(\alpha, p) + o(T/n)$, and $0 < J(\alpha, p) < \infty$.

Lemma (Relation between $J(\alpha, p)$ and $I(\alpha, p)$)

For any BMC, $J(\alpha, p) \leq I(\alpha, p)$. Furthermore, $I(\alpha, p) = 0$ if and only if there exists $i \neq j$ such that $p_{i,c} = p_{j,c}$ and $p_{c,i}/\alpha_i = p_{c,j}/\alpha_j$ for all $c \in \{1, \dots, K\}$.

This completes the proof. $\square$

Performance of the Spectral Clustering Algorithm

Step 1. We show that N^0 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 $N^0_{x,\cdot}$, $N^0_{y,\cdot}$ is at least $\Omega(\sqrt{T^2 D_N(\alpha, p)/n^3})$.

Step 2. We upper bound the error $\|\hat{R}^0 - N^0\|_{\mathrm{F}}$ using $\|\hat{N}_\Gamma - N\|$.

Performance of the Spectral Clustering Algorithm

Step 1. We show that N^0 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 $N^0_{x,\cdot}$, $N^0_{y,\cdot}$ is at least $\Omega(\sqrt{T^2 D_N(\alpha, p)/n^3})$.

Step 2. We upper bound the error $\|\hat{R}^0 - N^0\|_{\mathrm{F}}$ using $\|\hat{N}_\Gamma - N\|$.

Step 3. We prove that $\hat{R}$ also satisfies the separability property if $(n/T)\|\hat{N}_\Gamma - N\| \to 0$, as suggested by Step 1 and Step 2.

Step 4. Because of $\hat{R}^0$'s separability property, we must conclude that the number of misclassified states satisfies Theorem 6. Otherwise the separability property of Step 3 would contradict with Step 2.

Proposition (Spectral concentration of a noise matrix with <u>dependent</u> entries)

For any BMC, $\|\hat{N}_\Gamma - N\| = O_{\mathbb{P}}\Big(\sqrt{\frac{T}{n} \ln \frac{T}{n}}\Big)$.

Steps 1, 2, and 3

Lemma (Separability property)

For any $x, y \in \mathcal{V}$ *for which* $\sigma(x) \neq \sigma(y)$, $\|N^0_{x,\cdot} - N^0_{y,\cdot}\|_2 = \Omega\Big(\sqrt{\frac{T^2 D_N(\alpha,p)}{n^3}}\Big)$.

Lemma (Centered $\hat{R}$'s Frobenius norm and $\hat{N}$'s spectral norm)

$\|\hat{R}^0 - N^0\|_{\mathrm{F}} \leq \sqrt{16K}\|\hat{N}_\Gamma - N\|$.

Lemma (Inheritance of separability)

If $\|\hat{N}_\Gamma - N\| = o_{\mathbb{P}}(f(n,T))$ *for some* $f(n,T) = o(T/n)$ *and* $h(n,T)$ *is s.t.* $\omega((f(n,T))^2/n) = \big(h(n,T)\big)^2 = o(T^2 D_N(\alpha,p)/n^3)$, *then*

$$\|\hat{R}^0_{x,\cdot} - N^0_{x,\cdot}\|_2 = \Omega_{\mathbb{P}}\Big(\sqrt{\frac{T^2 D_N(\alpha,p)}{n^3}}\Big) \quad \textit{for any misclassified vertex} \quad x \in \mathcal{E}.$$

Step 4: Contradiction argument

The final step is almost immediate. Gathering Steps 1 – 3, we have:

$$\Omega_{\mathbb{P}}\Big(|\mathcal{E}|\frac{T^2 D_N(\alpha,p)}{n^3}\Big) \overset{\text{(i)}}{=} \|\hat{R}^0 - N^0\|_{\mathrm{F}}^2 \overset{\text{(ii)}}{\leq} 16K\|\hat{N}_\Gamma - N\|^2 \overset{\text{(iii)}}{=} O_{\mathbb{P}}\Big(\frac{T}{n}\ln\frac{T}{n}\Big),$$

where (i) stems from Lemma 15 (the terms $\|\hat{R}^0_{x,\cdot} - N^0_{x,\cdot}\|_2^2$ for $x \in \mathcal{V} \setminus \mathcal{E}$ can be added to form the Frobenius norm), (ii) comes from Lemma 14, and (iii) is from Proposition 6.

We deduce that $|\mathcal{E}|/n = O_{\mathbb{P}}((n/T)\ln(T/n))$. This concludes the proof.

Lemma

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, $x_n = O(y_n)$.

Performance of the Cluster Improvement Algorithm

Define $\mathcal{E}_{\mathcal{H}}^{[t]} = \mathcal{E}^{[t]} \cap \mathcal{H}$, where $\mathcal{H}$ is the largest set of states $x \in \Gamma$ that satisfy:

(H1) When $x \in \mathcal{V}_i$, for all $j \neq i$,

$$\sum_{k=1}^{K} \Big(\hat{N}_{x,\mathcal{V}_k} \ln \frac{p_{i,k}}{p_{j,k}} + \hat{N}_{\mathcal{V}_k,x} \ln \frac{p_{k,i}\alpha_j}{p_{k,j}\alpha_i} \Big) + \Big(\frac{\hat{N}_{\mathcal{V}_j,\mathcal{V}}}{\alpha_j n} - \frac{\hat{N}_{\mathcal{V}_i,\mathcal{V}}}{\alpha_i n} \Big) \geq \frac{T}{2n} I(\alpha, p).$$

(H2) $\hat{N}_{x,\mathcal{V}\setminus\mathcal{H}} + \hat{N}_{\mathcal{V}\setminus\mathcal{H},x} \leq 2 \ln ((T/n)^2)$.

