Collaboratively Learning the Best Option on Graphs, Using Bounded Local Memory

LILI SU*, CSAIL, EECS, MIT, USA
MARTIN ZUBELDIA, LIDS, EECS, MIT, USA
NANCY LYNCH, CSAIL, EECS, MIT, USA

We consider multi-armed bandit problems in social groups wherein each individual has bounded memory and shares the common goal of learning the best arm/option. We say an individual learns the best option if eventually (as $t \to \infty$) it pulls only the arm with the highest expected reward. While this goal is provably impossible for an isolated individual due to bounded memory, we show that, in social groups, this goal can be achieved easily with the aid of social persuasion (i.e., communication) as long as the communication networks/graphs satisfy some mild conditions. In this work, we model and analyze a type of learning dynamics which are well-observed in social groups. Specifically, under the learning dynamics of interest, an individual sequentially decides on which arm to pull next based on not only its private reward feedback but also the suggestion provided by a randomly chosen neighbor. To deal with the interplay between the randomness in the rewards and in the social interaction, we employ the *mean-field approximation* method. Considering the possibility that the individuals in the networks may not be exchangeable when the communication networks are not cliques, we go beyond the classic mean-field techniques and apply a refined version of mean-field approximation:

- Using coupling we show that, if the communication graph is connected and is either regular or has doubly-stochastic degree-weighted adjacency matrix, with probability $\to 1$ as the social group size $N \to \infty$, every individual in the social group learns the best option.
- If the minimum degree of the graph diverges as $N \to \infty$, over an arbitrary but given finite time horizon, the sample paths describing the opinion evolutions of the individuals are asymptotically independent. In addition, the proportions of the population with different opinions converge to the unique solution of a system of ODEs. Interestingly, the obtained system of ODEs are invariant to the structures of the communication graphs. In the solution of the obtained ODEs, the proportion of the population holding the correct opinion converges to 1 exponentially fast in time.

Notably, our results hold even if the communication graphs are highly sparse.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; **Social recommendation**; • **Theory of computation** → **Multi-agent learning**; **Social networks**;

---

*This is the corresponding author

Authors' addresses: Lili Su, CSAIL, EECS, MIT, 32 Vassar St, Cambridge, MA, 02139, USA, lilisu3@csail.mit.edu; Martin Zubeldia, LIDS, EECS, MIT, 32 Vassar St, Cambridge, Cambridge, MA, USA, zubeldia@mit.edu; Nancy Lynch, CSAIL, EECS, MIT, 32 Vassar St, Cambridge, MA, 02139, USA, lynch@csail.mit.edu.

---

1 INTRODUCTION

Individuals often need to make a sequence of decisions among a fixed finite set of options (alternatives), whose rewards/payoffs can be regarded as stochastic, for example:

- Human society: In many economic situations, individuals need to make a sequence of decisions among multiple options, such as when purchasing perishable products [5] and when designing financial portfolios [34]. In the former case, the options can be the product of the same kind from different sellers. In the latter, the options are different possible portfolios.
- Social insect colonies: Foraging and house-hunting are two fundamental problems in social insect colonies, and both of them have inspired counterpart algorithms in robotics [29]. During foraging, each ant/bee repeatedly refines its foraging areas to improve harvesting efficiency. House-hunting refers to the collective decision process in which the entire social group collectively identifies a high-quality site to immigrate to. For the success of house-hunting, individuals repeatedly scout and evaluate multiple candidate sites, and exchange information with each other to reach a collective decision.

Many of these sequential decision problems can be cast as *multi-armed bandit problems* [1, 9, 25]. These have been studied intensively in the centralized setting, where there is only one player in the system, under different notions of performance metrics such as pseudo-regret, expected regret, simple regret, etc. [1, 9, 9, 25, 26, 31]. Specifically, a K-armed bandit problem is defined by the reward processes of individual arms/options $(R_{k,i} : i \in \mathbb{Z}_+)$ for $k = 1, \cdots, K$, where $R_{k,i}$ is the reward of the i−th pull of arm k. At each stage, a player chooses one arm to pull and obtains some observable payoff/reward generated by the chosen arm. In the most basic formulation the reward process $(R_{k,i} : i \in \mathbb{Z}_+)$ of each option is stochastic and successive pulls of arm k yield *i.i.d.* rewards $R_{k,1}, R_{k,2}, \cdots$. Both asymptotically optimal algorithms and efficient finite-time order optimal algorithms have been proposed [1, 9, 31, 33]. These algorithms typically have some nontrivial requirements on individuals' memorization capabilities. For example, upper confidence bound (UCB) algorithm requires an individual to memorize the cumulative rewards of each arm he has obtained so far, the number of pulls of each arm, and the total number of pulls [1, 31]. Although this is not a memory-demanding requirement, nevertheless, this requirement cannot be perfectly fulfilled even by humans, let alone by social insects, due to bounded rationality of humans, and limited memory and inaccurate computation of social insects. In human society, when a customer is making a purchase decision of perishable products, he may recall only the brand of product that he is satisfied with in his most recent purchase. Similarly, in ant colonies, during house-hunting, an ant can memorize only a few recently visited sites.

In this paper, we capture the above memory constraints by assuming an individual has only bounded/finite memory. The problem of multi-armed bandits with *finite memory constraint* has been proposed by Robbins [31] and attracted some research attention [13, 14, 36]. The subtleties and pitfalls in making a good definition of memory were not identified until Cover's work [13, 14]. We use the memory assumptions specified in [13], which require that an individual's history be summarized by a finite-valued memory. The detailed description of this notion of memory can be found in Section 3. We say an individual learns the best option if eventually (as $t \to \infty$) it pulls only the arm with the highest expected reward.

For an isolated individual, learning the best option is provably impossible [13].[1] Nevertheless, successful learning is still often observed in social groups such as human society [5], social insect colonies [28] and swarm robotics [29]. This may be because in social groups individuals inevitably interact with others. In particular, in social groups individuals are able to, and tend to, take advantage

---

[1]A less restricted memory constraint – stochastic fading memory – is considered in [43], wherein similar negative results when memory decays fast are obtained.

of others' experience through observing their neighbors [2, 30]. Intuitively, it appears that as a result of this social interaction, the memory of each individual is "amplified", and this *amplified shared memory* is sufficient for the entire social group to collaboratively learn the best option.

Approach and key contributions: In this paper, we rigorously show that the above intuition is correct with a focus on the impact of the graph structures on the performance of collaboratively learning. We study the learning dynamics wherein an individual updates its local opinion about which arm is more likely to give a reward based on not only its private reward feedback but also the suggestions provided by randomly chosen neighbors. Concretely, we assume time is continuous and each individual has an independent Poisson clock with common rate. The Poisson clocks model is very natural and has been widely used [20, 24, 32, 35]. When an individual's local clock ticks, it attempts to perform an update immediately via two steps:

(1) **Sampling**: **If** the individual does not have any preference over the K arms yet, **then**:
   (a) With probability $\mu \in [0, 1]$, the individual pulls one of the K arms uniformly at random (uniform sampling).
   (b) With probability $1 - \mu$, the individual chooses one neighbor uniformly at random, and pulls the arm suggested by the chosen neighbor (peer recommendation); pulls no arm if the chosen neighbor does not have any preference over the K arms yet.
   else The individual chooses one neighbor uniformly at random, and pulls the arm suggested by the chosen neighbor (peer recommendation); pulls no arm if the chosen neighbor does not have any preference over the K arms yet.
(2) **Adopting**: **If** a reward is obtained by pulling the chosen arm, **then** the individual updates its preference to this arm.

Note that if the awake individual pulls no arm, it will not get a reward; thus, its preference is unchanged. Formal description can be found in Section 3.

