Probabilistic cellular automata with local transition matrices: synchronization, ergodicity, and inference

Erhan Bayraktar *[a], Fei Lu †[b], Mauro Maggioni ‡[b,c], Ruoyu Wu §[d], and Sichen Yang ¶[c]

[a]Department of Mathematics, University of Michigan, Ann Arbor, USA.
[b]Department of Mathematics, Johns Hopkins University, Baltimore, USA.
[c]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, USA.
[d]Department of Mathematics, Iowa State University, Ames, USA.

Abstract

We introduce a new class of probabilistic cellular automata that are capable of exhibiting rich dynamics such as synchronization and ergodicity and can be easily inferred from data. The system is a finite-state locally interacting Markov chain on a circular graph. Each site's subsequent state is random, with a distribution determined by its neighborhood's empirical distribution multiplied by a local transition matrix. We establish sufficient and necessary conditions on the local transition matrix for synchronization and ergodicity. Also, we introduce novel least squares estimators for inferring the local transition matrix from various types of data, which may consist of either multiple trajectories, a long trajectory, or ensemble sequences without trajectory information. Under suitable identifiability conditions, we show the asymptotic normality of these estimators and provide non-asymptotic bounds for their accuracy.

Key words. probabilistic cellular automata, synchronization, ergodicity, inference

Contents

*erhan@umich.edu
†feilu@math.jhu.edu
‡mauromaggionijhu@icloud.com
§ruoyu@iastate.edu
¶syang114@jhu.edu

1 Introduction

Interacting systems, including probabilistic cellular automata (PCA) [Too94, LMS90, LN18] and interacting particles systems (IPS) [Lig85, Dur07, Ald13, Gri18], have a wide range of applications in Physics, Computer Science and Electrical Engineering, Economics and Finance, Biology and Sociology, Epidemiology and Ecology. These applications drive a growing interest in studying the dynamics of these systems, and inference of model parameters from observational data. As Aldous [Ald13] pointed out, it is "the most broad-ranging currently active field of applied probability"; however, "it is easy to invent and simulate models, but hard to give rigorous proofs or to relate convincingly to real-world data". In this study, we introduce a new class of PCA that exhibits rich dynamics, yet can be easily inferred from observational data. We rigorously prove dynamical properties such as synchronization and ergodicity, and then construct computationally efficient estimators for which we prove properties such as asymptotic normality and non-asymptotic bounds.

We consider a new class of probabilistic cellular automata (PCA) on a N-node cyclic graph (V, E) and a finite alphabet $\mathcal{A}$, in which every site updates independently with a distribution determined by its neighborhood's empirical distribution multiplied by a local transition matrix.

The probabilistic cellular automaton we consider is a finite-state Markov chain:

$$X(t) = (X_1(t), \ldots, X_N(t)) =: X_{1:N}(t) \in \mathcal{A}^N, \quad \mathcal{A} = \{1, \ldots, K\} =: [K]$$

on a cyclic graph (V, E) with nodes indexed by $V = \{1, 2, \ldots, N\} =: [N]$ and with edges in E connecting nodes within distances n_v, that is, nodes n and n' are connected whenever $|n - n'| \leqslant n_v$ (modulo N[1]). Conditional on $X(t)$, each vertex n makes updates independently depending on its neighborhood [2]

$$V_n = \{n - n_v, \ldots, n, n+1, \ldots, n + n_v\} \quad \text{(modulo } N\text{, e.g., 0 is viewed as } N\text{)},$$

and the transition probability of $X(t)$ is in the form

$$\mathbb{P}\{X(t+1) = (x_1, \ldots, x_N) | X(t)\} = \prod_{n=1}^{N} \mathbb{P}\{X_n(t+1) = x_n | (X_{n'}(t))_{n' \in V_n}\}. \tag{1.1}$$

In particular, the local transition probability $\mathbb{P}\{X_n(t+1) = x_n | (X_{n'}(t))_{n' \in V_n}\}$ of the n-th vertex depends *linearly* on its neighborhood's empirical distribution:

$$\mathbb{P}\{X_n(t+1) = x_n | (X_{n'}(t))_{n' \in V_n}\} = \varphi_{n,t} \mathbf{T}(\cdot, x_n), \tag{1.2}$$

where $\mathbf{T} \in [0,1]^{K \times K}$ is a row stochastic matrix (that is, $\sum_{k=1}^K \mathbf{T}(j,k) = 1$ for each j) that we call *local transition matrix*, and $\varphi_{n,t} \in \mathbb{R}^{1 \times K}$ is the *local empirical distribution* of the vertex's neighborhood V_n at time t

$$\varphi_{n,t} = (\varphi_{n,t}(1), \ldots, \varphi_{n,t}(K)), \quad \text{with } \varphi_{n,t}(k) := \frac{1}{|V_n|} \sum_{j \in V_n} \delta_{X_j(t)}(k), \quad k \in [K], \tag{1.3}$$

where $|V_n|$ is the cardinality of V_n, and δ is the Kronecker delta function.

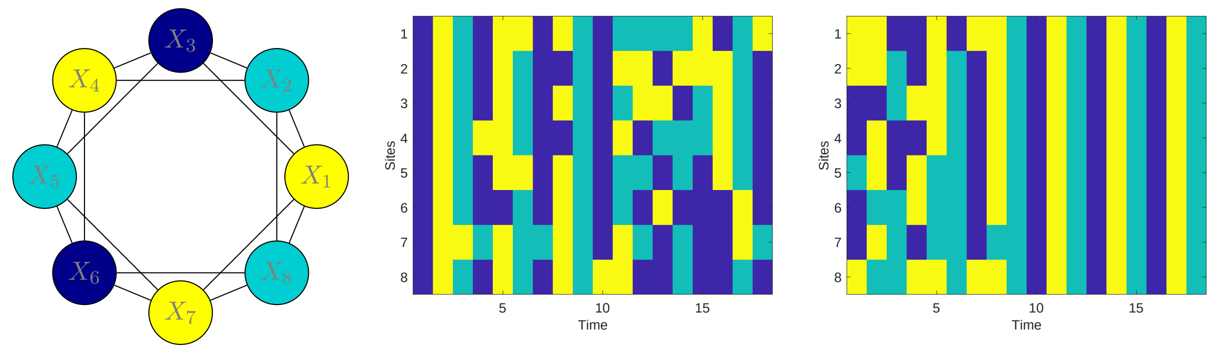
[1] RW: removed "plus 1".

[2] RW: moved here from below (1.3), rephrased "plus 1" after "modulo N".

Equivalently, the Markov chain $X(\cdot)$ has a *global transition matrix* $\mathbf{P} \in \mathbb{R}^{K^N \times K^N}$ determined by the local transition matrix $\mathbf{T}$:

$$\mathbf{P}(x_{1:N}, y_{1:N}) = \prod_{n=1}^{N} \phi_n \mathbf{T}_{\cdot, y_n} = \prod_{n=1}^{N} \frac{1}{|V_n|} \sum_{j \in V_n} \mathbf{T}_{j, y_n} \tag{1.4}$$

for any configurations $x_{1:N}, y_{1:N} \in [K]^N$, where $\phi_n(k) := \frac{1}{|V_n|} \sum_{j \in V_n} \delta_{x_j}(k)$ for $k \in [K]$ and $n \in [N]$. One can also view the system's evolution as follows: at each update step, every vertex samples a state from its neighbors uniformly at random, and then independently jumps according to the local transition matrix $\mathbf{T}$.



(a) PCA on a circle (b) From deterministic to stochastic (c) Synchronization

Figure 1: (a) The system of $N = 8$ agents on a graph with an alphabet $\mathcal{A} = \{\text{Yellow, Turquoise, and Blue}\}$ with $K = 3$. Each agent's transition depends linearly on the empirical distribution of its nearest neighbors with $n_v = 2$ agents on each side. (b) The system moves from deterministic to stochastic dynamics; see Example 2.3. (c) The system moves from stochastic to deterministic dynamics, reaching a synchronization; see Example 2.4.

The key feature of our model is the *linear* dependence of the local transition probability in (1.2) on the local empirical distribution, represented by a local transition matrix. Such a linear dependence significantly reduces the number of parameters describing the local transition probability, which has a size $K^{|V_n|+1}$ since it assigns a $\mathbb{R}^{1\times K}$-valued probability to each of the $K^{|V_n|}$ possible states of the neighborhood V_n. Without the linear dependence, the transition probability is overly complicated for analysis and requires a significant amount of data for its estimation. In contrast, with the linear dependence, our model has only K^2 parameters instead of $K^{|V_n|+1}$, significantly reducing the model complexity and the amount of data needed for inference.

The Markov chain can exhibit rich dynamics such as synchronization and ergodicity. Figure 1 illustrates two systems with $(N, K, n_v) = (8, 3, 2)$ and with different local transition matrices. Figure 1(b) shows a trajectory exhibiting a transition from deterministic to stochastic dynamics, and Figure 1(c) shows a trajectory exhibiting synchronization.

1.1 Main results

We first study how the local transition matrix determines important features of the global dynamics, by establishing sufficient and necessary conditions on the local transition matrix for synchronization and ergodicity. When the local transition matrix $\mathbf{T}$ is irreducible, Theorem 2.7 shows that the system will achieve a synchronization if and only if $\mathbf{T}$ is periodic with period K, and Theorem 2.10 shows that the system is exponentially ergodic if and only if $\mathbf{T}$ is aperiodic. In short,

$$\mathbf{T} \text{ irreducible} : \begin{cases} \text{periodic with period } K & \Leftrightarrow \text{synchronization} \\ \text{aperiodic} & \Leftrightarrow \text{exponential ergodicity.} \end{cases}$$

We then study the dependencies between the global transition probability matrix of the Markov chain and the local transition matrix. Theorem 3.2 shows that there is a 1-1 map between the global transition matrix $\mathbf{P} \in \mathbb{R}^{K^N \times K^N}$ and the local transition matrix; additionally, the invariant measure of $\mathbf{T}$ is the marginal distribution of the invariant measure of $\mathbf{P}$. Theorem 3.5 shows that $\mathbf{P}$ is Lipschitz in $\mathbf{T}$ under the total variation norm.

In Section 4, we introduce and study the properties of least squares estimators (LSEs) for inferring the local transition matrix from various types of data. The data may consist of multiple trajectories, a long trajectory, or ensemble sequences without trajectory information. Except for the case of a long trajectory, the system can be non-ergodic. These LSEs use the marginal distributions of each vertex and are more efficient to compute than the maximal likelihood estimator, which would involve non-convex optimization. We specify identifiability conditions for these LSEs and a non-identifiability for inference using stationary distribution in Section 4.4. We show the asymptotic normality of these estimators in Theorems 4.4 and 4.10 and provide non-asymptotic bounds for their accuracy in Section 4.5. Numerical tests show that the LSE with trajectory information is more accurate than the LSE without trajectory information, while both converge at the optimal rate $M^{-1/2}$, with M being the sample size.

1.2 Related work

Probabilistic cellular automata (PCA). PCA are large interacting discrete stochastic dynamical systems for the modeling of a wide range of physical and societal phenomena, and we refer to [Too94, LMS90, LN18] and the reference therein for the applications. Motivated by the applications, there is consistent interest in studying the ergodicity of the systems; see, e.g., Dawson [Daw75] for a system with interacting subsystems, Follmer and Horst [FH01] for the averaged process of an interacting Markov chain with an infinite set of sites, Bérard [Bér23] for the exponential ergodicity of a 1D PCA with a local transition kernel on a three-state alphabet, and Casse [Cas23] for the ergodicity of a PCA with binary alphabet via random walks. The innovation in our model above is introducing a local transition matrix, which enables efficient estimation, while leaving the system capable of producing rich dynamics, exhibiting phenomena such as synchronization or ergodicity.

Interacting particle systems (IPSs). IPSs are continuous-time Markov processes on certain spaces of configurations of finitely or infinitely many interacting particles. The state space can be either discrete, such as the stochastic Ising model or the voter model [Lig85, Dur07, Ald13, Gri18], or continuous in the form of stochastic differential equations [CM08, CDP18, LRW21]. The interaction rules, either short-range or long-range, are often specified by functions called interaction kernels/potentials [Lig85, CDP18] or rate functions [Ald13]. Thus, our local transition matrix can be viewed as a counterpart of these interaction kernels or rate functions.

Inference of the local transition matrix. The inference of the local transition matrix is akin to the nonparametric estimation of the interaction kernel of interacting particle systems in [LZTM19, LMT21, DMH22, LWLM24], where inference leads to a linear inverse problem and is solved by least squares. However, the estimators in those works maximize the likelihood; here, our LSEs are different from the maximal likelihood estimator, which would lead to a constrained non-convex optimization problem, as discussed in Section 4.1.1. Also, while the identifiability conditions are

Table 1: Notations.

$k \in \mathcal{A} = [K] = \{1, \ldots, K\}$	alphabet set for state values
$n \in [N] = \{1, 2, \ldots, N\}$	index of vertices/agents in the graph
$m \in [M],\ t \in [L]$	index of sample trajectories and index of time
$X(t) \in \mathcal{A}^N$	state of the Markov chain at time t
V_n and n_v	the n-th vertex's neighborhood V_n, consisting of $2n_v + 1$ vertices
$\varphi_{n,t} = (\varphi_{n,t}(1), \ldots, \varphi_{n,t}(K))$	empirical distribution in V_n at time t; $[0,1]^{1\times K}$-valued
$c_{n,t} = (c_{n,t}(1), \ldots, c_{n,t}(K))$	empirical distribution of $X_n(t)$: $c_{n,t}(k) := \delta_{X_n(t)}(k)$
$\mathbf{T} \in [0,1]^{K\times K}$	local transition matrix: $\sum_{k=1}^K \mathbf{T}_{jk} = 1, \forall j$
$\mathbf{P} \in [0,1]^{K^N \times K^N}$	(global) transition matrix of the Markov chain
$\lVert\cdot\rVert$, $\lVert\cdot\rVert_F$, $\lVert\cdot\rVert_{op}$	Euclidean, Frobenius and operator norms

specified based on the large sample limit case, in the same spirit as the coercivity condition on function spaces in [LLM+21, LL23], this study considers parametric inference, so the identifiability conditions are less restrictive.

We use the notations in Table 1 throughout the paper. We denote the entries of $\mathbf{T} \in \mathbb{R}^{K\times K}$ by $\mathbf{T}_{jk}$ with $j, k \in [K]$, and denote the entries of $\mathbf{P} \in \mathbb{R}^{K^N\times K^N}$ by $\mathbf{P}(x, y)$ with $x, y \in [K]^N$, where $x = x_{1:N} := (x_1, \ldots, x_N)$ with $x_i \in [K]$.

2 Dynamical properties: synchronization and ergodicity

This section studies the dynamical properties of the process $X(\cdot)$ in (1.1)-(1.2) as a Markov chain with K^N states, whose transition probability matrix $\mathbf{P} \in \mathbb{R}^{K^N\times K^N}$ is determined by the neighborhood and the local transition matrix $\mathbf{T} \in \mathbb{R}^{K\times K}$. In particular, we show the necessary and sufficient conditions on the local transition matrix for the Markov chain to achieve synchronization or ergodicity. We will study the relation between $\mathbf{P}$ and $\mathbf{T}$, including their associated invariant measures, in Section 3.

To study the Markov chain, we recall the following preliminaries about a finite-state Markov chain, denoted by $Z(t)$, which has states in $[n]$ and transition matrix $A \in \mathbb{R}^{n\times n}$.

- Irreducible matrix. The transition matrix $A \in \mathbb{R}^{n\times n}$ is called *irreducible* if $\forall i, j \in [n]^2$, $\exists t \in \mathbb{N}$, such that $(A^t)_{ij} > 0$.
- Period of a matrix. The period $d(i)$ of a state i is the greatest common divisor of all m such that $(A^m)_{ii} > 0$, i.e., $d(i) := gcd\{m \in \mathbb{N} : (A^m)_{ii} > 0\}$. When A is irreducible, the period of every state is the same and is called the period of A. The irreducible matrix A is called *aperiodic* if its period is one.
- Recurrence and Transience. A state i is called recurrent if $\mathbb{P}\{\tau_i < \infty | Z(0) = i\} = 1$, where $\tau_i = \min\{t \in \mathbb{N} : Z(t) = i\}$; in other words, the chain starting from this state returns to it in finite time with probability one. The state i is called transient if $\mathbb{P}\{\tau_i < \infty | Z(0) = i\} < 1$. The state is called positive recurrent if $\mathbb{E}[\tau_i | Z(0) = i] < \infty$.

In the following, we first present a few examples of the locally interacting Markov chain, showing that the Markov chain can have various dynamical properties. Then, we study the sufficient and necessary conditions for the system to synchronize or to be ergodic.

2.1 Examples: stochastic dynamics and synchronization

We introduce four examples: non-interacting agents, the smallest model, a system transitioning from deterministic to stochastic dynamics, and a system achieving synchronization.