Performance of the Cluster Improvement Algorithm

Define $\mathcal{E}_{\mathcal{H}}^{[t]} = \mathcal{E}^{[t]} \cap \mathcal{H}$, where $\mathcal{H}$ is the largest set of states $x \in \Gamma$ that satisfy:

(H1) When $x \in \mathcal{V}_i$, for all $j \neq i$,

$$\sum_{k=1}^{K} \Big(\hat{N}_{x,\mathcal{V}_k} \ln \frac{p_{i,k}}{p_{j,k}} + \hat{N}_{\mathcal{V}_k,x} \ln \frac{p_{k,i}\alpha_j}{p_{k,j}\alpha_i} \Big) + \Big(\frac{\hat{N}_{\mathcal{V}_j,\mathcal{V}}}{\alpha_j n} - \frac{\hat{N}_{\mathcal{V}_i,\mathcal{V}}}{\alpha_i n} \Big) \geq \frac{T}{2n} I(\alpha, p).$$

(H2) $\hat{N}_{x,\mathcal{V}\setminus\mathcal{H}} + \hat{N}_{\mathcal{V}\setminus\mathcal{H},x} \leq 2 \ln((T/n)^2)$.

Summing over all misclassified states that in $\mathcal{E}_{\mathcal{H}}^{[t+1]}$, we obtain

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

Step 1. Concentration implies that $E \approx -(T/n) I(\alpha, p) |\mathcal{E}_{\mathcal{H}}^{[t+1]}| + \|\hat{N}_\Gamma - N\| \sqrt{|\mathcal{E}_{\mathcal{H}}^{[t+1]}||\mathcal{E}_{\mathcal{H}}^{[t]}|}$.

Step 2. For large n, T, Step 1 + suboptimality $E \geq 0$ yields an iterative bound.

Improvement per iteration

Theorem

If $I(\alpha, p) > 0$ and $T = \omega(n)$, and $|\mathcal{E}_{\mathcal{H}}^{[t]}| = O_{\mathbb{P}}(e_n^{[t]})$ for some $0 < e_n^{[t]} = o(n)$, then

$$|\mathcal{E}_{\mathcal{H}}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]} = O\Big(e_n^{[t]}\Big(\frac{n}{T}f(n, T)\Big)^2\Big) = o(e_n^{[t]}).$$

Furthermore, there exists a strictly positive absolute constant C such that

$$|\mathcal{E}_{\mathcal{H}^{\mathrm{c}}}^{[t]}| \leq |\mathcal{H}^{\mathrm{c}}| = O_{\mathbb{P}}\Big(n \exp\Big(-C\frac{T}{n}I(\alpha, p)\Big) + n \exp\Big(-\frac{T}{n}\ln\frac{T}{n}\Big)\Big)$$

for all $t \in \mathbb{N}_0$.

Here, $f(n, T) = \sqrt{(T/n)\ln(T/n)}$.

Step 1: Concentration arguments

Substitute $u_x^{[t]}$'s definition to obtain after simplifying

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

Step 1: Concentration arguments

Substitute $u_x^{[t]}$'s definition to obtain after simplifying

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

Split it into E_1, E_2 centered around diff. objects that concentrate and U the remainder.

E.g. Define $E_1 = E_1^{\mathrm{out}} + E_1^{\mathrm{in}} + E_1^{\mathrm{cross}}$ with

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

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

Step 2: Exploiting suboptimality through a contradiction

Analyzing each term, you will find that:

Lemma

If $T = \omega(n)$, $|\mathcal{E}_{\mathcal{H}}^{[t]}| = O_{\mathbb{P}}(e_n^{[t]})$, and $|\mathcal{E}_{\mathcal{H}}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, then

$$-E_1 = \Omega_{\mathbb{P}}\Big(I(\alpha, p) \frac{T}{n} e_n^{[t+1]} \Big), \quad |U| = O_{\mathbb{P}}\Big(\sqrt{\frac{T}{n}} \Big(\ln \frac{T}{n} \Big) e_n^{[t+1]} \Big), \quad and$$

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

Step 2: Exploiting suboptimality through a contradiction

Analyzing each term, you will find that:

Lemma

If $T = \omega(n)$, $|\mathcal{E}_{\mathcal{H}}^{[t]}| = O_{\mathbb{P}}(e_n^{[t]})$, and $|\mathcal{E}_{\mathcal{H}}^{[t+1]}| \asymp_{\mathbb{P}} e_n^{[t+1]}$, then

$$-E_1 = \Omega_{\mathbb{P}}\Big(I(\alpha, p)\frac{T}{n} e_n^{[t+1]}\Big), \quad |U| = O_{\mathbb{P}}\Big(\sqrt{\frac{T}{n}}\Big(\ln\frac{T}{n}\Big) e_n^{[t+1]}\Big), \quad and$$

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

Suboptimality now implies that $-E_1 \leq |E_2| + |U|$ almost surely. Consequentially,

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

Rearranging when $e_n^{[t+1]} > 0$ completes the proof. □