Remark 1.1. Although pulling no arm seems counterintuitive from the exploitation prospective, in this paper, we focus on pure exploration, i.e., our algorithm is designed to *learn* which is the best arm. Thus, we focus on restrictions relevant to the learning dynamics such as the local memory to store information, the rate at which agents can communicate among themselves, and the underlying neighboring graph.

Our learning dynamics are similar to those studied in [12] but with the following key differences: (1) We consider general communication graphs; in contrast, only cliques are considered in [12]. (2) We consider continuous-time and asynchronous dynamics, whereas in [12] all individuals are required to make update simultaneously. (3) Under our learning rules, with high probability, every individual learns the best option; whereas in [12], the proportions of the population with the wrong opinions are bounded away from zero as long as $\mu > 0$. These differences are fundamental and require completely new analysis.

A key analytical challenge of our learning dynamics is to deal with the interplay of the randomness in the rewards and that in the social interaction. Comparing to the case when the communication graphs are cliques, this interplay is significantly complicated by the lack of exchangeability among the individuals on general communication graphs. Observing this, we go beyond the classic mean-field techniques and apply a refined version of mean-field approximation:

- We show that if the communication graph is connected and is either regular or has doubly-stochastic degree-weighted adjacency matrix, with probability $\rightarrow 1$ as the social group size $N \rightarrow \infty$, every individual in the social group learns the best option with local memory of $O(K)$ states.

For any fixed N, we couple the evolutions of the number of the individuals with the correct opinion with a standard biased random walk, whose success probability is well-understood.

- If the minimum degree of the graph diverges as $N \to \infty$, over an arbitrary but given finite time horizon, the sample paths describing the opinion evolutions of the individuals are asymptotically independent. In addition, the proportions of the population with different opinions converge to the unique solution of a system of ODEs. Interestingly, the obtained system of ODEs are invariant to the structures of the communication graphs. In the solution of the obtained ODEs, the proportion of the population holding the correct opinion converges to 1 exponentially fast in time.

  The key challenge in the analysis of general graphs is that due to the lack of exchangeability, one needs to keep track of the opinion evolution of each individual. This complicates the state description of the system.

Notably, our results hold even if the communication graphs are highly sparse.

It is easy to see that the time needed for the *entire* social group to learn the best option scales poorly in N – as the entire social group cannot learn the best option until every individual wakes up at least once. Fortunately, it turns out that in many applications such as social insect colonies, it suffices to know the convergence rate until a sufficiently large fraction of the population have the correct opinion; this can be obtained by exploring the transient system behaviors.

2 RELATED WORK

- **Multi-armed bandits with bounded memory:** Multi-armed bandits with finite memory has been proposed by Robbins [31] in the special setting where the bandit has only two arms. The goal there is to maximize the long-run proportion of heads obtained which is more relaxed than to identify the best arm asymptotically. Efforts have been made to extend and improve the results in the seminal work [31]. Cover [14] constructed a time-dependent deterministic allocation rule with two arms, which is shown to be asymptotically uniformly best among the class of time-dependent finite memory rules. However, the implicit assumption is that keeping track of time is costless. The subtleties and pitfalls in making a good definition of memory was not notified until Cover's work [14]. The finite-memory constraint defined by Cover and Hellman in the work [13] is more relevant to our setting. It is shown in [13] that the optimal value of the long-run proportion over all local update functions and all allocation rules is bounded away from one, and approaches one only with local memory size goes to infinity.

- **Plurality consensus:** Another line of work that is closely to our work is plurality consensus [3, 4, 15, 18, 21]. Plurality consensus problem considers N agents each initially supporting an option in $\{1, 2, \cdots, K\}$, and the goal is to learn, in a distributed fashion, the option with the largest initial support. The key differences between our problem and the plurality consensus problem are: (1) In our problem, the objective is to learn the option that yields the highest expected reward, not the most popular in the initial configuration. (2) In our problem, we have two sources of randomness – the randomness in the rewards and the randomness in social interaction – which interplay with each other; whereas in the plurality consensus we only have randomness in the social interaction. (3) Plurality consensus relies crucially on the initial configuration of the system; whereas in our problem, the system converges to the right absorbing state even if the best arm is not the most popular one initially.

- **Mean-field approximation:** Mean-field approximation has been a powerful tool for studying the behavior of large and complex stochastic models for a long time [23]. There are many different variants of mean-field models; thus, share some non-trivial similarity

[7, 17, 38, 41, 42]. In contrast to classic mean-field approximation where the nodes are exchangeable, as is the case with cliques, to the best of our knowledge, mean-field approximation on general networks/graphs is less well-understood. Recent years have witness a flurry of research on the mean-field approximation for queueing networks with general graphs. In the context of resource pooling, Tsitsiklis and Xu [39] considered the scenario where the servers and queues are connected through a bipartite graph. In the context of load balancing, Mukherjee et al. [27] studied the join-the-shortest queue (JSQ) policy on general graphs, and showed that this policy is optimal in some sense, as long as the expected degree diverges fast enough as a function of the number of servers. Similarly, Budhiraja et al. [10] studied the local *power-of-d* scheme on graphs, and showed that as long as the minimum degree of the graph diverges, and the degree-weighted adjacency matrix of the graph is asymptotically doubly-stochastic, the occupancy process converges to the same system of ODEs as that on the complete graph.

- **Propagation of chaos:** In general, a powerful stepping stone to establishing mean-field approximation results when the network/graph is not symmetric, is the asymptotic independence of the local state of the agents in the graph (this is also called "propagation of chaos"). This has been done in the context of weakly interacting particles [6], queueing systems [8, 10, 40], and in other general systems [19, 37].

3 MODEL

In this section, we formalize the social group, the learning goal, and the notion of bounded memory, describe the learning dynamics that we are interested in, and define a continuous-time Markov chain.

Social group. A social group consists of N homogeneous individuals/agents that are connected through an undirected graph $G_N = \left([N], \mathcal{E}^N\right)$, where $[N] := \{1, \cdots, N\}$ and $\mathcal{E}^N$ is a collection of edges among $[N]$. Each agent $i \in [N]$ has a set of neighbors V_i^N. Let

$$D_{\min}^N := \min_{i \in [N]} \left| V_i^N \right| \tag{1}$$

be the minimum degree of G_N. Let $\boldsymbol{D}^N$ be the diagonal matrix of degrees, and $\boldsymbol{A}^N$ be the adjacency matrix. With a little abuse of terminology, we say graph G_N is doubly-stochastic if the degree-weighted adjacency matrix

$$(\boldsymbol{D}^N)^{-1} \boldsymbol{A}^N$$

is doubly-stochastic. Equivalently, graph G_N is doubly-stochastic if

$$\sum_{j \in V_i^N} \frac{1}{|V_j^N|} = \sum_{j \in V_i^N} \frac{1}{|V_i^N|} = 1, \ \forall \, i \in [N].$$

Learning goal. The agents in the social group want to collaboratively solve the K-armed stochastic bandit problems, wherein the reward processes of the K arms/options are Bernoulli processes with parameters $p_1, \cdots, p_K$. If arm a_k is pulled at time t, then reward $R_t \sim \text{Bern}\,(p_k)$, i.e.,

$$R_t = \begin{cases} 1, & \text{with probability } p_k; \\ 0, & \text{otherwise.} \end{cases}$$

Initially the distribution parameters $p_1, \cdots, p_K$ are unknown to any agent. We assume the arm with the highest parameter p_k is unique. Without loss of generality, let a_1 be the unique best arm

and $p_1 > p_2 \geq \cdots p_K \geq 0$. Each agent has an independent Poisson clock[2] with common rate λ, and attempts to pull an arm immediately when its local clock ticks. We say an agent learns the best option if, as $t \to \infty$, it pulls only the arm with the highest expected reward, i.e., a_1.