Example 2.1 (Non-interacting agents) *When the agents do not interact, i.e., $n_v = 0$, they move independently according to a Markov chain with $\mathbf{T}$ as the probability transition matrix. That is, the process $X(t) = (X_1(t), X_2(t), \cdots, X_N(t))$ is a vector of N independent Markov chains with the same transition matrix $\mathbf{T}$.*

Example 2.2 (Smallest model: $(N, K) = (2, 2)$) *Consider the model with $(N, K, n_v) = (2, 2, 1)$. Since the neighborhood is the full network for each site, local empirical distributions are the same for all sites, i.e., we have $\phi_1 = \phi_2$ for all states. The following table shows the local empirical distributions and the global transition matrix with a local transition matrix $\mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{pmatrix}$:*

		Global transition probability $\mathbf{P}$			
x	$\phi_1 = \phi_2$	11	12	21	22
11	$(1, 0)$	$\mathbf{T}_{11}^2$	$\mathbf{T}_{11}\mathbf{T}_{12}$	$\mathbf{T}_{11}\mathbf{T}_{12}$	$\mathbf{T}_{12}^2$
12	$(0.5, 0.5)$	$\frac{1}{4}(\mathbf{T}_{11} + \mathbf{T}_{21})^2$	$\mathbf{P}(12, 12)$	$\mathbf{P}(12, 21)$	$\frac{1}{4}(\mathbf{T}_{12} + \mathbf{T}_{22})^2$
21	$(0.5, 0.5)$	$\frac{1}{4}(\mathbf{T}_{11} + \mathbf{T}_{21})^2$	$\mathbf{P}(21, 12)$	$\mathbf{P}(21, 21)$	$\frac{1}{4}(\mathbf{T}_{12} + \mathbf{T}_{22})^2$
22	$(0, 1)$	$\mathbf{T}_{21}^2$	$\mathbf{T}_{11}\mathbf{T}_{12}$	$\mathbf{T}_{11}\mathbf{T}_{12}$	$\mathbf{T}_{22}^2$

where $\mathbf{P}(12, 12) = \mathbf{P}(12, 21) = \mathbf{P}(21, 12) = \mathbf{P}(21, 21) = \frac{1}{4}(\mathbf{T}_{11} + \mathbf{T}_{21})(\mathbf{T}_{12} + \mathbf{T}_{22})$.

Example 2.3 (From deterministic to stochastic dynamics) *The system can change from deterministic to stochastic dynamics. Let $\mathbf{T}$ be such that $\mathbf{T}_{i,i+1} = 1$ for $i = 1, \ldots, K-1$ and $\mathbf{T}_{K,j} = \frac{1}{K}$ for $j = 1, \ldots, K$; that is,*

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \frac{1}{K} & \frac{1}{K} & \frac{1}{K} & \cdots & \frac{1}{K} \end{bmatrix}. \tag{2.1}$$

The system's dynamics will move from deterministic to stochastic if it starts with state $(1, \ldots, 1) \in \mathcal{A}^N$. Specifically, note that we have $\varphi_{n,0} = (1, 0, \cdots, 0) =: e_1 \in \mathbb{R}^{1\times K}$ for each $n \in [N]$. Then, the value of each site moves to the next value deterministically, i.e., $X(t) = (t+1, t+1, \cdots, t+1) \in \mathcal{A}^N$, for $t \leqslant K-1$. Correspondingly, we have $\varphi_{n,t} = e_{t+1} \in \mathbb{R}^{1\times K}$ for each $n \in [N]$ and $t \leqslant K-1$. When $t > K-1$, the move becomes stochastic as the last row of the local transition matrix injects the randomness. Figure 1(b) shows a typical trajectory of the system when $(N, K, n_v) = (8, 3, 2)$.

Example 2.4 (Synchronization: from stochastic to deterministic dynamics) *When $\mathbf{T}$ is a permutation matrix such that it is irreducible with period K, e.g.,*

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{2.2}$$

the Markov chain will achieve synchronization (see Theorem 2.7), in which all sites move from one state to another with the same period as $\mathbf{T}$*, as demonstrated in Figure 1*(c) *for a typical trajectory of the system when* $(N, K, n_v) = (8, 3, 2)$*. In particular, the deterministic dynamics of the Markov chain after synchronization is as follows. Without loss of generality, suppose that it starts from the state* $X(0) = (1, \ldots, 1) \in \mathcal{A}^N$*. The local empirical distributions are* $\varphi_{n,0} = (1, 0, \cdots, 0) =: e_1 \in \mathbb{R}^{1\times K}$ *for each* $n \in [N]$*. Then, all vertices move uniformly from one state to the next, i.e.,* $X(t) = (t+1, \ldots, t+1)$ *for* $t \leqslant K-1$*, and then repeats periodically, as shown in the following tabular.*

	$t=0$	$t=1$	$\cdots$	$t=K-1$	$t=K$	$\cdots$
$X(t)$	$(1,1,\cdots,1)$	$(2,2,\cdots,2)$	$\cdots$	$(K,K,\cdots,K)$	$(1,1,\cdots,1)$	$\cdots$
$\phi_{n,t}$	e_1	e_2	$\cdots$	e_K	e_1	$\cdots$

The corresponding local distributions are $\varphi_{n,t} = e_{t+1} \in \mathbb{R}^{1\times K}$ *for each* $n \in [N]$ *and* $t \leqslant K-1$*.*

2.2 Synchronization

We show first that the system with irreducible local transition matrix $\mathbf{T}$ will synchronize if and only if $\mathbf{T}$ is periodic with period K.

Definition 2.5 (Synchronization) *We say the system achieves a* synchronization *at time* t_0 *if all sites move identically after* t_0*, i.e.,* $X_1(t) = X_2(t) = \cdots = X_N(t)$ *for all* $t \geqslant t_0$*.*

The following proposition says that if $\mathbf{T}$ is irreducible and periodic, then the Markov chain $X(\cdot)$ eventually will be periodic.

Proposition 2.6 *Suppose that* $\mathbf{T}$ *is irreducible and periodic with period* $2 \leqslant d \leqslant K$*. Then*

(i) $[K]$ *can be decomposed as a finite disjoint union* $C_0 \cup \cdots \cup C_{d-1}$*, such that (setting* $C_d = C_0$*),* $\mathbf{T}_{jk} > 0$ *only if* $j \in C_r$ *and* $k \in C_{r+1}$ *for some* r*.*

(ii) For $X(\cdot)$*, the collection of states*

$$A := C_0^{\otimes N} \cup \cdots \cup C_{d-1}^{\otimes N} \subset [K]^N$$

is periodic with period d*. The collection of states* $[K]^N \backslash A$ *is transient.*

Proof. (i) This is a standard result; see e.g. [Nor98, Theorem 1.8.4].

(ii) It follows from (i) that states in A are periodic with period d. It remains to show that states in $[K]^N \backslash A$ are transient. For this, we will show that starting from $X(0) = x_{1:N}$ for any configuration $x_{1:N} = (x_1, \ldots, x_N) \in [K]^N \backslash A$, there is some $t_0 \in \mathbb{N}$ such that $\mathbb{P}\{X(t_0) \in A\} > 0$. Assume without loss of generality that $x_1 \in C_0$. Since $\mathbf{T}$ is irreducible, by part (i), there is some $c_1 \in C_1$ such that $\mathbf{T}_{c_0 c_1} > 0$. Therefore, recalling the neighborhood $V_1 = \{1, 2, \ldots, 1+n_v, N+1-n_v, \ldots, N-1, N\}$, there is a positive probability that $X(1)$ has some configuration in

$$\{(y_1, \ldots, y_N) : y_1 = \cdots = y_{1+n_v} = y_{N+1-n_v} = \cdots = y_N = c_1\},$$

i.e., the neighbors of node 1 have the same state c_1. Again since $\mathbf{T}$ is irreducible, by part (i), there is some $c_2 \in C_2$ such that $\mathbf{T}_{c_1 c_2} > 0$. So there is a positive probability that $X(2)$ has some configuration in

$$\{(y_1, \ldots, y_N) : y_1 = \cdots = y_{1+2n_v} = y_{N+1-2n_v} = \cdots = y_N = c_2\},$$

i.e., the neighbors of neighbors of node 1 have the same state c_2. Continuing in this manner, we see that there is a positive probability for X to jump after at most $\lceil \frac{N}{2n_v} \rceil$ steps to some state in A, i.e., $\mathbb{P}\left\{X(\lceil \frac{N}{2n_v} \rceil) \in A\right\} > 0$. ■

Using Proposition 2.6, we have the following characterization of when synchronization occurs.

Theorem 2.7 (Synchronization) *Suppose that $\mathbf{T}$ is irreducible. Then, the system will achieve a synchronization (with period K) if and only if the period of $\mathbf{T}$ is K.*

Proof. (i) We first prove the "if" direction. Suppose the period of $\mathbf{T}$ is K. Then the decomposition of $[K]$ in Proposition 2.6(i) must have the form that each C_r is a singleton. So the collection of states A in Proposition 2.6(ii) is actually just

$$\{(k, \ldots, k) \in [K]^N : k \in [K]\}. \tag{2.3}$$

Also by Proposition 2.6(ii), all states in $[K]^N \backslash A$ are transient and will eventually jump to some state in A. Therefore, the system will achieve a synchronization with period K.

(ii) Next, we prove the "only if" direction. Suppose the system will achieve synchronization. Then there exists some $k_0 \in [K]$ such that, starting from $(k_0, \ldots, k_0) \in [K]^N$, the system will jump only among the collection of states in (2.3). This implies that $\mathbf{T}_{k_0,k_1} = 1$ for some $k_1 \in [K]\backslash\{k_0\}$. This further implies, for each $r = 1, \ldots, K-2$, $\mathbf{T}_{k_r,k_{r+1}} = 1$ for some $k_{r+1} \in [K]\backslash\{k_0, \ldots, k_r\}$, as otherwise one has $k_{r+1} = k_j$ for some $0 \leqslant j \leqslant r$ and it contradicts with the assumption that $\mathbf{T}$ is irreducible. Finally, again by irreducibility of $\mathbf{T}$, we must have $\mathbf{T}_{k_{K-1},k_0} = 1$. Therefore, the period of $\mathbf{T}$ is K. ■

Remark 2.8 *The arguments in the proof of Proposition 2.6 also reveal that, when $\mathbf{T}$ is reducible, it is still possible that the system will reach a synchronization. For example:*

(i) If $\mathbf{T}$ has some transient states, such as $\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$, then eventually, the system will reach a synchronization and oscillate between $(2, \ldots, 2), (3, \ldots, 3) \in [K]^N$.

(ii) If $\mathbf{T}$ has more than one communication class, such as $\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$, then eventually the system will reach a synchronization and oscillate between the configurations $(1, \ldots, 1), (2, \ldots, 2) \in [K]^N$, or between $(3, \ldots, 3), (4, \ldots, 4) \in [K]^N$.

2.3 Ergodicity

We show that the system with irreducible local transition matrix $\mathbf{T}$ is ergodic if and only if $\mathbf{T}$ is (irreducible and) aperiodic.

Proposition 2.9 *The global transition matrix $\mathbf{P}$ is irreducible and aperiodic if and only if the stochastic matrix $\mathbf{T}$ is irreducible and aperiodic.*

Proof. (i) We first prove the "if" direction: We will show that $\mathbf{P}$ is irreducible and aperiodic when $\mathbf{T}$ is irreducible and aperiodic. It suffices to show that there exists l_0 such that for all $l \geqslant l_0$, $\mathbf{P}^l(x, y) > 0$ for all $x = (j_1, \ldots, j_N), y = (k_1, \ldots, k_N) \in [K]^N$.

Since $\mathbf{T}$ is irreducible and aperiodic, there exists some l_0 such that for all $l \geqslant l_0$ and $j, k \in [K]$, $(\mathbf{T}^l)_{jk} > 0$. Fix such an l. For each $n \in [n]$, $(\mathbf{T}^l)_{j_n k_n} > 0$ implies that there exists a sequence $i_1(j_n, k_n), \ldots, i_l(j_n, k_n) \in [K]$ such that $\mathbf{T}_{j i_1(j_n,k_n)} \mathbf{T}_{i_1(j_n,k_n) i_2(j_n,k_n)} \cdots \mathbf{T}_{i_l(j_n,k_n)k} > 0$. Let $x^{(t)} = (i_t(j_1, k_1), \ldots, i_t(j_N, k_N))$ for $1 \leqslant t \leqslant l$. Then,

$$\mathbf{P}^{(t)}(x, y) \geqslant \mathbf{P}(x, x^{(1)})\mathbf{P}(x^{(1)}, x^{(2)}) \cdots \mathbf{P}(x^{(l)}, y).$$

Meanwhile, noting that the neighborhood of the node n includes itself, we have $\phi_n(x_n) \geqslant \frac{1}{|V_n|}$ for any $x = x_{1:N}$. Thus, (1.4) implies that

$$\mathbf{P}(x_{1:N}, y_{1:N}) \geqslant \prod_{n=1}^{N} \frac{1}{|V_n|} \mathbf{T}_{x_n, y_n}$$

for any states $x_{1:N}$ and $y_{1:N}$. Thus, $\mathbf{P}(x^{(t)}, x^{(t+1)}) \geqslant \prod_{n=1}^{N} \frac{1}{|V_n^t|} \mathbf{T}_{x_n^{(t)}, x_n^{(t)}}$ for each t, where V_n^t is the neighborhood of the n-th agent at time t. Hence,

$$\mathbf{P}^{(l)}(x, y) \geqslant \prod_{n=1}^{N} \left(\mathbf{T}_{j_n i_1(j_n,k_n)} \mathbf{T}_{i_1(j_n,k_n) i_2(j_n,k_n)} \cdots \mathbf{T}_{i_l(j_n,k_n) k_n} \cdot \frac{1}{|V_n^l|} \right) > 0.$$

Thus, $\mathbf{P}$ is also irreducible and aperiodic.

(ii) Next, we prove the "only if" direction. Let $X(\cdot)$ be irreducible and aperiodic.

Suppose $\mathbf{T}$ is reducible. Then there exists $A \subset [K]$ with $A, A^c \neq \varnothing$ such that for all $j \in A, k \in A^c$ and $l \geqslant 1$, $\mathbf{T}_{jk}^{(l)} = 0$. Then for all $x = (j_1, \ldots, j_N) \in A^{\otimes N}, y = (k_1, \ldots, k_N) \in (A^c)^{\otimes N}$ and $l \geqslant 1$, $\mathbf{P}^l(x, y) = 0$, so $\mathbf{P}$ is reducible. This leads to a contradiction and hence $\mathbf{T}$ must be irreducible.

Suppose $\mathbf{T}$ is irreducible but periodic with period $d \geqslant 2$. Then, by Proposition 2.6(ii), $X(\cdot)$ is not irreducible, which is a contradiction. This completes the proof. ∎

RW: XXX Will check whether statements need to be revised if dropping the assumption of irreducible and aperiodic $\mathbf{T}$

Theorem 2.10 *Assume that the stochastic matrix $\mathbf{T}$ is irreducible and aperiodic. Then, $X(\cdot)$ is irreducible, aperiodic, and hence ergodic. In particular, there exists a unique stationary distribution π, all states are positive recurrent, $\lim_{t\to\infty} \mathbb{P}\{X(t) = \cdot\} = \pi(\cdot)$, and there exist $C \in (0, \infty)$, $\rho \in (0, 1)$ and $t_0 \in \mathbb{N}$ such that*

$$\|\mathbf{P}^t(x, \cdot) - \pi\|_{TV} \leqslant C\rho^t, \quad \forall\, t \geqslant t_0,\ x \in [K]^N. \tag{2.4}$$

Proof. By Proposition 2.9, we have that $X(\cdot)$ is irreducible and aperiodic. It then remains to prove (2.4), as the rest of the statements are classic results (see, e.g., Proposition A.1 for completeness).

Since $X(\cdot)$ is irreducible and aperiodic, there exists some $t_0 \in \mathbb{N}$ such that

$$\mathbf{P}^{t_0}(x, y) > 0, \quad \forall\, x, y \in [K]^N.$$

Therefore $|\mathbf{P}^{t_0}(x, y) - \mathbf{P}^{t_0}(x', y)| < \max\{\mathbf{P}^{t_0}(x, y), \mathbf{P}^{t_0}(x', y)\} < \mathbf{P}^{t_0}(x, y) + \mathbf{P}^{t_0}(x', y)$ for any $x, x', y \in [K]^N$. Hence, $\rho_0 := \frac{1}{2} \max_{x,x'} \sum_{y \in [K]^N} \max\{\mathbf{P}^{t_0}(x, y), \mathbf{P}^{t_0}(x', y)\} < \sum_{y \in [K]^N} [\mathbf{P}^{t_0}(x, y) +$

$\mathbf{P}^{t_0}(x',y)] = 1$ and

$$\begin{aligned}\|\mathbf{P}^{t_0}(x,\cdot)-\mathbf{P}^{t_0}(x',\cdot)\|_{TV} &= \sum_{y\in[K]^N} |\mathbf{P}^{t_0}(x,y)-\mathbf{P}^{t_0}(x',y)| \\ &< \sum_{y\in[K]^N} \max\{\mathbf{P}^{t_0}(x,y),\mathbf{P}^{t_0}(x',y)\} \leqslant 2\rho_0, \quad \forall\, x,x'\in[K]^N.\end{aligned}$$

It then follows from Proposition A.2 that

$$\|\mathbf{P}^{kt_0}(x,\cdot)-\pi\|_{TV} \leqslant 2\rho_0^k, \quad \forall\, k\in\mathbb{N},\ x\in[K]^N.$$

For $t\geqslant t_0$, writing $t=\lfloor\frac{t}{t_0}\rfloor t_0+(t-\lfloor\frac{t}{t_0}\rfloor t_0)$, we have

$$\|\mathbf{P}^t(x,\cdot)-\pi\|_{TV} \leqslant 2\rho_0^{\lfloor\frac{t}{t_0}\rfloor} = \frac{2}{\rho_0}\rho_0^{\lfloor\frac{t}{t_0}\rfloor+1} \leqslant \frac{2}{\rho_0}(\rho_0^{1/t_0})^t, \quad \forall\, x\in[K]^N.$$

This gives (2.4) with $C=\frac{2}{\rho_0}\in(0,\infty)$ and $\rho=\rho_0^{1/t_0}\in(0,1)$. ■

The following corollary is a particular case of Theorem 2.10, with $t_0=1$ in its proof.