Remark 3.1. Although we restrict ourselves to Bernoulli rewards, this can be easily extended to more general stochastic rewards. For example, if the rewards are exponential random variables with different means, and one wishes to learn the arm with the highest mean, one could transform these exponential rewards into Bernoulli ones by considering the indicator that the reward is above a fixed threshold. In that case, p_k is the probability that the reward of the k-th arm falls above the threshold.

Bounded memory. We assume that each agent has finite/bounded memory [13]. Measuring the memory size in terms of states, as an alternative to bits, is rather standard in many natural dynamics [13, 18]. We say an agent has a memory of size m if its experience is completely summarized by an m-valued variable $M \in \{0, \cdots, m-1\}$. As a result of this, an agent sequentially decides on which arm to pull next based on only (i) its memory state and (ii) the information it gets through social interaction. The memory state may be updated with the restriction that only (a) the current memory state, (b) the current choice of arm, and (c) the recently obtained reward, are used for determining the new state. In the learning dynamics considered in this paper, it suffices to have $m = K + 1$.

Learning dynamics. In the learning dynamics under consideration, each agent keeps two variables:

- a local memory variable M that takes values in $\{0, 1, \cdots, K\}$. If $M = 0$, the agent does not have any preference over the K arms; if $M = k \in \{1, \cdots, K\}$, it means that tentatively the agent prefers arm a_k over others.
- an arm choice variable c that takes values in $\{0, 1, \cdots, K\}$ as well. If $c = 0$, the agent pulls no arm; if $c = k \in \{1, \cdots, K\}$, the agent chooses arm a_k to pull next.

Note that M is the persistent memory while c is a temporary variable. There are a variety of choices in initializing M, either deterministically or randomly. Our results account for different choices of initialization. For the temporary variable c, initialize $c = M$.

When the clock at agent i ticks at time t, we say agent i obtains the memory refinement token. With such a token, agent i refines its memory M according to Algorithm 1 via a two-step procedure. The first two **if**–**else** clauses describe how to choose an arm to pull next: If an agent does not have any preference (i.e., $M = 0$), c is determined through a combination of uniform sampling and peer recommendation; otherwise, c is completely determined by peer recommendation, see Section 1 for the notions of *uniform sampling* and *peer recommendation*. Note that during peer recommendation, if the chosen peer does not have any preference over the K arms yet, the memory of the awake agent remains unchanged. The last **if**–**else** clause says that as long as the reward obtained by pulling the arm is 1 (i.e., $R_t = 1$), then $M \leftarrow c$; otherwise, M is unchanged.

A similar gossip model was used in [18]. In our model, different from [18], the reward information arrives at the system on the fly.

System state. Due to the lack of exchangeability among agents, one needs to keep track of the opinion evolution of each agent. For a given N, let $X^N(t) \in \{0, 1\}^{N \times (K+1)}$ denote the state of the system at time t defined as

$$X^N(t) := \left(X_1^N(t), \cdots, X_N^N(t) \right), \tag{2}$$

---

[2]The Poisson clocks model is very natural and has been widely adopted for modeling [20, 24, 32, 35] natural dynamics.

Algorithm 1: Collaborative Best Option Learning (at agent i)

Input: K, μ, and V_i^N;

Local variables: $M \in \{0, 1, \cdots, K\}$ and $c \in \{0, 1, \cdots, K\}$;

Initialize $c = M$;

When local clock ticks:

if $M = 0$ **then**

   | With probability μ, set c to be one of the K arms uniformly at random;

   | With probability $1 - \mu$, $c \leftarrow$ PeerRecommendation;

else

   | $c \leftarrow$ PeerRecommendation;

if $c = 0$ **then**

   | Pull no arm;

else

   | Pull arm c;

if $R_t = 1$ **then**

   | $M \leftarrow c$;

else

   | M unchanged;

PeerRecommendation()

Choose one neighbor i' from V_i^N uniformly at random;

return M'; %% M' is the memory state of i';

---

where

$$X_i^N(t) = \left(X_{i,0}^N(t), X_{i,1}^N(t), \cdots, X_{i,K}^N(t)\right)^\top \in \{0,1\}^{K+1} \tag{3}$$

represents agent i's opinion at time t with

$$X_{i,k}^N(t) = 1, \text{ and } X_{i,k'}^N(t) = 0, \forall k' \neq k$$

if $M = k$ at agent i at time t, i.e., the memory state of agent i at time t is k. Intuitively, if $X_{i,k}^N(t) = 1$, we say agent i prefers arm a_k at time t. Here the vectors under discussion are column vectors. Note that each state x is a $N \times (K + 1)$ matrix of entries either 0 or 1, and for each row $i \in [N]$, the entries sum up to 1. Throughout the process, $X_{i,k}^N(t)$ is "fluctuating" between 0 and 1 as the local clock ticks.

Under Algorithm 1, the evolution of the state $X^N(\cdot)$ is a continuous-time Markov chain with the following transition rates. For all $i \in [N]$, and for all $k = 1, \ldots, K$, we have

$$q_{x, x+e_{ik}-e_{i0}} = \lambda p_k \left(\frac{\mu}{K} + (1 - \mu) \frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} \mathbf{1}_{\{x_{\ell,k}=1\}} \right) \mathbf{1}_{\{x_{i,0}=1\}},$$

$$q_{x, x+e_{ik}-e_{ij}} = \lambda p_k \left(\frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} \mathbf{1}_{\{x_{\ell,k}=1\}} \right) \mathbf{1}_{\{x_{i,j}=1\}}, \quad \forall j \neq k,$$

where e_{ik} is the unit matrix with a single one in position (i,k), and zeros elsewhere. Here, $q_{x,x+e_{ik}-e_{i0}}$ is the transition rate for agent i to switch from no arm preference to preferring arm a_k, and $q_{x,x+e_{ik}-e_{ij}}$ is the transition rate for agent i to switch from preferring arm a_j to preferring arm a_k. Furthermore, the Markov chain $X^N(\cdot)$ admits the following sample path construction. For all $i \in [N]$ and for all $k = 0,\ldots,K$, we have

$$X_{i,k}^N(t) = X_{i,k}^N(0) + \int_{[0,K)\times[0,t]} \left(\mathbf{1}_{\left\{X_{i,k}^N(s^-)=0\right\}} \mathbf{1}_{\left\{y \in C_{i,k}^{N,+}(s^-)\right\}} - \mathbf{1}_{\left\{X_{i,k}^N(s^-)=1\right\}} \mathbf{1}_{\left\{y \in C_{i,k}^{N,-}(s^-)\right\}} \right) \mathcal{N}_i(dy,ds),$$

where $X_{i,k}^N(0)$ is the initial condition of $X_{i,k}^N(t)$, $\mathcal{N}_i$ is a two-dimensional Poisson process of rate λ over the set $[0,K) \times [0,\infty)$, and

$$C_{i,0}^{N,+}(t) := \emptyset,$$

$$C_{i,k}^{N,+}(t) := \left[k-1, \ k-1 + \frac{\mu p_k}{K} \mathbf{1}_{\left\{X_{i,0}^N(t)=1\right\}} + \left(1 - \mu \mathbf{1}_{\left\{X_{i,0}^N(t)=1\right\}}\right) \frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} p_k \mathbf{1}_{\left\{X_{\ell,k}^N(t)=1\right\}} \right),$$

$$C_{i,0}^{N,-}(t) := \bigcup_{j=1}^{K} \left[j-1, \ j-1 + \frac{\mu p_j}{K} + (1-\mu) \frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} p_j \mathbf{1}_{\left\{X_{\ell,j}^N(t)=1\right\}} \right),$$

$$C_{i,k}^{N,-}(t) := \bigcup_{\substack{j=1 \\ j \neq k}}^{K} \left[j-1, \ j-1 + \frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} p_j \mathbf{1}_{\left\{X_{\ell,j}^N(t)=1\right\}} \right).$$

This sample path construction is adopted for technical reasons, since it significantly simplifies the statements and proofs of the results on the transient system behaviors. Nevertheless, we now provide some intuition for this construction.