Corollary 2.11 *Suppose $\mathbf{T}_{jk}>0$ for all $j,k\in[K]$. Then all statements in Theorem 2.10 hold. In fact, there exists $\rho\in(0,1)$ such that*

$$\|\mathbf{P}^t(x,\cdot)-\pi\|_{TV} \leqslant 2\rho^t, \quad \forall\, t\geqslant 1,\ x\in[K]^N.$$

Remark 2.12 *The above results characterize the long-time behavior of the system when the local transition matrix $\mathbf{T}$ is irreducible. In particular, Theorem 2.7 shows that the system will achieve a synchronization if and only if $\mathbf{T}$ is periodic with period K, and Theorem 2.10 and Proposition 2.6 show that the system is exponentially ergodic if and only if $\mathbf{T}$ is aperiodic.*

3 Global and local transition matrices

3.1 Local and global transition matrices and associated invariant measures

We show that there exists a 1-1 map between the global and local transition matrices $\mathbf{P}$ and $\mathbf{T}$. Furthermore, we show that $\tilde{\pi}$ is the marginal of π, where $\pi\in\mathbb{R}^{K^N}$ and $\tilde{\pi}\in\mathbb{R}^K$ denote the unique invariant measures of $\mathbf{P}$ and $\mathbf{T}$, respectively. That is, the marginal distribution of the Markov chain's invariant measure is the same as the invariant measure of $\mathbf{T}$. However, the marginal distribution does not determine $\mathbf{T}$, as we show in Example 3.3 that there are two $\mathbf{T}$'s leading to the same $\tilde{\pi}$ and different π's; nor does the joint invariant distribution, as we show in Example 3.4 that there are two $\mathbf{T}$'s leading to the same $\tilde{\pi}$ and π.

Proposition 3.1 (Shift-invariance) *The transition matrix $\mathbf{P}$ is shift invariant. Therefore the invariant measure π is shift-invariant, namely, for each $(x_1,\dots,x_N)\in[K]^N$,*

$$\pi(x_1,\dots,x_N)=\pi(x_2,\dots,x_N,x_1).$$

Proof. Recall the neighborhood $V_n=\{n-n_v,\dots,n,n+1,\dots,n+n_v\}$. Clearly the graph $G=(V,E)$ is invariant under shift of the node indices $\{1,\dots,n\}$. As the transition of $X(\cdot)$ in (1.1) and (1.2) depends on states through the empirical distribution of neighborhood's states, clearly the transition matrix $\mathbf{P}$ is shift invariant. Since the invariant measure π is unique, we also have shift-invariance of π. ■

The next theorem shows that there is a 1-1 correspondence between the local and global transition matrices. In particular, $\mathbf{T}$ is determined by K^2 entries of $\mathbf{P}$.

Theorem 3.2 (1-1 map between P and T, $\tilde{\pi}$ as the marginal of π) *There is a 1-1 map between $\mathbf{T} \in \mathbb{R}^{K\times K}$ and $\mathbf{P} \in \mathbb{R}^{K^N \times K^N}$. Furthermore, denote by π_n the n-th marginal distribution of π, i.e., $\pi_n(k) = \lim_{t\to\infty} \mathbb{P}(X_n(t) = k)$, $\forall\, k = 1, \ldots, K$. Then $\pi_n = \tilde{\pi}$.*

Proof. Eq.(1.4) shows that $\mathbf{T}$ uniquely determines $\mathbf{P}$. It also implies that

$$\mathbf{T}_{jk} = \mathbf{P}(x_{1:N}, y_{1:N})^{1/N}, \text{ with } x_{1:N} = (j, \ldots, j),\, y_{1:N} = (k, \ldots, k).$$

Therefore there is a 1-1 map between $\mathbf{T} \in \mathbb{R}^{K\times K}$ and $\mathbf{P} \in \mathbb{R}^{K^N \times K^N}$.

Next, we prove that the marginal distribution π_n of the invariant measure π of $\mathbf{P}$ is the same as the invariant measure $\tilde{\pi}$ of $\mathbf{T}$. By shift-invariance in Proposition 3.1, $\pi_n = \pi_m$ for all $n, m \in [N]$. Since π is the invariant measure of $\mathbf{P}$, we have $\pi = \pi\mathbf{P}$. Applying (1.4), we can write

$$\pi(y_{1:N}) = \sum_{x_{1:N}\in[K]^N} \pi(x_{1:N}) \prod_{n=1}^{N} \frac{1}{|V_n|} \sum_{n'=n-n_v}^{n+n_v} \mathbf{T}_{x_{n'}y_n}.$$

Note that $\frac{1}{|V_n|}\sum_{n'=n-n_v}^{n+n_v} \mathbf{T}_{x_{n'}y_n}$ depends on $(y_{1:N})$ only through y_n. Summing over $(y_2, \ldots, y_N) \in \{1, \ldots, K\}^{N-1}$, we have

$$\pi_1(y_1) = \sum_{(x_{1:N})\in[K]^N} \pi(x_{1:N}) \frac{1}{|V_1|} \sum_{n'=1-n_v}^{1+n_v} \mathbf{T}_{x_{n'}y_1} = \frac{1}{|V_1|} \sum_{n'=1-n_v}^{1+n_v} \sum_{x_{n'}\in[K]} \pi_{n'}(x_{n'})\mathbf{T}_{x_{n'}y_1}.$$

By symmetry of π, we have

$$\pi_1(y_1) = \sum_{k\in\{1,\ldots,K\}} \pi_1(k)\mathbf{T}_{ky_1},$$

namely $\pi_1 = \tilde{\pi}$. ■

However, one generally does not have π as the product measure of $\tilde{\pi}$. In fact, π is not completely determined by $\tilde{\pi}$. As illustrated in the following example, one can have two different $\mathbf{T}$'s with the same $\tilde{\pi}$ but different π's.

Example 3.3 (Same marginal, different invariant measures, T and P's) *Consider $K = 2$, $N = 2$ and $V_n = \{1, 2\}$ without doubling counting the other vertex.*

(i) If $\mathbf{T} = \begin{bmatrix} 1/2 & 1/2 \\ 1 & 0 \end{bmatrix}$, *then* $\mathbf{P} = \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 9/16 & 3/16 & 3/16 & 1/16 \\ 9/16 & 3/16 & 3/16 & 1/16 \\ 1 & 0 & 0 & 0 \end{bmatrix}$, $\tilde{\pi}(1) = 2/3, \tilde{\pi}(2) = 1/3$, $\pi(1,1) = 10/21, \pi(1,2) = \pi(2,1) = 4/21, \pi(2,2) = 3/21$.

(ii) If $\mathbf{T} = \begin{bmatrix} 3/4 & 1/4 \\ 1/2 & 1/2 \end{bmatrix}$, *then* $\mathbf{P} = \begin{bmatrix} 9/16 & 3/16 & 3/16 & 1/16 \\ 25/64 & 15/64 & 15/64 & 9/64 \\ 25/64 & 15/64 & 15/64 & 9/64 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$, $\tilde{\pi}(1) = 2/3, \tilde{\pi}(2) = 1/3$, $\pi(1,1) = 14/31, \pi(1,2) = \pi(2,1) = 20/93, \pi(2,2) = 11/93$.

An interesting question is whether a $1-1$ map exists between π and $\mathbf{P}$ for the PCA. Clearly $\mathbf{P}$ uniquely determines π. The other direction is not true for a general Markov chain: there can be multiple transition matrices with the same invariant measure, as shown in Example 3.3, where two different $\mathbf{T}$'s leading to the same invariant measure $\widetilde{\pi}$. Given the special structure of $\mathbf{P}$ being determined by a local transition matrix $\mathbf{T} \in \mathbb{R}^{K\times K}$, which has only $K(K-1)$ unknowns, one may question whether the invariant measure $\pi \in \mathbb{R}^{K^N}$ can uniquely determine $\mathbf{T}$ and consequently $\mathbf{P}$. The following example demonstrates that the answer is no.

Example 3.4 (Same invariant measure, different T and P's) *Consider* $(N,K) = (2,2)$ *as in Example 3.3. If* $\mathbf{T} = \begin{bmatrix} 7/12 & 5/12 \\ 5/6 & 1/6 \end{bmatrix}$, *then* $\mathbf{P} = \begin{bmatrix} 49/144 & 35/144 & 35/144 & 25/144 \\ 289/576 & 119/576 & 119/576 & 49/576 \\ 289/576 & 119/576 & 119/576 & 49/576 \\ 25/36 & 5/36 & 5/36 & 1/36 \end{bmatrix}$, *and we still have* $\tilde{\pi}(1) = 2/3, \tilde{\pi}(2) = 1/3,\ \pi(1,1) = 14/31, \pi(1,2) = \pi(2,1) = 20/93, \pi(2,2) = 11/93$. *This example shares the same* $\tilde{\pi}$ *and* π *as in Example 3.3(ii) but it has different* $\mathbf{T}$ *and* $\mathbf{P}$.

3.2 Lipschitz dependence on the local transition matrix

We show the Lipschitz dependence of the global transition matrix and its invariant measure on the local transition matrix.

Theorem 3.5 (Lipschitz dependence on T) *Given* $\mathbf{T}^{(1)}$ *and* $\mathbf{T}^{(2)}$, *let* $\mathbf{P}_1, \mathbf{P}_2$ *be the corresponding global transition matrices and* π_1, π_2 *be the stationary measures. Then*

$$\|\mathbf{P}_1 - \mathbf{P}_2\|_1 \leqslant NK^{N-1} \min\{\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_1, K\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_{1,1}\},$$

where $\|A\|_1 = \sum_{ij} |A_{i,j}|$, *and* $\|\mathbf{T}\|_{1,1} = \sup_{\|\phi\|_1 \leqslant 1} \|\phi \mathbf{T}\|_1$ *is the* 1*-operator norm. Also,*

$$\|\pi_1 - \pi_2\|_{TV} \leqslant \frac{1}{2(1-\tau(\mathbf{P}_1))} \|\mathbf{P}_1 - \mathbf{P}_2\|_1 \leqslant NK^{N-1} \min\{\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_1, K\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_{1,1}\},$$

where $\tau(\mathbf{P}_1) := \frac{1}{2} \max_{x,x'} \|\mathbf{P}_1(x,\cdot) - \mathbf{P}_1(x',\cdot)\|_{TV}$.

Remark 3.6 *The above Lipschitz dependence on* $\mathbf{T}$ *for the invariant measure is local because the constant* $\frac{1}{2(1-\tau(\mathbf{P}_1))}$ *can depend on* $\mathbf{T}^{(1)}$. *Numerical tests, as shown in Figure 2, suggest that the bound for* $\mathbf{P}$ *is optimal, but the bound for* π *may not.*

The proof of Theorem 3.5 is based on the following lemma in [Wal20], which bounds the total variation between two invariant measures by the difference between their transition probability matrices. We refer to the general study on the perturbation of Markov chains in [Sch68, Sen88, HVdH84, CM01].

Lemma 3.7 *For two finite irreducible transition matrices* $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *with stationary distributions* π_1 *and* π_2,

$$\|\pi_1 - \pi_2\|_{TV} \leqslant \frac{1}{1-\tau(\mathbf{P}_1)} \sum_x \|\mathbf{P}_1(x,\cdot) - \mathbf{P}_2(x,\cdot)\|_{TV},$$

where $\tau(\mathbf{P}_1) := \frac{1}{2} \max_{x,x'} \|\mathbf{P}_1(x,\cdot) - \mathbf{P}_1(x',\cdot)\|_{TV}$.

Using this lemma and the observation that $\sum_x \|\mathbf{P}_1(x,\cdot) - \mathbf{P}_2(x,\cdot)\|_{TV} = \frac{1}{2}\|\mathbf{P}_1 - \mathbf{P}_2\|_1$, we have the following proof for the Lipschitz estimate.

Proof of Theorem 3.5. For $x = (x_1, \ldots, x_N) \in [K]^N$ and $y = (y_1, \ldots, y_N) \in [K]^N$, writing

$$z_n \equiv z_n(x,y) = \frac{1}{|V_n|} \sum_{n'=n-n_v}^{n+n_v} \mathbf{T}^{(1)}_{x_{n'},y_n}, \qquad w_n \equiv w_n(x,y) = \frac{1}{|V_n|} \sum_{n'=n-n_v}^{n+n_v} \mathbf{T}^{(2)}_{x_{n'},y_n},$$

we have

$$\mathbf{P}_1(x,y) = \prod_{n=1}^{N} z_n, \qquad \mathbf{P}_2(x,y) = \prod_{n=1}^{N} w_n.$$

So, by adding and subtracting terms,

$$\begin{aligned} |\mathbf{P}_1(x,y) - \mathbf{P}_2(x,y)| &= \left| \sum_{n=1}^{N} w_1 \cdots w_{n-1}(z_n - w_n) z_{n+1} \cdots z_N \right| \\ &\leqslant \sum_{n=1}^{N} |w_1 \cdots w_{n-1}(z_n - w_n) z_{n+1} \cdots z_N| \\ &= \sum_{n=1}^{N} w_1 \cdots w_{n-1} |z_n - w_n| z_{n+1} \cdots z_N. \end{aligned}$$

Noting that z_n, w_n depend on y only through y_n and $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$ are stochastic matrices, we have

$$\sum_{y_n \in [K]} z_n = 1, \qquad \sum_{y_n \in [K]} w_n = 1.$$

Therefore

$$\begin{aligned} \|\mathbf{P}_1 - \mathbf{P}_2\|_1 &= \sum_{x,y} |\mathbf{P}_1(x,y) - \mathbf{P}_2(x,y)| \leqslant \sum_{n=1}^{N} \sum_{x,y} w_1 \cdots w_{n-1} |z_n - w_n| z_{n+1} \cdots z_N \\ &= \sum_{n=1}^{N} \sum_{x} \left(\sum_{y_1} w_1\right) \cdots \left(\sum_{y_{n-1}} w_{n-1}\right) \left(\sum_{y_n} |z_n - w_n|\right) \left(\sum_{y_{n+1}} z_{n+1}\right) \cdots \left(\sum_{y_N} z_N\right) \\ &= \sum_{n=1}^{N} \sum_{x,y_n} |z_n - w_n| \leqslant \sum_{n=1}^{N} \sum_{x,y_n} \frac{1}{|V_n|} \sum_{n'=n-n_v}^{n+n_v} |\mathbf{T}^{(1)}_{x_{n'},y_n} - \mathbf{T}^{(2)}_{x_{n'},y_n}|. \end{aligned}$$

Noting that by symmetry, $\sum_x \frac{1}{|V_n|} \sum_{n'=n-n_v}^{n+n_v} \Delta \mathbf{T}_{x_{n'},y_n} = K^{N-1} \sum_{x_n} \Delta \mathbf{T}_{x_n,y_n}$, we can write the last term as

$$\sum_{n=1}^{N} K^{N-1} \sum_{x_n,y_n} |\mathbf{T}^{(1)}_{x_n,y_n} - \mathbf{T}^{(2)}_{x_n,y_n}| = \sum_{n=1}^{N} K^{N-1} \|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_1 = N K^{N-1} \|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_1.$$

Alternatively, note that

$$\sum_{y_n} |z_n - w_n| = \sum_{y_n} \phi_n(\cdot) |\mathbf{T}^{(1)}_{\cdot,y_n} - \mathbf{T}^{(2)}_{\cdot,y_n}| \leqslant \|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_{1,1}.$$

So, we also obtain

$$\|\mathbf{P}_1 - \mathbf{P}_2\|_1 \leqslant N K^N \|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_{1,1}.$$

By Lemma 3.7 with the fact $\tau(\mathbf{P}_1) \leqslant \frac{1}{2}$, we have the estimate on $\|\pi_1 - \pi_2\|_{TV}$. ■

Remark 3.8 (Bounds in Frobenius norm.) *If entries of $\mathbf{P}_1 - \mathbf{P}_2$ are of a similar order, then vaguely speaking one would expect*

$$\|\mathbf{P}_1 - \mathbf{P}_2\|_2 \approx CK^{-N}\|\mathbf{P}_1 - \mathbf{P}_2\|_1 \leqslant CNK^{-1}\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_1 \leqslant CN\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_2,$$

which agrees with Figure 2. But the above method only leads to a suboptimal estimate as follows:

$$\|\mathbf{P}_1 - \mathbf{P}_2\|_2 \leqslant N(KC_K)^{N/2}\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_2, \tag{3.1}$$

where $C_K := \max\left\{\max_{j\in[K]}\sum_{k\in[K]}(\mathbf{T}^{(1)}_{j,k})^2, \max_{j\in[K]}\sum_{k\in[K]}(\mathbf{T}^{(2)}_{j,k})^2\right\} \leqslant 1$. *To see this, one can adjust the argument in the above proof and get*

$$\begin{aligned}
|\mathbf{P}_1(x,y) - \mathbf{P}_2(x,y)|^2 &= \left|\sum_{n=1}^{N} w_1\cdots w_{n-1}(z_n - w_n)z_{n+1}\cdots z_N\right|^2 \\
&\leqslant N\sum_{n=1}^{N} |w_1\cdots w_{n-1}(z_n - w_n)z_{n+1}\cdots z_N|^2 \\
&= N\sum_{n=1}^{N} w_1^2\cdots w_{n-1}^2|z_n - w_n|^2 z_{n+1}^2\cdots z_N^2.
\end{aligned}$$

Since $\sum_{y_n} z_n^2 = \sum_{y_n}\left|\frac{1}{|V_n|}\sum_{n'=n-n_v}^{n+n_v}\mathbf{T}^{(1)}_{x_{n'},y_n}\right|^2 \leqslant \sum_{y_n}\frac{1}{|V_n|}\sum_{n'=n-n_v}^{n+n_v}\left|\mathbf{T}^{(1)}_{x_{n'},y_n}\right|^2$ *and similarly for* w_n, *i.e.,* $\sum_{y_n} z_n^2 \leqslant C_K, \quad \sum_{y_n} w_n^2 \leqslant C_K$, *we have*

$$\begin{aligned}
\|\mathbf{P}_1 - \mathbf{P}_2\|_2^2 &= \sum_{x,y}|\mathbf{P}_1(x,y) - \mathbf{P}_2(x,y)|^2 \leqslant N\sum_{n=1}^{N}\sum_{x,y} w_1^2\cdots w_{n-1}^2|z_n - w_n|^2 z_{n+1}^2\cdots z_N^2 \\
&\leqslant NC_K^{N-1}\sum_{n=1}^{N}\sum_{x,y_n}|z_n - w_n|^2 \leqslant NC_K^{N-1}\sum_{n=1}^{N}\sum_{x,y_n}\frac{1}{|V_n|}\sum_{n'=n-n_v}^{n+n_v}|\mathbf{T}^{(1)}_{x_{n'},y_n} - \mathbf{T}^{(2)}_{x_{n'},y_n}|^2.
\end{aligned}$$

Then, we obtain (3.1) by using symmetry to write the last term as

$$\begin{aligned}
NC_K^{N-1}\sum_{n=1}^{N}K^{N-1}\sum_{x_n,y_n}|\mathbf{T}^{(1)}_{x_n,y_n} - \mathbf{T}^{(2)}_{x_n,y_n}|^2 &= NC_K^{N-1}K^{N-1}\sum_{n=1}^{N}\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_2^2 \\
&= N^2(KC_K)^{N-1}\|\mathbf{T}^{(1)} - \mathbf{T}^{(2)}\|_2^2.
\end{aligned}$$

4 Inference of the local transition matrix

Inference is the first step in the application of the PCA model. In this section, we construct least squares estimators (LSEs) for inferring the local transition matrix from various types of data. The data may consist of multiple trajectories, a long trajectory, or ensemble sequences without trajectory information in Sections 4.1–4.3. For each of these cases, we specify identifiability conditions and prove that the estimators are asymptotically normal. Additionally, we show $\mathbf{T}$ is non-identifiable from the stationary distribution in Section 4.4. Furthermore, we establish non-asymptotic bounds for the estimators in Section 4.5. At last, numerical tests in Section 4.6 demonstrate that the LSE with trajectory information is more accurate than the LSE without trajectory information, while both converge at the rate $M^{-1/2}$, agreeing with the theory.

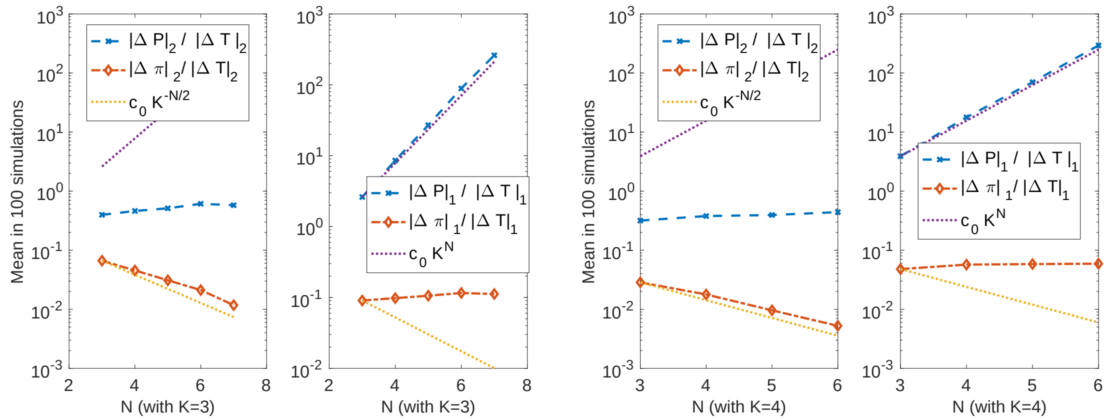


Figure 2: Mean ratios $\frac{\|\Delta\mathbf{P}\|_p}{\|\Delta\mathbf{T}\|_p}$ and $\frac{\|\Delta\pi\|_p}{\|\Delta\mathbf{T}\|_p}$ with $p = 1, 2$, where each (N, K)-pair is computed using 100 random $\mathbf{T}$ with entries sampled from uniform [0,1] followed by a row-normalization. Here the neighborhood has $n_v = \min\{3, \lfloor N/2 \rfloor\}$. Note that $\frac{\|\Delta\mathbf{P}\|_1}{\|\Delta\mathbf{T}\|_1} = O(K^N)$, agreeing with Theorem 3.5. Also, note that $\frac{\|\Delta\pi\|_1}{\|\Delta\mathbf{T}\|_1} = O(1)$, $\frac{\|\Delta\mathbf{P}\|_2}{\|\Delta\mathbf{T}\|_2} = O(1)$, and $\frac{\|\Delta\pi\|_2}{\|\Delta\mathbf{T}\|_2} = O(K^{-N/2})$,

4.1 LSE from multiple trajectories

Consider first the inference of $\mathbf{T} \in \mathbb{R}^{K\times K}$ from data consisting of M independent trajectories:

$$\textbf{Trajectory Data:} \qquad \{X^m(t), t = 1, \dots, L\}_{m=1}^M.$$

We estimate each column of $\mathbf{T}$ through least squares, followed by a normalization. The least squares estimator minimizes the loss function

$$\widehat{\mathbf{T}}_M = \underset{\mathbf{T}\in\mathbb{R}^{K\times K}}{\arg\min}\ \mathcal{E}(\mathbf{T}), \qquad \text{with } \mathcal{E}(\mathbf{T}) := \frac{1}{M}\frac{1}{L}\frac{1}{N}\sum_{m=1}^{M}\sum_{t=1}^{L}\sum_{n=1}^{N}\sum_{k=1}^{K}\left|c_{n,t}^m(k) - \varphi_{n,t-1}^m \mathbf{T}_{\cdot,k}\right|^2, \tag{4.1}$$

where $c_{n,t}^m \in \mathbb{R}^{1\times K}$ denotes the empirical distribution $X_n(t)^m$:

$$c_{n,t}^m = (c_{n,t}^m(1), \dots, c_{n,t}^m(K)) \in \mathbb{R}^{1\times K} \qquad \text{with } c_{n,t}^m(k) := \delta_{X_n^m(t)}(k), \tag{4.2}$$

and $\phi_{n,t}^m \in \mathbb{R}^{1\times K}$ is the empirical distribution of the states of sample $X(t)^m$ in the neighborhood V_n of the vertex n, as defined in (1.3). By solving the zero of this loss function's gradient with respect to each $\mathbf{T}_{\cdot,k}$, we obtain the least squares estimator from a system of K normal equations with a shared normal matrix $\mathbf{A}_M \in \mathbb{R}^{K\times K}$:

$$\begin{aligned} &\widehat{\mathbf{T}}_M(\cdot, k) = \mathbf{A}_M^\dagger \mathbf{b}_M(\cdot, k), \quad 1 \leqslant k \leqslant K, \\ &\mathbf{A}_M = \frac{1}{MLN}\sum_{t,n,m=1}^{L,N,M} (\varphi_{n,t-1}^m)^\top \varphi_{n,t-1}^m, \quad \mathbf{b}_M(\cdot, k) = \frac{1}{MLN}\sum_{t,n,m=1}^{L,N,M} \left(\varphi_{n,t-1}^m\right)^\top c_{n,t}^m(k). \end{aligned} \tag{4.3}$$

Here $\mathbf{A}_M^\dagger$ denotes the Moore-Penrose pseudo inverse of $\mathbf{A}_M$.

In practice, instead of using pseudo-inverse, we obtain $\widehat{\mathbf{T}}_M$ by using least squares with non-negative constraints and then row-normalize the solution. The non-negative constraints help to avoid possible negative entries caused by sampling error.

Identifiability from the large sample limit. To analyze the properties of the estimator, we first examine the inference problem in the large sample limit. Denote the large sample limit normal matrix and normal vectors by

$$\begin{aligned} \mathbf{A}_\infty &= \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N}\mathbb{E}[\varphi_{n,t-1}^\top \varphi_{n,t-1}], \\ \mathbf{b}_\infty(\cdot,k) &= \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N}\mathbb{E}[\varphi_{n,t-1}^\top c_{n,t}(k)], \quad 1 \leqslant k \leqslant K. \end{aligned} \tag{4.4}$$

Assumption 4.1 (Identifiability condition: multi-trajectory data) *The distribution of the samples satisfies that the matrix $\mathbf{A}_\infty$ in (4.4) is non-singular.*

Assumption 4.1 holds in general, as long as the random vectors $\{\phi_{n,t}\}$ span $\mathbb{R}^K$ with a positive probability, as the next lemma shows.

Lemma 4.2 *Assumption 4.1 holds except when there exists a fixed vector $v \in \mathbb{R}^{1\times K}$ such that $v\varphi_{n,t-1}^\top = 0$ a.s. for all n and l. In particular, when $\mathbf{T}$ is irreducible and aperiodic, Assumption 4.1 holds as long as $L \geqslant l_0 := \underset{l\geqslant 0}{\arg\min}\,\{l : \mathbf{P}^l(x,y) > 0, \forall x, y \in [K]^N\}$.*

Proof. Recall that the covariance matrix $\mathbb{E}[Z^\top Z]$ of a random vector $Z \in \mathbb{R}^{1\times K}$ is singular iff there exist a vector $v \in \mathbb{R}^{1\times K}$ such that $Zv^\top = 0$ a.s., which is true because $0 = v\mathbb{E}[Z^\top Z]v^\top = \mathbb{E}[|Zv^\top|^2]$ iff $Zv^\top = 0$ a.s.. Applying this fact to the random vectors $\varphi_{n,t-1}$, we obtain that $\mathbf{A}_\infty$ is singular iff there exists a fixed vector v such that $\varphi_{n,t-1}v^\top = 0$ a.s. for all n and l.

When the $\mathbf{T}$ is irreducible and aperiodic, $\mathbf{P}$ is also irreducible and aperiodic by Proposition 2.9. The number l_0 is well-defined and is finite. Then, no matter what the initial condition is, the states $\{X(t)\}_{t\leqslant l_0}$ visit all possible states with positive probability, so $\{\varphi_{n,t}\}_{t\leqslant l_0}$ span $\mathbb{R}^K$ with a positive probability. Hence, one can never find a v such that $\varphi_{n,t_0}v^\top = 0$ a.s. for all n. ∎

The exceptions are extreme. For the system in either Example 2.3 or Example 2.4, the normal matrix $\mathbf{A}_\infty$ is singular when $L \leqslant K-2$, and is non-singular once $L \geqslant K-1$. For the system in Example 2.3, since $\varphi_{n,t} = e_{l+1} \in \mathbb{R}^{K\times 1}$ for each $n \in [N]$ and $0 \leqslant l \leqslant K-2$, we can take $v = (0,\ldots,0,1) = e_K$ so that $v^\top\varphi_{n,t} = 0$ for all n and $l \leqslant K-2$. In this case, $\mathbf{A}_\infty$ is singular for any $L \leqslant K-2$. On the other hand, if $L \geqslant K$, the resulting normal matrix $\mathbf{A}_\infty$ becomes full rank.

Lemma 4.3 *Under Assumption 4.1, $\mathbf{T} = \mathbf{A}_\infty^{-1}\mathbf{b}_\infty$ with $\mathbf{A}_\infty$ and $\mathbf{b}_\infty$ in (4.4).*

Proof. For each $1 \leqslant k \leqslant K$, recall that $\varphi_{n,t-1}(k) = \frac{1}{|V_n|}\sum_{j\in V_n}\delta_{X_j(t-1)}(k)$ in (1.3) and $c_{n,t}(k) = \delta_{X_n(t)}(k)$ in (4.2). Then,

$$\mathbb{E}[\varphi_{n,t-1}^\top c_{n,t}(k)] = \mathbb{E}\left[\varphi_{n,t-1}^\top \mathbb{E}[\delta_{X_n(t)}(k) \mid X(t-1)]\right] = \mathbb{E}\left[\varphi_{n,t-1}^\top\varphi_{n,t-1}\mathbf{T}_{\cdot,k}\right],$$

where the second equality follows from $\mathbb{E}[\delta_{X_n(t)}(k) \mid X(t-1)] = \mathbb{P}\{X_n(t) = k|X(t-1)\} = \varphi_{n,t-1}\mathbf{T}_{\cdot,k}$ by (1.2). Hence,

$$\mathbf{b}_\infty(\cdot,k) = \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N}\mathbb{E}[\varphi_{n,t-1}^\top c_{n,t}(k)] = \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N}\mathbb{E}\left[\varphi_{n,t-1}^\top\varphi_{n,t-1}\right]\mathbf{T}_{\cdot,k}.$$

In other words, $\mathbf{b}_\infty = \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N}\mathbb{E}\left[\varphi_{n,t-1}^\top\varphi_{n,t-1}\right]\mathbf{T} = \mathbf{A}_\infty\mathbf{T}$. Therefore, $\mathbf{T} = \mathbf{A}_\infty^{-1}\mathbf{b}_\infty$, where $\mathbf{A}_\infty$ is invertible by Assumption 4.1. ∎

Asymptotic normality. We show next that the LSE is asymptotically normal.

Theorem 4.4 *Under Assumption 4.1, for each $k \in [K]$, the estimator $\widehat{\mathbf{T}}_M(\cdot, k)$ in (4.3) is asymptotically normal; that is,*

$$\sqrt{M}\left(\widehat{\mathbf{T}}_M(\cdot, k) - \mathbf{T}(\cdot, k)\right) \to \mathcal{N}(0, \mathbf{A}_\infty^{-1}\Sigma_k\mathbf{A}_\infty^{-1}), \tag{4.5}$$

where the covariance matrix $\Sigma_k \in \mathbb{R}^{K\times K}$ is,

$$\Sigma_k = \frac{1}{L^2N^2}\sum_{t,t'=1}^{L}\sum_{n,n'=1}^{N}\mathbb{E}[(c_{n,t}(k)\varphi_{n,t-1} - \mathbb{E}[c_{n,t}(k)\varphi_{n,t-1}])\left(c_{n',t'}(k)\varphi_{n',t'-1}^\top - \mathbb{E}[c_{n',t'}(k)\varphi_{n',t'-1}^\top]\right)]. \tag{4.6}$$

Proof. For $1 \leqslant k \leqslant K$, denote

$$\mathbf{A}_{L,N}^m := \frac{1}{LN}\sum_{t,n}^{L,N}(\varphi_{n,t-1}^m)^\top\varphi_{n,t-1}^m\,, \quad \mathbf{b}_{L,N}^m(\cdot, k) : = \frac{1}{LN}\sum_{t,n}^{L,N}c_{n,t}(k)\varphi_{n,t-1}^m\,, \tag{4.7}$$

where $c_{n,t}(k)$ is defined in (4.2). Note that $\{\mathbf{A}_{L,N}^m\}_{m=1}^M$ and $\{\mathbf{b}_{L,N}^m(\cdot,k)\}_{m=1}^M$ are independent identically distributed, and $\mathbf{A}_M = \frac{1}{M}\sum_{m=1}^M\mathbf{A}_{L,N}^m$ and $\mathbf{b}_M(\cdot, k) = \frac{1}{M}\sum_{m=1}^M\mathbf{b}_{L,N}^m(\cdot, k)$. Thus, $\mathbf{A}_M \xrightarrow{a.s.} \mathbf{A}_\infty$ by the strong law of large numbers, and

$$\sqrt{M}(\mathbf{b}_M(\cdot, k) - \mathbf{b}_\infty(\cdot, k)) \xrightarrow{d} \mathcal{N}(0, \Sigma_k)$$

for each k, where the matrix Σ_k in (4.6) is the covariance matrix of $\mathbf{b}_{L,N}^m(\cdot, k)$:

$$\Sigma_k = \mathbb{E}[\left(b_{L,N}^m(\cdot, k) - \mathbb{E}[b_{L,N}^m(\cdot, k)]\right)\left(b_{L,N}^m(\cdot, k) - \mathbb{E}[b_{L,N}^m(\cdot, k)]\right)^\top].$$

Then, by Lemma 4.5, we have $\mathbf{A}_M^\dagger\mathbf{b}_M(\cdot, k) - \mathbf{A}_\infty^{-1}\mathbf{b}_\infty(\cdot, k) \xrightarrow{d} \mathcal{N}(0, \mathbf{A}_\infty^{-1}\Sigma_k\mathbf{A}_\infty^{-1})$. Now, the asymptotic normality of the LSE in (4.5) follows from the definition of $\mathbf{T}_M(\cdot, k) = \mathbf{A}_M^\dagger\mathbf{b}_M(\cdot, k)$ in (4.3) and Lemma 4.3. ∎

The following lemma is a slight extension of Slutsky's theorem; we will use it repeatedly to study the asymptotic normality of least squares estimators. Its proof is included for completeness.