Intuition of sample path construction. For $i = 1,\ldots,N$, the two-dimensional Poisson process $\mathcal{N}_i(dy,ds)$ of rate λ drives the changes in opinion of the i-th agent, where the second coordinate represents time and the first coordinate is an auxiliary parameter. In particular, note that $\mathcal{N}_i([0,K],ds)$ is a Poisson process of rate $K\lambda$ in time and, conditioned on being an event at time s, $\mathcal{N}_i(dy,\{s\})$ is a uniform probability measure over the interval $[0,K]$ independent from the measures associated to other events. We use the second coordinate to determine the time at which the opinion of the i-th agent (given by $X^N(\cdot)$) may change its state, and the first coordinate to determine how it changes (if it changes at all).

Remark 3.2. Recall that the Poisson clock of each agent has rate λ (which is K times slower than the rate of $\mathcal{N}_i([0,K],ds)$), which is compensated by thinning the faster process. This this thinning has the effect of making the probability of each transition (given that there is an event in the process $\mathcal{N}_i([0,K],ds)$) K times smaller than they would be with the Poisson clock of rate λ.

Suppose that there is an event with coordinates (y,s) in $\mathcal{N}_i$, and that agent i had opinion k right before time s (i.e., we have $X_{i,k}^N(s^-) = 1$ and $X_{i,j}^N(s^-) = 0$, for all $j \neq k$). Then, the i-th agent changes its opinion from k to $k' \neq k$ (i.e., we have $X_{i,k'}^N(s) = 1$ and $X_{i,j}^N(s^-) = 0$, for all $j \neq k'$) if and only if $y \in C_{i,k'}^{N,+}$. If $k = 0$, this jump happens with probability

$$\frac{1}{K} \left[\frac{\mu p_{k'}}{K} + (1-\mu) \frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} p_{k'} \mathbf{1}_{\left\{X_{\ell,k'}^N(t)=1\right\}} \right],$$

which is $1/K$ times the sum of two terms: one is the probability of pulling an arm at random (μ), times the probability of choosing k' ($1/K$), times the probability of success ($p_{k'}$); the other one is the probability of asking a neighbor for its opinion ($1 - \mu$), times the probability of success ($p_{k'}$). Note that the probability is exactly K times smaller than the probability of this transition when there is a tick in the local Poisson clock of an agent, which is a consequence of our underlying thinning. On the other hand, if $k \neq 0$, this jump happens with probability

$$\frac{1}{K} \left[\frac{1}{|V_i^N|} \sum_{\ell \in V_i^N} p_{k'} \mathbf{1}_{\left\{X_{\ell,k'}^N(t)=1\right\}} \right],$$

which is the fraction of neighbors with opinion k', times the probability of success $p_{k'}$, times $1/K$. Once again, this probability is K times smaller than it would be, which is a consequence of our underlying thinning.

4 MAIN RESULTS

In this section, we present the main results regarding *learnability* – the probability that the entire social group learns the best option – (subsection 4.1), and regarding the transient of opinion evolution before reaching a consensus (subsection 4.2). Furthermore, in subsection 4.3 we highlight the differences and similarities between the assumptions for the two types of results.

4.1 Learnability

For $k = 0, \ldots, K$, let

$$Z_k^N(t) := \sum_{i=1}^{N} X_{i,k}^N(t)$$

be the number of agents that prefer arm a_k at time t. We define the success event E^N as:

$$E^N \triangleq \{\text{every agent eventually learns the best option}\}$$
$$= \left\{ \lim_{t \to \infty} Z_1^N(t) = N \right\}. \tag{4}$$

Given the transition rates of the Markov chain $X^N(\cdot)$, we have that, for $k = 0, \cdots, K$, the process $Z_k^N(\cdot)$ jumps upwards with rate

$$\sum_{i=1}^{N} \mathbf{1}_{\left\{X_{i,k}^N(t)=0\right\}} \lambda p_k \left(\frac{\mu}{K} \mathbf{1}_{\left\{X_{i,0}^N(t)=1\right\}} + \left(1 - \mu \mathbf{1}_{\left\{X_{i,0}^N(t)=1\right\}}\right) \frac{1}{D_i^N} \sum_{j \in V_i^N} \mathbf{1}_{\left\{X_{j,k}^N(t)=1\right\}} \right),$$

and downwards with rate

$$\sum_{i=1}^{N} \mathbf{1}_{\left\{X_{i,k}^N(t)=1\right\}} \lambda \left(\frac{\mu}{K} \mathbf{1}_{\left\{X_{i,0}^N(t)=1\right\}} + \left(1 - \mu \mathbf{1}_{\left\{X_{i,0}^N(t)=1\right\}}\right) \frac{1}{D_i^N} \sum_{j \in V_i^N} \sum_{k':1 \leq k' \leq K, \& k' \neq k} \mathbf{1}_{\left\{X_{j,k'}^N(t)=1\right\}} p_{k'} \right).$$

It is easy to see that $Z_k^N(t) = N$ is an absorbing state of the Markov chain $X^N(\cdot)$, for all $k = 1, \ldots, K$.

For ease of exposition, we first consider the simplified scenario wherein the memory states of the individuals are initialized deterministically, and there exists a positive fraction of agents with the correct opinion – initially preferring the arm a_1 (i.e., the best arm). It turns out that the probability that eventually every agent pulls only arm a_1 increases exponentially with N. This is formalized in the following result.

Theorem 4.1. *Suppose that the graph G_N is connected, and suppose that $Z_1^N(0) \geq c_0 N$ for some $c_0 \in (0,1]$.*

- *If G_N is regular, then, for any $\mu \in [0,1]$, we have*

$$\mathbb{P}\left(E^N\right) \geq 1 - \left(\frac{p_1}{p_2}\right)^{-c_0 N}.$$

- *If G_N is doubly-stochastic, i.e., if*

$$\sum_{j \in V_i^N} \frac{1}{|V_j^N|} = 1, \qquad \forall i \in [N],$$

then, for any $\mu \in [0, 1 - p_2/p_1]$, we have

$$\mathbb{P}\left(E^N\right) \geq 1 - \left(\frac{(1 - \mu + \frac{\mu}{K})p_1}{p_2}\right)^{-c_0 N}.$$

The proof consists in coupling the process $Z_1^N(\cdot)$ with a smaller but simpler biased random walk, and using classical results on biased random walk to lower bound the probability of hitting N. The proof is deferred to Appendix A.

Remark 4.2. Intuitively, in a doubly-stochastic graph the "outflow" of agents equals the "inflow" of agents. [3] In particular, every regular graph is also doubly-stochastic.

Note that Theorem 4.1 says that the probability of "every agent eventually learns the best option" grows to 1 exponentially fast as the group size N increases. In order to maximize the lower bound in Theorem 4.1, we can simply choose $\mu = 0$ – removing the uniform sampling. However, as we will show next, it is crucial to have $\mu > 0$ when $Z_1^N(0) = 0$.

Generalizations. In Theorem 4.1 we restrict our attention to initial conditions where a positive proportion of agents prefer arm a_1. This result can be easily generalized to include initial conditions where $Z_1^N(0) = 0$ and $Z_0^N(0) \geq c_0 N$, for some $c_0 \in (0,1]$. The key to this generalization is the following lemma.