Lemma 4.5 *Suppose that $A_M \xrightarrow{a.s.} A$ and $b_M - b \xrightarrow{d} N(0, B)$ as $M \to \infty$, where A, B are two symmetric strictly positive definite matrices. Then, $A_M^\dagger b_M$ is asymptotically normal, i.e., $A_M^\dagger b_M - A^{-1}b \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$.*

Proof. First, we have $A_M^\dagger \xrightarrow{a.s.} A^{-1}$ since A is invertible and $A_M \xrightarrow{a.s.} A$. Specifically, the almost sure convergence of A_M implies that $\mathbb{P}\{\lim_{M\to\infty}\|A_M - A\| = 0\} = 1$. Then, Weyl's inequality $|\lambda_{min}(A_M) - \lambda_{min}(A)| \leqslant \|A_M - A\|$, where $\lambda_{min}(A)$ denotes the minimal eigenvalue of A, implies that $\mathbb{P}\{\lim_{M\to\infty}\lambda_{min}(A_M) = \lambda_{min}(A) > 0\} = 1$. Thus, $A_M^\dagger \xrightarrow{a.s.} A^{-1}$.

Next, combining $A_M^\dagger \xrightarrow{a.s.} A^{-1}$ with $\sqrt{M}(b_M - b) \xrightarrow{d} \mathcal{N}(0, B)$, we have, by Slutsky's theorem, $A_M^\dagger(b_M - b) \xrightarrow{d} A^{-1}\mathcal{N}(0, B) = \mathcal{N}(0, A^{-1}BA^{-1})$ and $A_M^\dagger b \xrightarrow{a.s.} A^{-1}b$. Therefore, using that $A_M^\dagger b_M = A_M^\dagger(b_M - b) + A_M^\dagger b$, we obtain $A_M^\dagger b_M - A^{-1}b \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1})$. ∎

4.1.1 Comparison with MLE and LSE using stochasticity

We discuss two other estimators, an LSE using stochasticity and the maximal likelihood estimator, and compare them with the above LSE. We show that the LSE using stochasticity is similar to the above LSE in theory, but the above LSE is computationally more stable and efficient. The maximal likelihood estimator (MLE) involves a constrained non-convex optimization, making it less effective than the LSE.

Least squares estimator using stochasticity. We can also estimate $\widetilde{\mathbf{T}}$, the first $K-1$ columns of $\mathbf{T}$, by least squares, since $\mathbf{T}$ is a stochastic matrix. We call this estimator *LSE using stochasticity.*

In comparison with the LSE above, this estimator has two advantages: (1) it does not need the extra normalization step, and (2) it has $K(K-1)$ parameters to be estimated, removing the redundancy. However, these advantages may be offset by its computational drawbacks when K is large: it solves a linear system with $K(K-1)$ equations, with a $K(K-1)\times K(K-1)$ normal matrix prone to be ill-conditioned, while the above LSE solves K linear systems with K equations, which can be done in parallel. Moreover, as shown below, the two estimators share the same identifiability condition and asymptotic behavior. Thus, the LSE is preferred in practice.

Specifically, the array $\widetilde{\mathbf{T}} \in \mathbb{R}^{(K-1)\times K}$ is the first $K-1$ column of $\mathbf{T}$. Note that

$$\varphi^m_{n,t}\mathbf{T} = \left[\varphi^m_{n,t}\widetilde{\mathbf{T}},\ 1-\varphi^m_{n,t}\widetilde{\mathbf{T}}\mathbf{1}^\top\right]$$

where $\mathbf{1} = (1,1,\ldots,1) \in \mathbb{R}^{1\times(K-1)}$, and the last column follows from that

$$[\varphi^m_{n,t}\mathbf{T}](K) = \sum_{j=1}^{K}\mathbf{T}_{K,j}\varphi^{j,m}_{n,t} = \sum_{j=1}^{K}(1-\sum_{k=1}^{K-1}\mathbf{T}_{k,j})\varphi^{j,m}_{n,t} = 1-\varphi^m_{n,t}\widetilde{\mathbf{T}}\mathbf{1}^\top.$$

Then, the loss function becomes

$$\mathcal{E}(\mathbf{T}) = \mathcal{E}(\widetilde{\mathbf{T}}) := \frac{1}{MLN}\sum_{t,n,m=1}^{L,N,M}\left(\left\|c^{1:K-1,m}_{n,t} - \varphi^m_{n,t-1}\widetilde{\mathbf{T}}\right\|^2_{\mathbb{R}^{K-1}} + |c^{K,m}_{n,t} - 1 + \varphi^m_{n,t-1}\widetilde{\mathbf{T}}\mathbf{1}^\top|^2\right).$$

This loss function is quadratic, so its minimizer can be found from the zeros of its gradient. Thus, we write $\widetilde{\mathbf{T}}$ in vector form $\overrightarrow{T} \in \mathbb{R}^{(K-1)K\times 1}$ and write the gradient of the loss function as

$$\nabla_{\overrightarrow{T}}\mathcal{E}(\widetilde{\mathbf{T}}) = \widetilde{\mathbf{A}}_M\overrightarrow{T} - \widetilde{\mathbf{b}}_M = \mathbf{0} \in \mathbb{R}^{(K-1)K},$$

where the normal matrix and vectors are, with a notation $s_{kj} := (k-1)K + j$ and similarly $s_{k'j'}$ for $1 \leqslant k,k' \leqslant K-1$ and $1 \leqslant j,j' \leqslant K$,

$$\begin{aligned}\widetilde{\mathbf{A}}_M(s_{kj}, s_{k'j'}) &= \frac{1}{MLN}\sum_{m=1}^{M}\sum_{t=1}^{L}\sum_{n=1}^{N}(1+\delta_{kk'})\varphi^{j',m}_{n,t-1}\varphi^{j,m}_{n,t-1},\\ \widetilde{\mathbf{b}}_M(s_{kj}) &= \frac{1}{MLN}\sum_{m=1}^{M}\sum_{t=1}^{L}\sum_{n=1}^{N}\left(1+c^{k,m}_{n,t} - c^{K,m}_{n,t}\right)\varphi^{j,m}_{n,t-1}.\end{aligned} \tag{4.8}$$

The resulting estimator is

$$\widehat{\mathbf{T}}^{lse}_{MLN} \leftrightarrow \widehat{\overrightarrow{T}} = \widetilde{\mathbf{A}}^{\dagger}_M\widetilde{\mathbf{b}}_M. \tag{4.9}$$

The invertibility of the matrix $\widetilde{\mathbf{A}}_M \in \mathbb{R}^{(K-1)K\times(K-1)K}$ is the same as the matrix $\mathbf{A}_M \in \mathbb{R}^{K\times K}$ in (4.3). We can write

$$\widetilde{\mathbf{A}}_M = \mathbf{A}_M \otimes \mathbf{B},$$

where $\mathbf{A}\otimes\mathbf{B}$ is a Kronecker product

$$\mathbf{A}\otimes\mathbf{B} = \begin{bmatrix} 2\mathbf{A} & \mathbf{A} & \cdots & \mathbf{A} \\ \mathbf{A} & 2\mathbf{A} & \cdots & \mathbf{A} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A} & \mathbf{A} & \cdots & 2\mathbf{A} \end{bmatrix}, \quad \text{with } \mathbf{B} = \begin{bmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{bmatrix} \in \mathbb{R}^{(K-1)\times(K-1)}.$$

Recall that the Kronecker product $\mathbf{A}\otimes\mathbf{B}$ is invertible iff both $\mathbf{A}$ and $\mathbf{B}$ are invertible, and $(\mathbf{A}\otimes\mathbf{B})^{-1} = \mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$. Note that $\mathbf{B}$ is invertible with eigenvalues $\lambda_1^B = K$ and $\lambda_k^B = 1$ for $k = 2,\ldots,K-1$. Hence, $\widetilde{\mathbf{A}}_M$ is nonsingular iff $\mathbf{A}_M$ is nonsingular.

Maximal likelihood estimator. We can also estimate the local transition matrix by maximizing the likelihood of data:

$$\widehat{\mathbf{T}} = \underset{\mathbf{T}\text{ stochastic}}{\arg\min}\ \mathcal{E}_{mle}(\mathbf{T}), \quad \text{where } \mathcal{E}_{mle}(\mathbf{T}) := \frac{1}{M}\frac{1}{L}\frac{1}{N}\sum_{m=1}^{M}\sum_{t=1}^{L}\sum_{n=1}^{N} -\log(\varphi_{n,t-1}^m \mathbf{T}_{\cdot,X_n^m(t)}). \tag{4.10}$$

Note that, unlike the LSE, it is necessary to consider the optimization with respect to a stochastic matrix $\mathbf{T}$, because otherwise, the likelihood has no maximum.

The derivative of the loss function can be computed directly. Using the fact that $\varphi_{n,t}^m\mathbf{T}_{\cdot,K} = \sum_{j=1}^{K-1}\varphi_{n,t}^m(j)(1-\mathbf{T}_{jK})$, we have, for each $1\leqslant k\leqslant K-1$ and $j\in[K]$,

$$\frac{\partial\mathcal{E}_{mle}}{\partial\mathbf{T}_{Kj}}(\mathbf{T}) = \frac{1}{M}\frac{1}{L}\frac{1}{N}\sum_{m=1}^{M}\sum_{t=1}^{L}\sum_{n=1}^{N}\left[-\frac{1}{\varphi_{n,t}^m\mathbf{T}_{\cdot,k}}\delta_{X_n^m(t)}(k)\varphi_{n,t}^{j,m} + \frac{1}{\varphi_{n,t}^m\mathbf{T}_{\cdot,K}}\delta_{X_n^m(t)}(K)\varphi_{n,t}^{j,m}\right].$$

Clearly, even with the above gradient, the uniqueness of the minimizer for the constrained optimization of a nonconvex function is relatively complicated for analysis. Thus, the asymptotic normality of the MLE is non-trivial since it relies on uniqueness. Also, while optimization algorithms can easily compute the minimizer, it remains open to provide a performance guarantee.

4.2 LSE for a single long trajectory for ergodic systems

Suppose that the system is ergodic, and we estimate the local transition matrix from data consisting of a long trajectory:

$$\textbf{A long trajectory Data:} \qquad \{X(t), t = 1,\ldots,L\}.$$

Under the ergodicity assumption, the estimation is the same as the previous case with $M = 1$, and we define the estimator by

$$\begin{aligned} \widehat{\mathbf{T}}_L(\cdot,k) &= \mathbf{A}_L^\dagger \mathbf{b}_L(\cdot,k), \quad 1\leqslant k\leqslant K, \\ \mathbf{A}_L &= \frac{1}{LN}\sum_{t,n=1}^{L,N}\varphi_{n,t-1}^\top\varphi_{n,t-1}, \quad \mathbf{b}_L(\cdot,k) = \frac{1}{LN}\sum_{t,n=1}^{L,N}(\varphi_{n,t-1})^\top c_{n,t}(k). \end{aligned} \tag{4.11}$$

Here $\mathbf{A}_L^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{A}_L$.

Similar to the previous section, the large sample limit helps us specify the identifiability condition. Denote the large sample limit normal matrix and normal vectors by

$$\begin{aligned}\mathbf{A}_\infty &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\varphi_{n,1}^\top\varphi_{n,1}] = \lim_{L\to\infty}\frac{1}{NL}\sum_{t,n=1}^{L,N}\varphi_{n,t}^\top\varphi_{n,t}\,,\\ \mathbf{b}_\infty(\cdot,k) &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\varphi_{n,1}^\top c_{n,1}(k)] = \lim_{L\to\infty}\frac{1}{NL}\sum_{t,n=1}^{L,N}\varphi_{n,t}^\top c_{n,t}(k), \quad 1\leqslant k\leqslant K\,,\end{aligned} \tag{4.12}$$

where the expectation is with respect to the stationary measure of the Markov chain.

Theorem 4.6 *Assume that $\mathbf{A}_\infty$ in (4.12) is non-singular. Then, for each $k\in[K]$, the estimator $\widehat{\mathbf{T}}_L(\cdot,k)$ in (4.11) is asymptotically normal:*

$$\sqrt{L}\left(\widehat{\mathbf{T}}_L(\cdot,k)-\mathbf{T}(\cdot,k)\right)\to\mathcal{N}(0,\mathbf{A}_\infty^{-1}\Sigma_k\mathbf{A}_\infty^{-1}), \tag{4.13}$$

as $L\to\infty$, where the covariance matrix Σ_k is, with $\overline{\Phi_n(k)} := \lim_{L\to\infty}\frac{1}{L}\sum_{t=1}^{L}[c_{n,t}(k)\varphi_{n,t-1}]$,

$$\Sigma_k = \lim_{L\to\infty}\frac{1}{L^2N^2}\sum_{t,t'=1}^{L}\sum_{n,n'=1}^{N}\left(c_{n,t}(k)\varphi_{n,t-1}-\overline{\Phi_n(k)}\right)\left(c_{n',t'}(k)\varphi_{n',t'-1}^\top-\overline{\Phi_n(k)}^\top\right). \tag{4.14}$$

Proof. We omit the proof since it is nearly identical to the case of multiple trajectories except for applying the law of large numbers and the central limit theorem for an ergodic trajectory. ■

4.3 LSE from ensemble data without trajectory

Another interesting setting is when the observations are $M_{n,t}$ samples of $X(t)$ for each time t, but these samples may come from different trajectories. We call this setting as *ensemble data without trajectory information* and denote the data by

$$\textbf{Ensemble Data:}\quad \{X_n^m(t), m=1,\ldots,M_{n,t}\}_{n=1,l=0}^{N,\,L};\quad M=\min_{1\leqslant n\leqslant N,1\leqslant t\leqslant L}\{M_{n,t}\}.$$

We assume that $M=\min_{1\leqslant n\leqslant N,1\leqslant t\leqslant L}\{M_{n,t}\}$ is large, and study error bounds with respect to M.

We estimate $\mathbf{T}$ by least squares matching the empirical marginal densities of each site. Recall that the marginal density of site n at time t is, for $1\leqslant k\leqslant K$,

$$\begin{aligned}p_{n,t}(k) &= \mathbb{P}\{X_n(t)=k\}\\ &= \mathbb{E}[\mathbb{P}\{X_n(t)=k|X(t-1)\}] = \mathbb{E}[\varphi_{n,t-1}\mathbf{T}_{\cdot,k}] = \mathbb{E}[\varphi_{n,t-1}]\mathbf{T}_{\cdot,k}.\end{aligned} \tag{4.15}$$

Thus, our estimator is based on empirical approximations of $p_{n,t}(k)$ and $\mathbb{E}[\varphi_{n,t-1}]$:

$$\begin{aligned}\widehat{p}_{n,t,M}(k) &:= \frac{1}{M_{n,t}}\sum_{m=1}^{M_{n,t}}\delta_{X_n^m(t)}(k), \quad 1\leqslant k\leqslant K,\\ \widehat{\varphi}_{n,t-1,M} &:= \frac{1}{M_{n,t-1}}\sum_{m=1}^{M_{n,t-1}}\varphi_{n,t-1}^m\in\mathbb{R}^{1\times K},\end{aligned} \tag{4.16}$$

for $1\leqslant t\leqslant L$. Note that they are determined by the empirical distributions at each time, and there is no need for sample trajectories. The sample sizes $\{M_{n,t}\}$ do not have to be the same at different times, as long as their minimum M is large enough to make these empirical approximations reasonably accurate.

Our least squares estimator, called *ensemble LSE*, minimizes the discrepancy between the empirical approximations $\widehat{p}_{n,t+1,M}$ and $\widehat{\varphi}_{n,t-1,M}$ in (4.16):

$$\widehat{\mathbf{T}}^e_M = \underset{\mathbf{T}\in\mathbb{R}^{K\times K}}{\arg\min} \sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{t=1}^{L} \|\widehat{p}_{n,t,M}(k) - \widehat{\varphi}_{n,t-1,M}\mathbf{T}_{\cdot,k}\|^2. \tag{4.17}$$

The ensemble LSE is solved by

$$\begin{aligned}
\widehat{\mathbf{T}}^e_M(\cdot,k) &= (\mathbf{A}^e_M)^{\dagger}\mathbf{b}^e_M(\cdot,k), && 1\leqslant k\leqslant K,\\
\mathbf{A}^e_M &= \frac{1}{LN}\sum_{t,n=1}^{L,N} \widehat{\varphi}^{\top}_{n,t-1,M}\widehat{\varphi}_{n,t-1,M} \in \mathbb{R}^{K\times K}, &&\\
\mathbf{b}^e_M(\cdot,k) &= \frac{1}{LN}\sum_{t,n=1}^{L,N} \widehat{\varphi}^{\top}_{n,t-1,M}\widehat{p}_{n,t,M}(k) \in \mathbb{R}^{K\times 1}, && 1\leqslant k\leqslant K.
\end{aligned} \tag{4.18}$$

Similar to the multi-trajectory LSE in Section 4.1, in practice, we obtain the ensemble LSE by least squares with non-negative constraints, followed by row-normalization.