Lemma 4.3. *Suppose that $Z_0^N(0) \geq c_0 N$ for some $c_0 \in (0,1]$, and that $\mu > 0$. Then, for any $C \in (0,1)$, we have*

$$\mathbb{P}\left(Z_1^N\left(\frac{1}{\lambda}\right) \geq (1 - C)\frac{\mu c_0 p_1}{eK}N\right) \geq 1 - \exp\left(-2\left(\frac{C\mu c_0 p_1}{eK}\right)^2 N\right).$$

The intuition behind Lemma 4.3 is that when t is sufficiently small, successful memory state updates of the agents that initially have no preference over the K arms mainly rely on uniform sampling rather than peer recommendation. Thus, successful memory updates of those agents are likely to be *independent* of each other, and have some nice concentration properties. We justify this intuition in Appendix A.2.

Using Lemma 4.3, we can couple our original random walk and a standard random walk at time $\frac{1}{\lambda}$, and obtain an exponential learnability result analogous to Theorem 4.1.

Furthermore, we can generalize Theorem 4.1 and Lemma 4.3 for the case where the initial conditions $\{X_i^N(0)\}_{i \in [N]}$ are i.i.d. with respect to a fixed distribution $q = (q_0, q_1, \ldots, q_K)$.

---

[3]In fact, this assumption is one of the standard assumptions on reaching average consensus.

Lemma 4.4. *Suppose that $\{X_i^N(0)\}_{i \in [N]}$ are i.i.d. with respect to a fixed distribution $q = (q_0, q_1, \ldots, q_K)$, with $q_0 + q_1 > 0$. Then, for every $C \in (0, 1)$, we have*

$$\mathbb{P}\left(Z_0^N(0) + Z_1^N(0) \geq (1-C)(q_0 + q_1)N\right) \geq 1 - \exp\left(-2C^2(q_0 + q_1)^2 N\right).$$

The proof is a simple application of Hoeffding's inequality, and is omitted.

Remark 4.5. Although the results in this section require strict regularity conditions on all connectivity graphs (i.e., for all N), these could be loosened to asymptotic regularity conditions with appropriate concentration bounds. However, this only complicates the analysis and we believe that it does not further illustrate the regularity requirements of these results.

4.2 Transient System Behaviors

In addition to *learnability*, it is also important to characterize the transient behavior of our learning dynamics, i.e., at a given time t, what fraction of agents are preferring the best arm, the second best arm, etc. This is because in applications such as biology, chemistry, and networking, knowing the trajectories of the dynamics of interests usually provide fundamental insights on those systems. This subsection is devoted to characterizing this transient system behaviors in the case where the number of agents is very large. In this case, we will show that the dynamics are largely independent of the graph structure, as long as some mild conditions are satisfied.

First, we will establish a local approximation result. For $i \in [N]$, and for $k = 0, \ldots, K$, we define the coupled processes

$$X_{i,k}(t) = X_{i,k}^N(0) + \int_{[0,K)\times[0,t]} \left(\mathbf{1}_{\{X_{i,k}(s^-)=0\}}\mathbf{1}_{\{y \in C_{i,k}^+(s^-)\}} - \mathbf{1}_{\{X_{i,k}(s^-)=1\}}\mathbf{1}_{\{y \in C_{i,k}^-(s^-)\}}\right)\mathcal{N}_i(dy, ds). \quad (5)$$

where

$$C_{i,0}^+(t) := \emptyset,$$

$$C_{i,k}^+(t) := \left[k-1, \; k-1 + \frac{\mu p_k}{K}\mathbf{1}_{\{X_{i,0}(t)=1\}} + p_k\left(1 - \mu\mathbf{1}_{\{X_{i,0}(t)=1\}}\right)\mathbb{P}\left(X_{i,k}(t) = 1\right)\right), \qquad \forall k \geq 1,$$

$$C_{i,0}^-(t) := \bigcup_{j=1}^{K}\left[j-1, \; j-1 + \frac{\mu p_j}{K} + (1-\mu)p_j\mathbb{P}\left(X_{i,j}(t) = 1\right)\right),$$

$$C_{i,k}^-(t) := \bigcup_{\substack{j=1 \\ j \neq k}}^{K}\left[j-1, \; j-1 + p_j\mathbb{P}\left(X_{i,j}(t) = 1\right)\right), \qquad \forall k \geq 1.$$

Note that the processes $X^N(\cdot)$ and $X(\cdot)$ are coupled through the initial conditions $X(0) = X^N(0)$ and through the underlying Poisson processes $\mathcal{N}_i$ (they are all the same). However, unlike $C_{i,k}^{N,-}(t)$ and $C_{i,k}^{N,+}(t)$, the sets $C_{i,k}^-(t)$ and $C_{i,k}^+(t)$ only depend on the local state. This corresponds to a setting where an agent, instead of asking a neighbor for its state, it draws a new state (independent from the states of its neighbors), according to the distribution $\left(\mathbb{P}\left(X_{i,0}(t) = 1\right), \ldots, \mathbb{P}\left(X_{i,K}(t) = 1\right)\right)$.

We now show that, if the minimum degree of the graph G_N is large enough, and if the initial conditions are i.i.d., then the coordinate processes $\{X_i^N(\cdot)\}_{i \in [N]}$ can be approximated by the i.i.d. coordinate processes $\{X_i(\cdot)\}_{i \in [N]}$ defined above. This is formalized in the following result.

Theorem 4.6. *Fix some time $T > 0$. Suppose that the initial conditions $\{X_i^N(0)\}_{i \in [N]}$ are i.i.d. with respect to a fixed distribution $q = (q_0, \ldots, q_K)$, such that $\mathbb{P}(X_{i,k}^N(0) = 1) = q_k$, for all $k = 0, \ldots, K$,*

and for all $i \in [N]$. *Then*

$$\max_{i \in [N]} \mathbb{E} \left[\sup_{0 \le t \le T} \sum_{k=0}^{K} \left(X_{i,k}^N(t) - X_{i,k}(t) \right)^2 \right] \le \frac{16(4+\lambda)\lambda KT(K+1)}{\sqrt{D_{\min}^N}} \exp \left(48(4+\lambda)\lambda(K+1)T \right).$$

The proof is deferred to Appendix B.

Remark 4.7. Theorem 4.6 states that, for D_{min}^N large enough, the coordinate processes $\{X_i^N(\cdot)\}_{i \in [N]}$ can be approximated by the i.i.d. coordinate processes $\{X_i(\cdot)\}_{i \in [N]}$. In particular, if $D_{min}^N \to \infty$, this implies that the original coordinate processes are asymptotically independent.

Using this local approximation result and i.i.d. processes, we establish a convergence result on $Z_k^N(t)$. For any given N, and for $k = 0, \ldots, K$, let

$$Y_k^N(t) := \frac{1}{N} Z_k^N(t) = \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t)$$

be the fraction of agents that prefer arm a_k at time t. [4] We will show that over any fixed finite time horizon $[0, T]$, $Y^N(\cdot) = \left(Y_0^N(\cdot), \ldots, Y_K^N(\cdot) \right)$ converges uniformly to the unique[5] solution $y(\cdot)$ of the following ODEs.

$$\dot{y}_0(t) = -y_0(t)\lambda \frac{\mu}{K} \sum_{j=1}^{K} p_j - y_0(t)\lambda \sum_{j=1}^{K} (1-\mu)p_j y_j(t), \tag{6}$$

$$\dot{y}_k(t) = y_0(t)\lambda \frac{\mu}{K} p_k + y_k(t)\lambda \left((1-\mu)p_k y_0(t) + \sum_{j=1}^{K} (p_k - p_j)y_j(t) \right),$$
$$\forall k \ge 1. \tag{7}$$

This convergence result is formalized in the following theorem.

Theorem 4.8. *Fix some time* $T > 0$. *Suppose that the initial conditions* $\{X_i^N(0)\}_{i \in [N]}$ *are i.i.d. with respect to a fixed distribution* $q = (q_0, \ldots, q_K)$, *such that* $\mathbb{P}(X_{i,k}^N(0) = 1) = q_k$, *for all* $k = 0, \ldots, K$, *and for all* $i \in [N]$. *Furthermore, suppose that*

$$\lim_{N \to \infty} D_{\min}^N = \infty.$$

Then,

$$\lim_{N \to \infty} \mathbb{E} \left[\sup_{0 \le t \le T} \sum_{k=0}^{K} \left(Y_k^N(t) - y_k(t) \right)^2 \right] = 0,$$

where $y(t)$ *is the solution to the ODEs defined by equations* (6) *and* (7) *with initial condition* $y(0) = q$.

The proof is deferred to Appendix C.

This theorem implies that, for N large enough, the process $Y^N(t)$ closely tracks a deterministic and smooth trajectory. We will illustrate this via a simulation result at the end of this subsection. Such approximations are desired because the analysis of an ODEs system is relatively easier than that of the original stochastic system. Indeed, in a great variety of fields, such as biology, epidemic theory, physics, and chemistry [23], differential equations are used directly to model the *macroscopic*

---

[4]Recall that we say an agent prefers arm a_k at time t is its memory state is $M = k$ at time t.
[5]The uniqueness follows from the fact that the drift is a simple polynomial.

level system dynamics that are *arguably* caused by the *microscopic* level agents interactions in the system.

Remark 4.9. Note that Theorem 4.8 concerns the asymptotic aggregate behavior of the agents, which is unaffected by the behavior of a vanishingly small fraction of them. As a result, the assumptions on i.i.d. initial conditions and on diverging minimum degree need not hold for *every* node. It can be checked that as long as the assumptions hold for a fraction of agents C_N with $C_N/N \to 1$, Theorem 4.8 still holds.

Note that the above ODE system is similar to the antisymmetric Lotka-Volterra equation [22] with $\mu = 0$. The Lotka-Volterra equation is a typical replicator dynamics, where if $y_k(0) = 0$ for some k, it remains to be zero throughout the entire process. In contrast, even when $y_1(0) = 0$, the solution of our ODE system converges exponentially fast to $(0, 1, 0, \ldots, 0)$ in time as long as $\mu > 0$ and $y_0(0) > 0$.

Theorem 4.10. *Let y be the solution of the ODEs in (6) and (7). We have the following.*

- *Suppose $y_1(0) > 0$. Then, we have*

$$y_1(t) \geq 1 - \frac{1}{\frac{y_1(0)}{1 - y_1(0)} \exp(Rt) + 1}, \qquad \forall\, t \geq 0,$$

  where

$$R = \lambda \times \min \left\{ \left(1 - \mu + \frac{\mu}{K}\right) p_1, \; p_1 - p_2 \right\}.$$

- *Suppose $y_1(0) = 0$ and $y_0(0) > 0$. Then, for every $c \in (0, 1)$, we have*

$$y_1(t) \geq 1 - \frac{1}{\frac{(1-c)y_0(0)}{K - (1-c)y_0(0)} \exp(R(t - t_c)) + 1}, \qquad \forall\, t \geq \bar{t}_c,$$

  where

$$\bar{t}_c \triangleq \frac{\log \frac{1}{c}}{\lambda \frac{\mu}{K} \sum_{i=1}^{K} p_i}.$$

The proof of the second part involves showing that at time $\bar{t}_c$, we have $y_1(\bar{t}_c) \geq \frac{(1-c)y_0(0)}{K}$, and then obtaining a suitable lower bound that holds from that point going forward. The proof of the first part is a special case with $\bar{t}_c = 0$. The proof is in Appendix D.

An immediate corollary of the previous theorem is the asymptotic convergence of the trajectories to the desired state.

Corollary 4.11. *Let y be the solution of the ODEs in (6) and (7) with $y_0(0) + y_1(0) > 0$. We have that*

$$\lim_{t \to \infty} y(t) = (0, 1, 0, \ldots, 0).$$

The results in Theorems 4.8, 4.10, and Corollary 4.11 are illustrated in Figure 1, where some typical sample paths are drawn. Here, we choose $N = 200$, $K = 2$, $\lambda = 1$, $\mu = 0.2$, $p_1 = 0.8$, $p_2 = 0.4$, and initial conditions $Y^N = (1, 0, \cdots, 0)$ (i.e., every agent starts with no preference over the arms). In Figure 1, each of the component in Y^N goes to their corresponding equilibrium states exponentially fast. In particular, these typical sample paths have two slightly different behaviors stages: At the first stage, $Y_1^N(t)$ and $Y_2^N(t)$ both increase up to the point where $Y_1^N(t) + Y_2^N(t) \approx 1$ – noting that $Y_2^N(t)$ grows much slower than $Y_1^N(t)$. At the second stage, until entering their equilibrium states, $Y_1^N(t)$ is increasing and $Y_2^N(t)$ is decreasing. More importantly, Y_0^N, Y_1^N, and Y_2^N track their corresponding deterministic and smooth trajectories, which converge to the desired $(0, 1, \cdots, 0)$ in time.
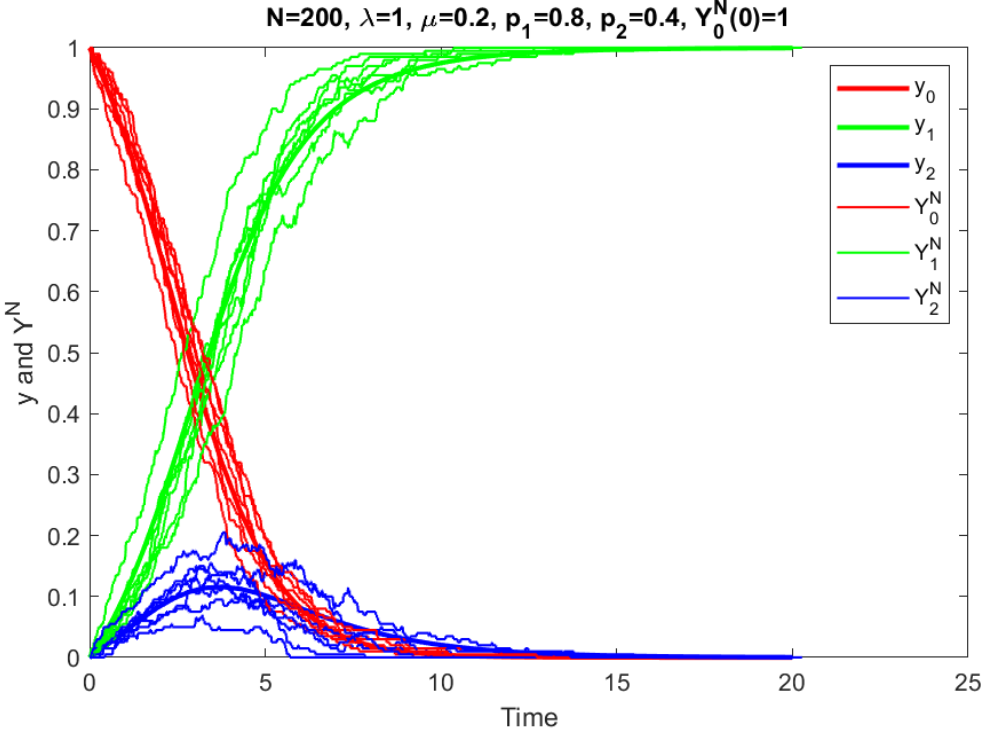
Fig. 1. Simulation with $N = 200$, $\lambda = 1$, $\mu = 0.2$, $p_1 = 0.8$, $p_2 = 0.4$, and initial condition $(1,0,0)$.