This LSE can be viewed as a generalized moment estimator since the entries in the normal matrix and normal vector are approximations of moments. We will show that the estimator is asymptotically normal under a new identifiability condition.

Identifiability in the large sample limit. Denote the large sample limit normal matrix and normal vectors by

$$\begin{aligned}
\mathbf{A}^e_\infty &= \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N} \mathbb{E}[\varphi_{n,t-1}]^{\top}\mathbb{E}[\varphi_{n,t-1}],\\
\mathbf{b}^e_\infty(\cdot,k) &= \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N} \mathbb{E}[\varphi_{n,t-1}]^{\top}p_{n,t}(k), \quad 1\leqslant k\leqslant K.
\end{aligned} \tag{4.19}$$

Assumption 4.7 (Identifiability condition: ensemble data) *The distribution of the samples satisfies that the matrix $\mathbf{A}^e_\infty$ in (4.19) is nonsingular.*

Lemma 4.8 *Under Assumption 4.7, $\mathbf{T} = (\mathbf{A}^e_\infty)^{-1}\mathbf{b}^e_\infty$.*

Proof. It follows directly from (4.15) and the defintion of $\mathbf{A}^e_\infty$ and $\mathbf{A}^e_\infty$ in (4.19). ∎

This assumption puts constraints on both the distribution of the process and the local empirical distributions $\{\varphi_n\}$ that depend on the neighborhood size of interaction. Three factors can contribute to the identifiability: a non-symmetric initial distribution between sites, a neighborhood that can lead to varying local empirical distributions, and a process that varies in time. For example, $\mathbf{A}^e_\infty$ can be full rank if $\{\mathbb{E}[\varphi_{n,0}]\}_{n=1}^N$ has rank K, which relies on a diverse initial distribution and local empirical distribution. Example 4.9 below shows an extreme case that $\{\varphi_n\}$ are the same for all sites, and we rely on the distribution at different times to attain a full-rank norm matrix.

Example 4.9 (Full-network neighborhood) *The local empirical distributions are the same for all sites if the neighborhood is the whole network for each agent. For example, the model in Example 2.2 has $\varphi_1 = \varphi_2$ for all states $\mathbf{x}$. Thus, for full-network neighborhood, we have $\mathbb{E}[\varphi_{n,t}] = \mathbb{E}[\varphi_{n',t}]$ for any $n \neq n'$. Then, we have $\mathbf{A}^e_\infty = \frac{1}{L}\sum_{t=1}^{L}\mathbb{E}[\varphi_{1,l-1}]^\top\mathbb{E}[\varphi_{1,l-1}]$ and it is full rank only if $\{\mathbb{E}[\varphi_{1,l-1}]\}_{t=1}^{L}$ has rank K. As discussed later, the normal matrix has rank 1 when the process is stationary.*

Asymptotic normality. We show next that the ensemble LSE is asymptotically normal.

Theorem 4.10 *Under Assumption 4.7, for each $k \in [K]$, the estimator $\widehat{\mathbf{T}}^e_M(\cdot,k)$ in (4.18) is asymptotically normal; that is,*

$$\sqrt{M}\left(\widehat{\mathbf{T}}^e_M(\cdot,k) - \mathbf{T}(\cdot,k)\right) \to \mathcal{N}(0, (\mathbf{A}^e_\infty)^{-1}\Sigma^e_k(\mathbf{A}^e_\infty)^{-1}),$$

where the covariance matrix Σ^e_k is,

$$\Sigma^e_k = \frac{1}{L^2N^2}\sum_{n,t=0}^{n,t-1} p_{n,t}(k)^2\mathbb{E}[\varphi_{n,t-1}]^\top\mathbb{E}[\varphi_{n,t-1}]) \in \mathbb{R}^{K\times K}. \tag{4.20}$$

Proof. The proof is based on the asymptotic properties of the empirical approximations of $\widehat{p}_{n,t,M}(k)$ and $\widehat{\varphi}_{n,t-1,M}$ defined in (4.16).

First, by the strong Law of Large Numbers,

$$\widehat{\varphi}_{n,t-1,M} \xrightarrow{a.s.} \mathbb{E}[\varphi_{n,t-1}]$$

as $M \to \infty$, for each n, l. Also, by Central Limit Theorem,

$$\widehat{p}_{n,t,M}(k) - p_{n,t}(k) \xrightarrow{d} \mathcal{N}(0, \sigma_{n,t}(k)), \tag{4.21}$$

for each n, l, where the variance $\sigma_{n,t}(k)$ follows from (recall that $\widehat{p}_{n,t,M}(k) = \frac{1}{M_{n,t}}\sum_{m=1}^{M_{n,t}}\delta_{X_n^m(t)}(k)$)

$$\begin{aligned}\sigma_{n,t}(k) &= \mathbb{E}[\widehat{p}_{n,t,M}(k)\widehat{p}_{n,t,M}(k)] = \mathbb{E}[\frac{1}{M_{n,t}^2}\sum_{m,m'=1}^{M_{n,t}}\delta_{X_n^m(t)}(k)\delta_{X_n^{m'}(t)}(k)] \\ &= \mathbb{E}[\delta_{X_n^m(t)}(k)]\mathbb{E}[\delta_{X_n^{m'}(t)}(k)] = p_{n,t}(k)^2.\end{aligned}$$

Next, we study the asymptotical properties of the normal matrix and vector in (4.18). Since $\widehat{\varphi}_{n,t-1,M} \xrightarrow{a.s.} \mathbb{E}[\varphi_{n,t-1}]$ for each n, l, the normal matrice must also converge a.s., i.e.,

$$\mathbf{A}^e_M \xrightarrow{a.s.} \mathbf{A}^e_\infty = \frac{1}{LN}\sum_{t=1}^{L}\sum_{n=1}^{N}\mathbb{E}[\varphi_{n,t-1}]^\top\mathbb{E}[\varphi_{n,t-1}],$$

where $\mathbf{A}^e_\infty$ is defined in (4.19). Meanwhile, by Slutsky's theorem and (4.21), we have

$$\widehat{\varphi}^\top_{n,t-1,M}[\widehat{p}_{n,t,M}(k) - p_{n,t}(k)] \xrightarrow{d} \mathcal{N}(0, p_{n,t}(k)^2\mathbb{E}[\varphi_{n,t-1}]^\top\mathbb{E}[\varphi_{n,t-1}])$$

for each $n \in [N], 0 \leqslant t \leqslant L-1$. Then, since $\left(\widehat{\varphi}_{n,t-1,M}^{\top} - \mathbb{E}[\widehat{\varphi}_{n,t-1}^{\top}]\right) p_{n,t}(k) \xrightarrow{a.s.} 0$, we have

$$\widehat{\varphi}_{n,t-1,M}^{\top} \widehat{p}_{n,t,M}(k) - \mathbb{E}[\varphi_{n,t-1}^{\top}] p_{n,t}(k) \xrightarrow{d} \mathcal{N}(0, p_{n,t}(k)^2 \mathbb{E}[\varphi_{n,t-1}]^{\top} \mathbb{E}[\varphi_{n,t-1}]).$$

Consequently,

$$\mathbf{b}_M^e(\cdot,k) - \mathbf{b}_\infty^e(\cdot,k) = \frac{1}{LN} \sum_{t=1}^{L} \sum_{n=1}^{N} \widehat{\varphi}_{n,t-1,M}^{\top} \widehat{p}_{n,t,M}(k) - \mathbb{E}[\varphi_{n,t-1}^{\top}] p_{n,t}(k) \xrightarrow{d} \mathcal{N}(0, \Sigma_k^e),$$

with Σ_k^e defined in (4.20).

Then, by Lemma 4.5 and the invertibility of $\mathbf{A}_\infty^e$ in Assumption 4.7, we have $(\mathbf{A}_M^e)^{\dagger} \mathbf{b}_M^e(\cdot,k) - (\mathbf{A}_\infty^e)^{-1} \mathbf{b}_\infty^e(\cdot,k) \xrightarrow{d} \mathcal{N}(0, (\mathbf{A}_\infty^e)^{-1} \Sigma_k^e (\mathbf{A}_\infty^e)^{-1})$. This, together with Lemma 4.8, proves the asymptotic normality of the ensemble LSE $\mathbf{T}_M^e(\cdot,k)$. ■

4.4 Non-identifiability from stationary distribution

Inference from the stationary distribution is challenging since the information is limited. It is well known that a stationary distribution of a Markov chain does not determine its transition matrix, i.e., $\mathbf{P}$ is not determined by π. Similarly, the local transition matrix in our model is under-determined by the stationary distribution. Example 3.4 shows that when $(N,K) = (2,2)$, there are two $\mathbf{T}$'s leading to the same invariant measure. Theorem 4.11 shows that, in general, the ensemble-LSE is under-determined by the marginal invariant distribution, even though $\mathbf{T}$ has only $K(K-1)$ unknowns and the invariant measure π has K^N entries.

We start with a few basic facts when the process is stationary. By the shift-invariance in Proposition 3.1, the marginal distributions of all vertices are the same, and so are the expectation of the local empirical distributions, i.e.,

$$p_{n,t} = p_{1,0}, \quad \mathbb{E}[\varphi_{n,t}] = \mathbb{E}_\pi[\varphi_{1,0}], \quad \forall l \geqslant 0,\, n \in [N]. \tag{4.22}$$

Thus, the large sample limit of the loss function in (4.17) is

$$\begin{aligned} \mathcal{E}_\infty(\mathbf{T}) &= \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{t=1}^{L} \| p_{n,t,\infty}(k) - \varphi_{n,t-1,\infty} \mathbf{T}_{\cdot,k} \|^2 \\ &= \sum_{k=1}^{K} | p_{1,0}(k) - \mathbb{E}[\varphi_{1,0}] \mathbf{T}_{\cdot,k} |^2 = \| p_{1,0} - \mathbb{E}[\varphi_{1,0}] \mathbf{T} \|_2^2. \end{aligned} \tag{4.23}$$

Theorem 4.11 (Non-identifiability from the stationary distributions) *Given only the invariant measure, the local transition matrix* $\mathbf{T}$ *is under-determined, i.e., the loss function in* (4.23) *has multiple minimizers, either when* $K > 2$ *or when* $K = 2$ *with* $p_{1,0} = \mathbb{E}[\varphi_{1,0}]$.

Proof. Note that when given only the stationary measure, the local empirical distributions $\{\varphi_n\}$ can only be used via their expectations. Then, by (4.22), one can only estimate $\mathbf{T}$ from: (1) the discrepancy between the distributions $p_{1,0}$ and $\mathbb{E}[\varphi_{1,0}]\mathbf{T}$; and (2) the fact that $p_{1,0} = p_{1,0}\mathbf{T}_{true}$ by Theorem 3.2. Thus, the identifiability of $\mathbf{T}$ is equivalent to the uniqueness of the minimizer to the loss function

$$\mathcal{E}_{local}(\mathbf{T}) = \| p_{1,0} - p_{1,0}\mathbf{T} \|_2^2 + \| p_{1,0} - \mathbb{E}[\varphi_{1,0}]\mathbf{T} \|_2^2. \tag{4.24}$$

Since $\mathcal{E}_{local}(\mathbf{T})$ is quadratic in $\mathbf{T}$, it suffices to study the invertibility of its Hessian

$$\text{Hess}(\mathcal{E}_{local}) = p_{1,0}^\top p_{1,0} + \mathbb{E}[\varphi_{1,0}]^\top \mathbb{E}[\varphi_{1,0}]. \tag{4.25}$$

Here, the Hessian is with respect to $\mathbf{T}_{\cdot,k}$ and they are the same for all $k \in [K]$. The Hessian matrix has rank 2 when $p_{1,0} \neq \mathbb{E}[\varphi_{1,0}]$; and it has rank 1 when $p_{1,0} = \mathbb{E}[\varphi_{1,0}]$. Thus, there are multiple minimizers to $\mathcal{E}_{local}(\mathbf{T})$, i.e., $\mathbf{T}$ is under-determined, either when $K > 2$ or when $K = 2$ with $p_{1,0} = \mathbb{E}[\varphi_{1,0}]$. ∎

The above non-identifiability is rooted in the limited information in use: only the marginal distributions are used. Loss functions other than the quadratic loss function in (4.24), such as those based on the Kullback-Leibler divergence, total variation, or Wasserstein distances between $p_{1,0}$ and $\mathbb{E}[\varphi_{1,0}]\mathbf{T}$, will also have the same issue.

4.5 Non-asymptotic bounds for the LSEs

We establish non-asymptotic bounds for the multi-trajectory LSE $\widehat{\mathbf{T}}_M$ in (4.3) and the ensemble LSE $\widehat{\mathbf{T}}_M^e$ in (4.18). Roughly speaking, for small ϵ ad δ, both estimators are ϵ-close to the true local transition matrix with a probability of at least $1-\delta$ when the sample size is of order $M = O(\frac{K^2}{\epsilon^2 \lambda_{min}(\mathbf{A}_\infty)^2} \ln \frac{1}{\delta})$, but the constant for $\widehat{\mathbf{T}}_M^e$ is much larger due to the absence of trajectory information.

Theorem 4.12 *Let $\mathbf{T}$ be the true local transition matrix. For any $\epsilon, \delta \in (0,1)$, let $\alpha = \frac{\epsilon}{4}\lambda_{min}(\mathbf{A}_\infty)$ and $s = \frac{1}{2}\lambda_{min}(\mathbf{A}_\infty)\min\{1, \frac{\epsilon}{2\|\mathbf{T}\|_F}\}$. The following non-asymptotic bounds hold.*

(a) Under Assumption 4.1, the multi-trajectory LSE $\widehat{\mathbf{T}}_M$ in (4.3) satisfies

$$\mathbb{P}\{\|\widehat{\mathbf{T}}_M - \mathbf{T}\|_F > \epsilon\} < \delta \tag{4.26}$$

if the sample size satisfies $M > M_{\epsilon,\delta} := \max\{\frac{48K^2+8\alpha K}{3\alpha^2}\ln\frac{6K^2}{\delta}, \frac{6+2s}{3s^2}\ln\frac{6K}{\delta}\}$.

(b) Under Assumption 4.8 and assume $M_{n,t} \equiv M$, the ensemble LSE $\widehat{\mathbf{T}}_M^e$ in (4.18) satisfies

$$\mathbb{P}\{\|\widehat{\mathbf{T}}_M^e - \mathbf{T}\|_F > \epsilon\} < \delta \tag{4.27}$$

if M satisfies $M > M_{\epsilon,\delta}^e := \max\{\frac{384K^2+16\alpha K}{3\alpha^2}\ln\frac{12NLK}{\delta}, \frac{288K^2+8Ks}{3s^2}\ln\frac{6NLK}{\delta}\}$.

The proof is based on the concentration bounds for the normal matrices and vectors in the next lemma. These bounds highlight that the trajectory-based normal matrice and vector approach their large sample limits faster than those without using trajectory information.

Lemma 4.13 (Concentration for normal matrices and vectors) *For any $s > 0$, the following concentration bounds hold for the normal matrix $\mathbf{A}_M$ and vector $\mathbf{b}_M$ in for the multi-trajectory LSE in (4.3), and $\mathbf{A}_M^e$ and $\mathbf{b}_M^e$ for the ensemble LSE (4.18).*

(a) Under Assumption 4.1, we have

$$\mathbb{P}\{\|\mathbf{A}_\infty - \mathbf{A}_M\|_{op} > s\} < 2K\exp\left(-\frac{Ms^2/2}{1+s/3}\right),$$
$$\mathbb{P}\{\|\mathbf{b}_\infty - \mathbf{b}_M\|_F > s\} < 2K^2\exp\left(-\frac{Ms^2}{16K^2+8sK/3}\right).$$

(b) Under Assumption 4.8, we have

$$\mathbb{P}\left\{\|\mathbf{A}_{\infty}^{e}-\mathbf{A}_{M}^{e}\|_{op}>s\right\}<2NLK\exp\left(-\frac{Ms^2/2}{288K^2+8Ks}\right),$$
$$\mathbb{P}\left\{\|\mathbf{b}_{\infty}^{e}-\mathbf{b}_{M}^{e}\|_{F}>s\right\}<4NLK\exp\left(-\frac{Ms^2}{128K^2+16Ks/3}\right).$$

Proof. These bounds follow from applying Bernstein's inequalities.