4.3 Learnability vs transient and interchange of limits

Recall that, for the learnability results in subsection 4.1, we need the graph G_N to be connected and either doubly-stochastic or regular, and we need the initial conditions be such that $Z_0^N(0) + Z_1^N(0) > cN$ (holds either deterministically or with high probability) for some positive constant $c > 0$. Under these conditions, we have

$$\lim_{N \to \infty} \lim_{t \to \infty} Y_1^N(t) = \lim_{N \to \infty} \mathbb{1}_{E_N} = 1.$$

Moreover, for the transient approximation result and the convergence to the best state presented in subsection 4.2, we need the graph G_N to have diverging minimum degree, and i.i.d. initial conditions with $q_0 + q_1 > 0$. Under these conditions, we have

$$\lim_{t \to \infty} \lim_{N \to \infty} Y_1^N(t) = \lim_{t \to \infty} y_1(t) = 1.$$

When we have all the assumptions at the same time, we can interchange the order of these limits. This is depicted in the commutative diagram of Figure 2.

On the other hand, the fact that we have different assumptions is not an artifact of our analysis, but it stems from fundamental differences in the dynamics. This is shown in the following examples.

Example 1. Let G_N be the circular graph with N agents. Suppose that the initial condition is such that there are $N/2$ agents that prefer arm a_1 and $N/2$ agents that prefer arm a_2. Then, Theorem 4.1 states that the probability that all agents eventually learn the best arm converges to 1 exponentially fast in N. However, if the initial condition is such that all agents prefer arm a_1 are all next to
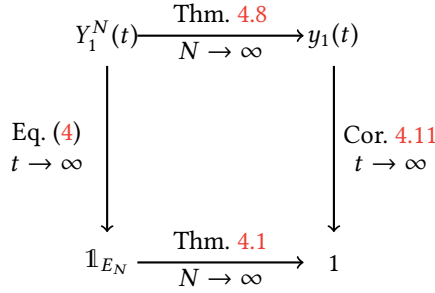
Fig. 2. Interchange of limits.

each other, then there are only 4 agents that have a neighbor with a different preferred arm than theirs (the two pair of agents in the boundary of preferred arms). It can be checked that $Z_1^N(\cdot)$ is a birth-death process with birth rate $2\lambda p_1$ and death rate $2\lambda p_2$. As a result, the approximation result given by Theorem 4.8 does not hold. Not only the transient is now linear instead of exponential, but the speed of convergence to the best arm is N times slower.

Example 2. Let G_N be the graph consisting of $N/\log\log(N)$ connected components consisting of complete graphs with $\log\log(N)$ agents. If the initial conditions are i.i.d. with $q_0 + q_1 > 0$, then Theorem 4.8 holds and we have the usual exponential convergence to the desired state. However, if $q_i > 0$ for some $i > 1$, it can be checked that all agents in at least one connected component will prefer arm a_i, with high probability. Since those agents will prefer arm a_i for all time, not all agents will eventually learn the best arm, with high probability.

Finally, the assumption on the diverging minimum degree in Theorem 4.8 is necessary in order to have local processes $\{X_i^N(\cdot)\}_{i\in[N]}$ that are asymptotically independent. Indeed, if there was a uniform upper bound on the minimum degree, there would be at least one agent with uniformly bounded degree. For those agents, the probability of interacting with any given neighbor in any given interval of time $[0,T]$ would be uniformly lower bounded.

Without this asymptotic independence, we are unable to obtain the aggregate behavior of the system, and thus we are unable to check whether the absence of the assumption on the minimum degree is necessary for the conclusions of Theorem 4.8 to hold.

5 CONCLUDING REMARKS

We studied the collaborative multi-armed bandit problems in social groups wherein each agent suffers finite memory constraint [13]. In contrast to isolated agents [13] for whom learning the best option is impossible, we showed that with the aid of social persuasion even if the communication graphs can be highly sparse, the probability of collaboratively learning the best option goes to 1 exponentially fast in N. We also characterized the transient system behaviors. In particular, we showed that if the minimum degree of the graph diverges as $N \to \infty$, sample paths describing the opinion evolutions of the individuals are asymptotically independent. Additionally, over an arbitrary but given finite time horizon, the proportions of population with different opinions converge to the unique solution of a system of ODEs that are invariant to the structures of the communication graphs. In the solution of the obtained ODEs, the proportion of the population holding the correct opinion converges to 1 exponentially fast in time. The key challenge in the analysis of general graphs is that due to the lack of exchangeability, one needs to keep track of the opinion evolution of each individual. This complicates the state description of the system.

ACKNOWLEDGEMENTS

REFERENCES

[1] Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*, 2-3 (2002), 235–256.

[2] Bandura, A. Social-learning theory of identificatory processes. *Handbook of socialization theory and research 213* (1969), 262.

[3] Becchetti, L., Clementi, A., Natale, E., Pasquale, F., and Silvestri, R. Plurality consensus in the gossip model. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms* (2015), Society for Industrial and Applied Mathematics, pp. 371–390.

[4] Becchetti, L., Clementi, A., Natale, E., Pasquale, F., Silvestri, R., and Trevisan, L. Simple dynamics for plurality consensus. *Distributed Computing 30*, 4 (2017), 293–306.

[5] Bergemann, D., and Vlimki, J. Learning and Strategic Pricing. *Econometrica 64*, 5 (1996), 1125–1149.

[6] Bhamidi, S., Budhiraja, A., and Wu, R. Weakly interacting particle systems on inhomogeneous random graphs. *Stochastic Processes and their Applications* (2018).

[7] Bramson, M. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems 30*, 1-2 (1998), 89–140.

[8] Bramson, M., Lu, Y., and Prabhakar, B. Asymptotic independence of queues under randomized load balancing. *Queueing Systems 71* (2012), 247–292.

[9] Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning 5*, 1 (2012), 1–122.

[10] Budhiraja, A., Mukherjee, D., and Wu, R. Supermarket model on graphs. *arXiv preprint arXiv:1712.07607* (2017).

[11] Budhiraja, A., Mukherjee, D., and Wu, R. Supermarket model on graphs. *arXiv:1712.07607* (2017).

[12] Celis, L. E., Krafft, P. M., and Vishnoi, N. K. A distributed learning dynamics in social groups. In *Proceedings of the ACM Symposium on Principles of Distributed Computing* (New York, NY, USA, 2017), PODC '17, ACM, pp. 441–450.

[13] Cover, T., and Hellman, M. The two-armed-bandit problem with time-invariant finite memory. *IEEE Transactions on Information Theory 16*, 2 (March 1970), 185–195.

[14] Cover, T. M. A note on the two-armed bandit problem with finite memory. *Information and Control 12*, 5 (1968), 371–377.

[15] Draief, M., and Vojnović, M. Convergence speed of binary interval consensus. *SIAM Journal on control and Optimization 50*, 3 (2012), 1087–1109.

[16] Eberle, A. Stochastic Analysis. Available onine: http://www.mi.uni-koeln.de/stochana/ws1617/Eberle_StochasticAnalysis2015.pdf, 2015.

[17] Gamarnik, D., Tsitsiklis, J. N., and Zubeldia, M. Delay, memory, and messaging tradeoffs in distributed service systems. *Stochastic Systems 1*, 7 (2017), 1–54.

[18] Ghaffari, M., and Parter, M. A polylogarithmic gossip algorithm for plurality consensus. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing* (2016), ACM, pp. 117–126.

[19] Graham, C., and Méléard, S. Chaos hypothesis for a system interacting through shared resources. *Probab. Theory Relat. Fields 100* (1994), 157–173.