Part (a). First, Note that $\mathbf{A}_{\infty}-\mathbf{A}_{M}=\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{A}_{L,N}^{m}-\mathbb{E}[\mathbf{A}_{L,N}^{m}]\right)$, where $\{\mathbf{A}_{L,N}^{m}\}_{m=1}^{M}$, defined in (4.7), is a sequence of symmetric identically distributed random matrices with mean zero and

$$\begin{aligned}\sigma_1^2&=\|\mathbb{E}[\left(\mathbf{A}_{L,N}^{m}-\mathbb{E}[\mathbf{A}_{L,N}^{m}]\right)^2]\|_{op}=\left\|\mathbb{E}[(\mathbf{A}_{L,N}^{m})^2]-\mathbb{E}[\mathbf{A}_{L,N}^{m}]^2\right\|_{op}^2\\&\leqslant\mathbb{E}\left\|\mathbf{A}_{L,N}^{m}\right\|_{op}^2+\left\|\mathbb{E}[\mathbf{A}_{L,N}^{m}]\right\|_{op}^2\leqslant2\mathbb{E}\|\mathbf{A}_{L,N}^{m}\|_{op}^2.\end{aligned}$$

Meanwhile, since $\mathbf{A}_{L,N}^{m}$ is symmetric,

$$\begin{aligned}\|\mathbf{A}_{L,N}^{m}\|_{op}&=\sup_{u\in\mathbb{R}^{K\times1},\|u\|=1}u^{\top}\mathbf{A}_{L,N}^{m}u=\sup_{u\in\mathbb{R}^{K},\|u\|=1}\frac{1}{LN}\sum_{t,n=1}^{L,N}u^{\top}(\varphi_{n,t-1}^{m})^{\top}\varphi_{n,t-1}^{m}u\\&=\sup_{u\in\mathbb{R}^{K},\|u\|=1}\frac{1}{LN}\sum_{t,n=1}^{L,N}|\varphi_{n,t-1}^{m}u|^2\leqslant1,\end{aligned}$$

where the inequality follows from $|\varphi_{n,t-1}^{m}u|^2\leqslant\|u\|$ since each entry of $\varphi_{n,t-1}^{m}$ is in $[0,1]$. Consequently, $\sigma_1^2\leqslant2$. Applying the matrix Bernstein's inequality (see Theorem A.4), we obtain the bound for $\mathbb{P}\left\{\|\mathbf{A}_{\infty}-\mathbf{A}_{M}\|_{op}>s\right\}$.

Next, recall that by the definition of $\mathbf{b}_M$ in (4.7), we have

$$[\mathbf{b}_{\infty}-\mathbf{b}_{M}](k,k')=\frac{1}{M}\sum_{m=1}^{M}\xi_{k,k'}^{m},\quad\xi_{k,k'}^{m}:=\left(\mathbb{E}[\mathbf{b}_{L,N}^{m}(k,k')]-\mathbf{b}_{L,N}^{m}(k,k')\right)$$

with $\mathbf{b}_{L,N}^{m}(k,k')=\frac{1}{LN}\sum_{t,n}^{L,N}c_{n,t}^{m}(k)\varphi_{n,t-1}^{m,k'}$. Using the fact that $c_{n,t}^{m}(k)\in\{0,1\}$ and $\varphi_{n,t-1}^{m,k'}\in[0,1]$, we have $|\xi_{k,k'}^{m}|\leqslant2$ and $\mathbb{E}[|\xi_{k,k'}^{m}|^2]\leqslant4$. Thus, Bernstein's inequality (see Theorem A.3) implies

$$\mathbb{P}\left\{|[\mathbf{b}_{\infty}-\mathbf{b}_{M}](k,k')|>\frac{s}{K}\right\}<2\exp\left(-\frac{Ms^2/(2K^2)}{8+4s/(3K)}\right)=2\exp\left(-\frac{Ms^2}{16K^2+8sK/3)}\right).$$

Meanwhile, note that $\|\mathbf{b}_{\infty}-\mathbf{b}_{M}\|_F^2=\sum_{k,k'=1}^{K,K}|\mathbf{b}_{\infty}-\mathbf{b}_{M}](k,k')|^2>s^2$ holds true if $|\mathbf{b}_{\infty}-\mathbf{b}_{M}](k,k')|>\frac{s}{K}$ for all k,k'. Hence,

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{b}_{\infty}-\mathbf{b}_{M}\|_F>s\right\}&\leqslant\mathbb{P}\left\{\bigcup_{k,k'}\{|[\mathbf{b}_{\infty}-\mathbf{b}_{M}](k,k')|>\frac{s}{K}\}\right\}\\&\leqslant\sum_{k,k'}\mathbb{P}\left\{|[\mathbf{b}_{\infty}-\mathbf{b}_{M}](k,k')|>\frac{s}{K}\right\}\leqslant2K^2\exp\left(-\frac{Ms^2}{16K^2+8sK/3)}\right).\end{aligned}$$

Part (b). The normal matrix $\mathbf{A}_M^e = \frac{1}{LN}\sum_{t,n=1}^{L,N} \widehat{\varphi}_{n,t-1,M}^\top \widehat{\varphi}_{n,t-1,M} \in \mathbb{R}^{K\times K}$ and vector $\mathbf{b}_M^e = \frac{1}{LN}\sum_{t,n=1}^{L,N} \widehat{\varphi}_{n,t-1,M}^\top \widehat{p}_{n,t,M}(k)$ in (4.18) require additional treatments since they involve products of the averages in samples. To remove these products, we use their upper bounds, which leads to a multiplicative factor NL in the upper bounds for the probabilities.

First, we show that

$$
\begin{aligned}
\widetilde{\varphi}_{n,t,M} &:= \widehat{\varphi}_{n,t,M} - \mathbb{E}[\widehat{\varphi}_{n,t,M}], \quad \|\widetilde{\varphi}_{n,t,M}\|_{\mathbb{R}^{1\times K}} \leqslant \min\{2, \sum_{k=1}^K |\widetilde{\varphi}_{n,t,M}(k)|\}; \\
\widetilde{p}_{n,t,M} &:= \widehat{p}_{n,t,M} - \mathbb{E}[\widehat{\varphi}_{n,t,M}], \quad \|\widetilde{p}_{n,t,M}\|_{\mathbb{R}^{1\times K}} \leqslant \min\{2, \sum_{k=1}^K |\widetilde{p}_{n,t,M}(k)|\}.
\end{aligned} \tag{4.28}
$$

Note that $\|a\|_{\mathbb{R}^{1\times K}} = \left(\sum_{k=1}^K a_k^2\right)^{1/2} \leqslant \sum_{k=1}^K |a_k|$. Thus, we only need to prove that $\|\widetilde{\varphi}_{n,t,M}\| \leqslant 2$ and $\|\widetilde{p}_{n,t,M}\| \leqslant 2$. Recall that $\widehat{\varphi}_{n,t,M} = \frac{1}{M}\sum_{m=1}^M \varphi_{n,t}^m$ and $\widehat{p}_{n,t,M}(k) = \frac{1}{M}\sum_{m=1}^M \delta_{X_n^m(t)}(k)$ in (4.16). Since $\varphi_{n,t}^m$ is a probability distribution, its entries are non-negative, so $|\varphi_{n,t}^m(k) - \mathbb{E}[\varphi_{n,t}^m](k)| \leqslant \varphi_{n,t}^m(k) + \mathbb{E}[\varphi_{n,t}^m(k)]$. As a result,

$$
\begin{aligned}
\|\widetilde{\varphi}_{n,t,M}\| \leqslant \sum_{k=1}^K |\widetilde{\varphi}_{n,t,M}(k)| &= \sum_{k=1}^K \left| \frac{1}{M}\sum_{m=1}^M [\varphi_{n,t}^m(k) - \mathbb{E}[\varphi_{n,t}^m](k) \right| \\
&\leqslant \sum_{k=1}^K \frac{1}{M}\sum_{m=1}^M \left|[\varphi_{n,t}^m(k) - \mathbb{E}[\varphi_{n,t}^m](k)\right| \leqslant 2,
\end{aligned}
$$

where the last equality uses the facts that $\left|[\varphi_{n,t}^m(k) - \mathbb{E}[\varphi_{n,t}^m](k)\right| \leqslant \varphi_{n,t}^m(k) + \mathbb{E}[\varphi_{n,t}^m](k)$ and $\sum_{k=1}^K \varphi_{n,t}^m(k) = 1$. Similarly, the bound holds for $\|\widetilde{p}_{n,t,M}\|$.

Next, we show the concentration bound for $\|\mathbf{A}_M^e - \mathbf{A}_\infty^e\|$. Note that

$$
\begin{aligned}
\mathbf{A}_M^e - \mathbf{A}_\infty^e &= \frac{1}{NL}\sum_{t,n=1}^{L,N} \widehat{\varphi}_{n,t-1,M}^\top \widehat{\varphi}_{n,t-1,M} \\
&= \frac{1}{NL}\sum_{t,n=1}^{L,N} \left(\widetilde{\varphi}_{n,t-1,M} \widetilde{\varphi}_{n,t-1,M}^\top + \widetilde{\varphi}_{n,t-1,M}^\top \mathbb{E}[\widetilde{\varphi}_{n,t-1,M}] + \mathbb{E}[\widetilde{\varphi}_{n,t-1,M}]^\top \widetilde{\varphi}_{n,t-1,M} \right).
\end{aligned}
$$

Then, using the fact that for any $u, v \in \mathbb{R}^{1\times K}$, $\|u^\top v\|_{op} = \sup_{c\in\mathbb{R}^{1\times K}, \|c\|=1} cu^\top vc^\top \leqslant \|u\|\|v\|$, we have

$$
\begin{aligned}
\|\mathbf{A}_M^e - \mathbf{A}_\infty^e\| &\leqslant \frac{1}{NL}\sum_{t,n=1}^{L,N} \|\widetilde{\varphi}_{n,t-1,M}^\top \widetilde{\varphi}_{n,t-1,M}\| + \|\widetilde{\varphi}_{n,t-1,M}^\top \mathbb{E}[\widetilde{\varphi}_{n,t-1,M}]\| + \|\mathbb{E}[\widetilde{\varphi}_{n,t-1,M}]^\top \widetilde{\varphi}_{n,t-1,M}\| \\
&\leqslant \frac{1}{NL}\sum_{t,n=1}^{L,N} (\|\widetilde{\varphi}_{n,t-1,M}\|^2 + 2\|\widetilde{\varphi}_{n,t-1,M}\| \|\mathbb{E}[\widetilde{\varphi}_{n,t-1,M}]\|) \\
&\leqslant \frac{6}{NL}\sum_{t,n=1}^{L,N} \|\widetilde{\varphi}_{n,t-1,M}\| \leqslant \frac{6}{NL}\sum_{t,n=1}^{L,N}\sum_{k=1}^K |\widetilde{\varphi}_{n,t-1,M}(k)|,
\end{aligned}
$$

where the last two inequalities follow from (4.28). Applying the Bernstein's inequality as above, we obtain

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{A}^e_M - \mathbf{A}^e_\infty\|_{op} > s\right\} &\leqslant \mathbb{P}\left\{\frac{6}{NL}\sum_{t,n=1}^{L,N}\sum_{k=1}^{K}|\widetilde{\varphi}_{n,t-1,M}(k)| > s\right\}\\ &\leqslant \sum_{t,n=1}^{L,N}\sum_{k}\mathbb{P}\left\{|\widetilde{\varphi}_{n,t-1,M}(k)| > \frac{s}{6K}\right\} \leqslant 2NLK\exp\left(-\frac{Ms^2}{288K^2+8sK)}\right).\end{aligned}$$

Lastly, we consider $\|\|\mathbf{b}^e_M - \mathbf{b}^e_\infty\|_F\|$. Using the fact that $\|u^\top v\|_F^2 = \|u\|^2\|v\|^2$, we have

$$\begin{aligned}\|\mathbf{b}^e_M - \mathbf{b}^e_\infty\| &\leqslant \frac{1}{NL}\sum_{t,n=1}^{L,N}\|\widetilde{\varphi}^\top_{n,t-1,M}\widetilde{p}_{n,t-1,M}\| + \|\widetilde{\varphi}^\top_{n,t-1,M}\mathbb{E}[\widetilde{p}_{n,t-1,M}]\| + \|\mathbb{E}[\widetilde{\varphi}_{n,t-1,M}]^\top\widetilde{p}_{n,t-1,M}\|\\ &\leqslant \frac{1}{NL}\sum_{t,n=1}^{L,N}(\|\widetilde{\varphi}_{n,t-1,M}\|\|\widetilde{p}_{n,t-1,M}\| + \|\widetilde{\varphi}_{n,t-1,M}\|\|\mathbb{E}[\widetilde{p}_{n,t-1,M}]\| + \|\mathbb{E}[\widetilde{\varphi}_{n,t-1,M}]\|\|\widetilde{p}_{n,t-1,M}\|)\\ &\leqslant \frac{2}{NL}\sum_{t,n=1}^{L,N}(\|\widetilde{\varphi}_{n,t-1,M}\| + \|\widetilde{p}_{n,t-1,M}\|) \leqslant \frac{2}{NL}\sum_{t,n=1}^{L,N}\sum_{k=1}^{K}(|\widetilde{\varphi}_{n,t-1,M}(k)| + |\widetilde{p}_{n,t-1,M}(k)|),\end{aligned}$$

where the last two inequalities follow from (4.28). Applying Bernstein's inequality, we obtain

$$\begin{aligned}\mathbb{P}\left\{\|\mathbf{b}^e_M - \mathbf{b}^e_\infty\|_{op} > s\right\} &\leqslant \mathbb{P}\left\{\frac{2}{NL}\sum_{t,n=1}^{L,N}\sum_{k=1}^{K}(|\widetilde{\varphi}_{n,t-1,M}(k)| + |\widetilde{p}_{n,t-1,M}(k)|) > s\right\}\\ &\leqslant \sum_{t,n=1}^{L,N}\sum_{k}(\mathbb{P}\left\{|\widetilde{\varphi}_{n,t-1,M}(k)| > \frac{s}{4K}\right\} + \mathbb{P}\left\{|\widetilde{p}_{n,t-1,M}(k)| > \frac{s}{4K}\right\}\\ &\leqslant 4NLK\exp\left(-\frac{Ms^2}{128K^2+\frac{16sK}{3}}\right).\end{aligned}$$

This completes the proof. ■

Proof of Theorem 4.12. The proof is based on the Bernstein concentration inequalities for the normal matrices and the normal vectors. For Part (a), note that

$$\begin{aligned}\|\widehat{\mathbf{T}}_M - \mathbf{T}\|_F = \|\mathbf{A}^\dagger_M\mathbf{b}_M - \mathbf{A}^{-1}_\infty\mathbf{b}_\infty\|_F &\leqslant \|\mathbf{A}^\dagger_M\mathbf{b}_M - \mathbf{A}^\dagger_M\mathbf{b}_\infty\|_F + \|\mathbf{A}^\dagger_M\mathbf{b}_\infty - \mathbf{A}^{-1}_\infty\mathbf{b}_\infty\|_F\\ &\leqslant \|\mathbf{A}^\dagger_M\|_{op}\|\mathbf{b}_M - \mathbf{b}_\infty\|_F + \|\mathbf{A}^\dagger_M\|_{op}\|(\mathbf{A}_\infty - \mathbf{A}_M)\|_{op}\|\mathbf{T}\|_F,\end{aligned}$$

where in the last inequality we have used the fact that $\|(\mathbf{A}^\dagger_M - \mathbf{A}^{-1}_\infty)\mathbf{b}_\infty\|_F = \|\mathbf{A}^\dagger_M(\mathbf{A}_\infty - \mathbf{A}_M)\mathbf{A}^{-1}_\infty\mathbf{b}_\infty\|_F \leqslant \|\mathbf{A}^\dagger_M\|_{op}\|(\mathbf{A}_\infty - \mathbf{A}_M)\mathbf{T}\|_F$. Hence, we have

$$\|\widehat{\mathbf{T}}_M - \mathbf{T}\| \leqslant \frac{1}{2}\epsilon + \frac{1}{2}\epsilon = \epsilon, \text{ on } E_1 \cap E_2 \cap E_3,$$

where we denote by E_1, E_2, E_3 the following events:

$$\begin{aligned}E_1 &:= \left\{\|\mathbf{A}^\dagger_M\|_{op} \leqslant C\right\}, \quad \text{with } C = 2\|\mathbf{A}^{-1}_\infty\|_{op}\\ E_2 &:= \left\{\|(\mathbf{A}_M - \mathbf{A}_\infty)\| \leqslant \frac{\epsilon}{2C\|\mathbf{T}\|_F}\right\},\\ E_3 &:= \left\{\|(\mathbf{b}_M - \mathbf{b}_\infty)\| \leqslant \frac{\epsilon}{2C}\right\}.\end{aligned}$$

Thus, $\{\|\widehat{\mathbf{T}}_M - \mathbf{T}\| > \epsilon\} \subset E_1^c \cup E_2^c \cup E_3^c$. Then, if we can prove the following bounds

$$\mathbb{P}\{E_1^c\} < \frac{\delta}{3}, \quad \mathbb{P}\{E_2^c\} < \frac{\delta}{3}, \quad \mathbb{P}\{E_3^c\} < \frac{\delta}{3} \tag{4.29}$$

for $M \geqslant M_{\epsilon,\delta}$, we can conclude (4.26) by noting that

$$\mathbb{P}\{\|\widehat{\mathbf{T}}_M - \mathbf{T}\| > \epsilon\} < \mathbb{P}\{E_1^c \cup E_2^c \cup E_3^c\} \leqslant \mathbb{P}\{E_1^c\} + \mathbb{P}\{E_2^c\} + \mathbb{P}\{E_3^c\} < \delta.$$

In the following, we prove the three bounds in (4.29) by Bernstein's inequalities. Note that

$$\|\mathbf{A}_M^\dagger\|_{op} = \lambda_{min}(\mathbf{A}_M)^{-1}, \quad \|\mathbf{A}_\infty\|_{op} = \lambda_{min}(\mathbf{A}_\infty).$$

Thus, $E_1^c = \left\{\lambda_{min}(\mathbf{A}_M)^{-1} > 2\lambda_{min}(\mathbf{A}_\infty)^{-1}\right\} = \{\lambda_{min}(\mathbf{A}_M) < \frac{1}{2}\lambda_{min}(\mathbf{A}_\infty)\} \subset \left\{|\lambda_{min}(\mathbf{A}_\infty) - \lambda_{min}(\mathbf{A}_M)| > \frac{1}{2}\lambda_{min}(\mathbf{A}_\infty)\right\}$. Meanwhile, by Weyt's inequality, $|\lambda_{min}(\mathbf{A}_\infty) - \lambda_{min}(\mathbf{A}_M)| \leqslant \|\mathbf{A}_\infty - \mathbf{A}_M\|_{op}$. Hence,

$$\mathbb{P}\{E_1^c\} < \mathbb{P}\left\{\|\mathbf{A}_\infty - \mathbf{A}_M\|_{op} > \frac{1}{2}\lambda_{min}(\mathbf{A}_\infty)\right\},$$

which can be bounded by matrix Bernstein's inequality. Similarly,

$$\mathbb{P}\{E_2^c\} = \mathbb{P}\left\{\|\mathbf{A}_\infty - \mathbf{A}_M\|_{op} > \frac{\epsilon}{2C\|\mathbf{T}\|_F} = \frac{\epsilon}{4\|\mathbf{T}\|_F}\lambda_{min}(\mathbf{A}_\infty)\right\}.$$

Thus, with $s = \frac{1}{2}\lambda_{min}(\mathbf{A}_\infty)\min\left\{1, \frac{\epsilon}{2\|\mathbf{T}\|_F}\right\}$, Lemma 4.13 implies,

$$\max\left\{\mathbb{P}\{E_1^c\}, \mathbb{P}\{E_2^c\}\right\} < \mathbb{P}\left\{\|\mathbf{A}_\infty - \mathbf{A}_M\|_{op} > s\right\} < 2K\exp\left(-\frac{Ms^2/2}{1+s/3}\right).$$

Similarly, with $\alpha = \frac{\epsilon}{2C} = \frac{\epsilon}{4}\lambda_{min}(\mathbf{A}_\infty)$, Lemma 4.13 implies,

$$\mathbb{P}\{E_3^c\} = \mathbb{P}\left\{\|\mathbf{b}_\infty - \mathbf{b}_M\|_F > \alpha\right\} < 2K^2\exp\left(-\frac{M\alpha^2}{16K^2 + 8K\alpha/3}\right).$$

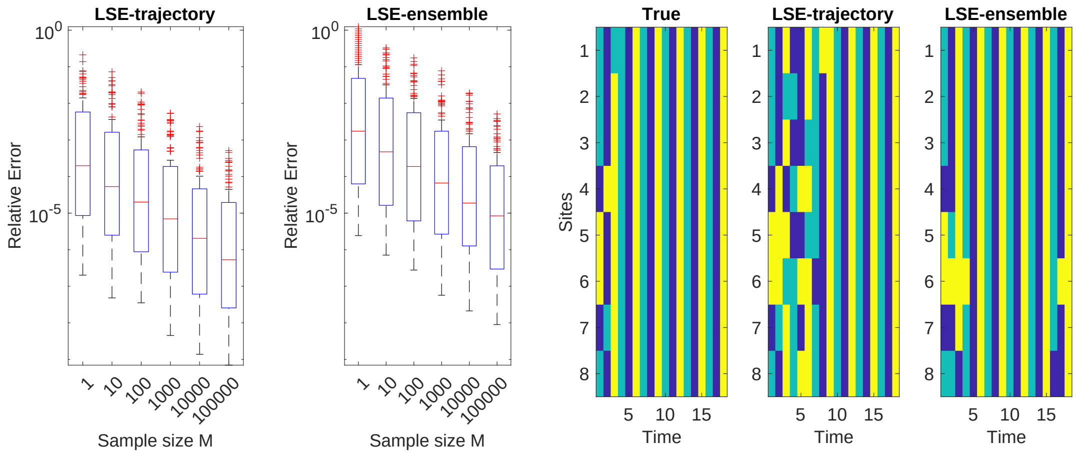
Hence, to obtain (4.29), we set M to satisfy both $2K\exp\left(-\frac{Ms^2/2}{1+s/3}\right) < \frac{\delta}{3}$ and $2K^2\exp\left(-\frac{M\alpha^2}{16K^2+8K\alpha/3}\right) < \frac{\delta}{3}$, which lead to the lower bound for M.

Part (b). The proof is the same as the above for Part (a). ■

4.6 Numerical examples

Numerical tests show that the estimators converge as sample size increases at the rate $M^{-1/2}$, agreeing with the theory. They also show that the sampling error may lead to estimators missing the periodic property of the local transition matrix and hence the synchronization; thus, additional techniques, such as an application of a threshold or a sparse condition, are needed to preserve the additional properties of the local transition matrix.

Figure 3(a) examines the convergence in sample size for the multi-trajectory LSE in Section 4.1 and the ensemble LSE in Section 4.3. That is, these estimators are obtained by first solving the normal equations by least squares with non-negative constraints and then row-normalizing the resulting solutions. Here we consider a system with $(N, K, n_v) = (8, 3, 3)$. The figure shows the box plots of the relative errors of the estimators in 100 independent simulations with increasing sample



(a) Convergence in sample size (b) Synchronization Prediction

Figure 3: **(a)**: Box plot of relative errors of the LSE estimators in 100 simulations for a system with $(N, K, n_v) = (8, 3, 3)$. The estimators converge at the same rate, but the multi-trajectory LSE is much more accurate than the ensemble LSE. **(b)**: A prediction of synchronization for Example 2.4 with $\mathbf{T}$ estimated from $M = 10^3$ trajectories with length $L = 100$. The sampling error in LSE-ensemble leads to a system without synchronizations.

size. Here we consider a randomly generated matrix $\mathbf{T} = \begin{bmatrix} 0.4719 & 0.0315 & 0.4966 \\ 0.1385 & 0.6118 & 0.2497 \\ 0.2895 & 0.4999 & 0.2107 \end{bmatrix}$, and use it to generate 5×10^5 sample trajectories with $L = 100$. Then, we randomly draw M samples out of them for 100 times.

The results show that the estimators converge as sample size increases at the rate $M^{-1/2}$, agreeing with Theorems 4.4 and 4.10. Additionally, the multi-trajectory LSE is more accurate than the ensemble LSE; since both estimators use the same dataset each time, the better accuracy comes from the additional trajectory information.

Figure 3(b) tests the effects of the sampling error in predicting the synchronization. Here the true and estimated local transition matrices are:

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \widehat{\mathbf{T}}_M = \begin{bmatrix} 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.0000 \\ 0.9962 & 0.0038 & 0.0000 \end{bmatrix}, \widehat{\mathbf{T}}_M^e = \begin{bmatrix} 0.0079 & 0.9921 & 0.0000 \\ 0.0182 & 0.0000 & 0.9818 \\ 0.9926 & 0.0074 & 0.0000 \end{bmatrix},$$

The estimators are estimated using $M = 10^2$ sample trajectories with $L = 10$. Due to the sampling error, both estimated local transition matrices are not periodic; thus, their systems will not synchronize. Figure 3(b) shows that the more accurate multi-trajectory LSE leads to numerical synchronization, while the ensemble LSE cannot maintain the synchronized motion due to the large estimation error. In practice, when the system is known to synchronize, we can apply thresholding or specification techniques to preserve the additional properties of the local transition matrix and achieve synchronization.

5 Future work

Many venues are to be explored beyond the scope of the present work.

The first venue is to study the new PCAs on general finite graphs. The graph can be more complex than the cyclic graph in this study, for example, a graph with a non-binary weight matrix for edges, or a high-dimensional lattice. The dynamical properties, such as synchronization and ergodicity, and the inference of the local transition matrix, can be studied similarly. Additionally, it is of great interest to jointly infer the local transition matrix and the weight matrix of the graph, as studied in [LWLM24] for interacting particle systems on graphs.

Another venue is to study the new PCAs on infinite graphs. Concerning the dynamical properties, one can study the ergodicity and the critical phenomena by extending the results in [Too94, LMS90, Cas23, Bér23, FH01]. Concerning the inference of the local transition matrix, one may consider the asymptotic and non-asymptotic properties of the estimator when the data is a single trajectory with $N \to \infty$, for which [DMH23] has established similar results for interacting particle systems and [BZ24] considered this problem for graphon particle systems; see, e.g., [BCW23]. An interesting parameter to estimate in a similar context would be the size n_v of each neighborhood.

A Preliminaries on Markov chain and concentration inequalities

A.1 Properties of Markov chains

Suppose $X(t)$ is a finite-state Markov chain with transition matrix $\mathbf{P}$. We recall the following general results about Markov chains.

Proposition A.1
- *There is a stationary distribution π. (because $I - P$, where P denotes the transition matrix, is not of full rank)*
- *All states are positive recurrent; see* [Dur19, Theorem 1.30].
- *Suppose $X(t)$ is aperiodic. Then $\lim_{t\to\infty} \mathbb{P}\{X(t) = \cdot\} = \pi$; see* [Dur19, Theorem 1.19].
- $\lim_{t\to\infty} \frac{1}{t}\sum_{t=1}^{t} \mathbb{P}\{X(t) = \cdot\} = \pi$*; see* [Dur19, Theorem 1.23].
- *Suppose $\int |f|\, d\pi < \infty$. Then $\lim_{t\to\infty} \frac{1}{t}\sum_{t=1}^{t} f(X(t)) = \int f\, d\pi$; see* [Dur19, Theorem 1.22].

Regarding exponential convergence to the stationary distribution, we have the following result taken from [Kul15, Theorem 1.3].

Proposition A.2 *Suppose there is some $\rho < 1$ such that*

$$\|P(x,\cdot) - P(x',\cdot)\|_{TV} \leqslant 2\rho, \quad \forall\, x, x' \in \mathcal{A}^N. \tag{A.1}$$

Then $\|P_t(x,\cdot) - P_t(x',\cdot)\|_{TV} \leqslant 2\rho^t, \quad \forall\, t \geqslant 1,\ x, x' \in \mathcal{A}^N$. In addition,

$$\|P_t(x,\cdot) - \pi\|_{TV} \leqslant 2\rho^t, \quad \forall\, t \geqslant 1,\ x, x' \in \mathcal{A}^N.$$

A.2 Concentration inequalities

Theorem A.3 (Bernstein's Inequality) (see e.g., [Ver18, Theorem 2.8.4]) *Let $X_1, \ldots, X_M$ be independent zero-mean random variables. Suppose that $|X_i| \leqslant c$ almost surely, for all i. Then for all positive t, $\mathbb{P}\left(|\sum_{i=1}^M X_i| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2/2}{\mathrm{Var}(\sum_i X_i) + \frac{1}{3}ct}\right)$.In particular, when $\{X_i\}$ are iid., we have $\mathbb{P}\left(|\frac{1}{M}\sum_{i=1}^M X_i| \geqslant t\right) \leqslant 2\exp\left(-\frac{Mt^2/2}{\mathrm{Var}(X_1) + \frac{1}{3}ct}\right)$.*

Theorem A.4 (Matrix Bernstein's inequality) ([Ver18, Theorem 5.4.1] or [Tro15, Theorem 6.1.1]) *Let $\{X_i\}_{i=1}^M \subset \mathbb{R}^{n\times n}$ be independent mean zero symmetric random matrices such that $\|X_i\|_{op} \leqslant c$ almost surely for all i. Then, for every $t \geqslant 0$, we have $\mathbb{P}(\|\sum_{i=1}^M X_i\|_{op} \geqslant t) \leqslant 2n\exp\left(-\frac{t^2/2}{\sigma^2+ct/3}\right)$, where $\sigma^2 = \|\sum_{i=1}^M \mathbb{E}[X_i^2]\|_{op}$. Additionally, when $\{X_i\}$ are identically distributed, we have*

$$\mathbb{P}(\|\frac{1}{M}\sum_{i=1}^M X_i\|_{op} \geqslant t) \leqslant 2n\exp\left(-\frac{Mt^2/2}{\|\mathbb{E}[X_1^2]\|_{op}+ct/3}\right).$$

Acknowledgement This work is supported by the following grants: NSF-1913243, FA9550-21-1-0317, and FA9550-23-1-0445.

References

[Ald13] David Aldous. Interacting particle systems as stochastic social dynamics. *Bernoulli*, 19(4):1122–1149, 2013.

[BCW23] Erhan Bayraktar, Suman Chakraborty, and Ruoyu Wu. Graphon mean field systems. *Ann. Appl. Probab.*, 33(5):3587–3619, 2023.

[Bér23] Jean Bérard. Coupling from the past for exponentially ergodic one-dimensional probabilistic cellular automata. *Electronic Journal of Probability*, 28:1–17, 2023.

[BZ24] Erhan Bayraktar and Hongyi Zhou. Non-parametric estimates for graphon mean-field particle systems, 2024.

[Cas23] Jérôme Casse. Ergodicity of some probabilistic cellular automata with binary alphabet via random walks. *Electronic Journal of Probability*, 28:1–17, 2023.

[CDP18] Patrick Cattiaux, Fanny Delebecque, and Laure Pédèches. Stochastic Cucker–Smale models: Old and new. *Ann. Appl. Probab.*, 28(5):3239–3286, 2018.

[CM01] Grace E Cho and Carl D Meyer. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335(1-3):137–150, 2001.

[CM08] Felipe Cucker and Ernesto Mordecki. Flocking in noisy environments. *Journal de Mathématiques Pures et Appliquées*, 89(3):278–296, 2008.

[Daw75] DA Dawson. Information flow in graphs. *Stochastic Processes and their applications*, 3(2):137–151, 1975.

[DMH22] Laetitia Della Maestra and Marc Hoffmann. Nonparametric estimation for interacting particle systems: McKean-Vlasov models. *Probability Theory and Related Fields*, pages 1–63, 2022.

[DMH23] Laetitia Della Maestra and Marc Hoffmann. The lan property for mckean–vlasov models in a mean-field regime. *Stochastic Processes and their Applications*, 155:109–146, 2023.

[Dur07] Richard Durrett. *Random graph dynamics*, volume 200. Cambridge University Press, 2007.

[Dur19] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge University Press, 2019.

[FH01] Hans Föllmer and Ulrich Horst. Convergence of locally and globally interacting Markov chains. *Stochastic Processes and Their Applications*, 96(1):99–121, 2001.

[Gri18] Geoffrey Grimmett. *Probability on graphs: random processes on graphs and lattices*, volume 8. Cambridge University Press, 2018.

[HVdH84] Moshe Haviv and Ludo Van der Heyden. Perturbation bounds for the stationary probabilities of a finite Markov chain. *Advances in Applied Probability*, 16(4):804–818, 1984.

[Kul15] Alexei Kulik. *Introduction to Ergodic rates for Markov chains and processes: with applications to limit theorems*, volume 2. Universitätsverlag Potsdam, 2015.

[Lig85] Thomas Milton Liggett. *Interacting particle systems*, volume 2. Springer, 1985.

[LL23] Zhongyang Li and Fei Lu. On the coercivity condition in the learning of interacting particle systems. *Stochastics and Dynamics*, page 2340003, 2023.

[LLM+21] Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, and Cheng Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135–163, 2021.

[LMS90] Joel L Lebowitz, Christian Maes, and Eugene R Speer. Statistical mechanics of probabilistic cellular automata. *Journal of statistical physics*, 59:117–170, 1990.

[LMT21] Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *Foundations of Computational Mathematics*, pages 1–55, 2021.

[LN18] Pierre-Yves Louis and Francesca R Nardi. Probabilistic cellular automata. *Emergence, Complexity, Computation*, 27, 2018.

[LRW21] Daniel Lacker, Kavita Ramanan, and Ruoyu Wu. Locally interacting diffusions as Markov random fields on path space. *Stochastic Processes and their Applications*, 140:81–114, 2021.

[LWLM24] Quanjun Lang, Xiong Wang, Fei Lu, and Mauro Maggioni. Interacting particle systems on networks: joint inference of the network and the interaction kernel. *arXiv preprint arXiv:2402.08412*, 2024.

[LZTM19] Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. Nonparametric inference of interaction laws in systems of agents from trajectory data. *Proc. Natl. Acad. Sci. USA*, 116(29):14424–14433, 2019.

[Nor98] James R Norris. *Markov chains*, volume 2. Cambridge University Press, 1998.

[Sch68] Paul J Schweitzer. Perturbation theory and finite Markov chains. *Journal of Applied Probability*, 5(2):401–413, 1968.

[Sen88] Eugene Seneta. Perturbation of the stationary distribution measured by ergodicity coefficients. *Advances in Applied Probability*, 20(1):228–230, 1988.

[Too94] Andrei Toom. On critical phenomena in interacting growth systems. part i: General. *Journal of statistical physics*, 74:91–109, 1994.

[Tro15] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[Wal20] Neil Walton. Pertubation of Markov chains. *https://appliedprobability.blog/2020/08/07/perturbation-of-markov-matrices/*, 2020.