[20] Hajek, B. *Random processes for engineers*. Cambridge University Press, 2015.

[21] Kempe, D., Dobra, A., and Gehrke, J. Gossip-based computation of aggregate information. In *Foundations of Computer Science, 2003. Proceedings. 44th Annual IEEE Symposium on* (2003), IEEE, pp. 482–491.

[22] Knebel, J., Weber, M. F., Krüger, T., and Frey, E. Evolutionary games of condensates in coupled birth–death processes. *Nature communications 6* (2015).

[23] Kurtz, T. G. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of applied Probability 7*, 1 (1970), 49–58.

[24] Kurtz, T. G. *Approximation of population processes*. SIAM, 1981.

[25] Lai, T. L., and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*, 1 (1985), 4–22.

[26] Mannor, S., and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research 5*, Jun (2004), 623–648.

[27] Mukherjee, D., Borst, S. C., and van Leeuwaarden, J. S. Asymptotically optimal load balancing topologies. *Proceedings of the ACM on Measurement and Analysis of Computing Systems 2*, 1 (2018), 14.

[28] Nakayama, K., Hisakado, M., and Mori, S. Nash equilibrium of social-learning agents in a restless multiarmed bandit

game. *Scientific Reports 7* (2017).

[29] PINI, G., BRUTSCHY, A., FRANCESCA, G., DORIGO, M., AND BIRATTARI, M. Multi-armed bandit formulation of the task partitioning problem in swarm robotics. In *ANTS* (2012), Springer, pp. 109–120.

[30] RENDELL, L., BOYD, R., COWNDEN, D., ENQUIST, M., ERIKSSON, K., FELDMAN, M. W., FOGARTY, L., GHIRLANDA, S., LILLICRAP, T., AND LALAND, K. N. Why copy others? insights from the social learning strategies tournament. *Science 328*, 5975 (2010), 208–213.

[31] ROBBINS, H. A sequential decision problem with a finite memory. *Proceedings of the National Academy of Sciences 42*, 12 (1956), 920–923.

[32] ROSS, S. M. *Introduction to probability models*. Academic press, 2014.

[33] SHAHRAMPOUR, S., NOSHAD, M., AND TAROKH, V. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing 65*, 16 (2017), 4281–4292.

[34] SHEN, W., WANG, J., JIANG, Y.-G., AND ZHA, H. Portfolio choices with orthogonal bandit learning. In *IJCAI* (2015), p. 974.

[35] SHWARTZ, A., AND WEISS, A. *Large deviations for performance analysis: queues, communication and computing*, vol. 5. CRC Press, 1995.

[36] SMITH, C. V., AND PYKE, R. The robbins-isbell two-armed-bandit problem with finite memory. *The Annals of Mathematical Statistics* (1965), 1375–1386.

[37] SZNITMAN, A.-S. Topics in propagation of chaos. In *Lecture Notes in Mathematics*, vol. 1464. Springer Berlin Heidelberg, 1991, pp. 165–251.

[38] TSITSIKLIS, J. N., AND XU, K. On the power of (even a little) resource pooling. *Stochastic Systems 2* (2012), 1–66.

[39] TSITSIKLIS, J. N., AND XU, K. Flexible queueing architectures. *Operations Research 65*, 5 (2017), 1398–1413.

[40] WANG, W., HARCHOL-BALTER, M., JIANG, H., SCHELLER-WOLF, A., AND SRIKANT, R. Delay Asymptotics and Bounds for Multi-Task Parallel Jobs. ArXiv preprint: arXiv:1710.00296, 2018.

[41] WORMALD, N. C. Differential equations for random processes and random graphs. *The annals of applied probability* (1995), 1217–1235.

[42] XU, J., AND HAJEK, B. The supermarket game. *Stochastic Systems 3*, 2 (2013), 405–441.

[43] XU, K., AND YUN, S.-Y. Reinforcement with fading memories. In *Sigmetrics* (2018).

A PROOF OF THEOREM 4.1

The main idea in proving Theorem 4.1 is to couple the process $Z_1^N(\cdot)$ with a standard biased random walk whose probability of hitting N (i.e., success probability) is well understood.

Despite the fact that the $N \times (K+1)$-dimensional Markov chain $X^N(\cdot)$ is hard to directly analyze, the evolution of the number of agents with the correct opinion (i.e., preferring the best arm a_1) $Z_1^N(\cdot)$ has the following nice property.

Lemma A.1. *Suppose that G_N is connected and doubly-stochastic. If there is a jump in $Z_1^N(\cdot)$ at time t, the probability of moving upwards is lower bounded as*

$$\mathbb{P}\Big(Z_1^N(t) = Z_1^N(t^-) + 1 \ \Big| \ Z_1^N(t) \neq Z_1^N(t^-)\Big) \geq \frac{p_1(1-\mu) + \frac{\mu}{K}p_1}{p_1(1-\mu) + \frac{\mu}{K}p_1 + p_2}.$$

In addition, if G_N is regular, then

$$\mathbb{P}\Big(Z_1^N(t) = Z_1^N(t^-) + 1 \ \Big| \ Z_1^N(t) \neq Z_1^N(t^-)\Big) \geq \frac{p_1}{p_1 + p_2}.$$

We prove this lemma in Appendix A.1.

Let

$$\{T_l\}_{l\geq 1} \triangleq \Big\{t \geq 0 : Z_1^N(t) \neq Z_1^N(t^-)\Big\}$$

be the set of random times when there is a jump in the process $Z_1^N(\cdot)$. Without loss of generality, we assume that $T_1 \leq T_2 \leq \cdots$. Note that the set $\{T_l\}_{l\geq 1}$ might be finite since $X^N(\cdot)$ might be absorbed after a finite number of jumps. For notational convenience, denote the random walk

$$W(l) \triangleq Z_1^N(T_l).$$

Clearly, $W(\cdot)$ is a random walk that represents the evolution of the number of agents with the correct opinion (i.e., preferring the best arm a_1). Thus, E^N can be rewritten as

$$E^N = \left\{ \lim_{l \to \infty} W(l) = N \right\}.$$

It follows from Lemma A.1 that

$$\mathbb{P}\left(W(l+1) = W(l) + 1 \mid \text{the } (l+1)\text{-st jump of } Z_1^N(\cdot) \text{ exists}\right)$$

$$\geq \begin{cases} \frac{p_1}{p_1 + p_2}, & \text{if } G_N \text{ is regular,} \\ \frac{p_1(1-\mu) + \frac{\mu}{K} p_1}{p_1(1-\mu) + \frac{\mu}{K} p_1 + p_2}, & \text{if } G_N \text{ is doubly-stochastic.} \end{cases}$$

When G_N is regular, the random walk $W(\cdot)$ is biased towards moving upwards. When G_N is only doubly-stochastic, we need to have $\mu \leq 1 - p_2/p_1$ in order to guarantee that $W(\cdot)$ is also biased towards moving upwards.

Consider the standard random walk $\left(\widehat{W}(l) : l \in \mathbb{Z}_+\right)$ such that if $\widehat{W}(l) = 0$ or $\widehat{W}(l) = N$, then $\widehat{W}(l+1) = \widehat{W}(l)$; otherwise,

$$\widehat{W}(l+1) = \begin{cases} \widehat{W}(l) + 1 & \text{with probability } p^*; \\ \widehat{W}(l) - 1 & \text{with probability } 1 - p^*, \end{cases} \tag{8}$$

where

$$p^* = \begin{cases} \frac{p_1}{p_1 + p_2}, & \text{if } G_N \text{ is regular;} \\ \frac{p_1(1-\mu) + \frac{\mu}{K} p_1}{p_1(1-\mu) + \frac{\mu}{K} p_1 + p_2}, & \text{if } G_N \text{ is doubly-stochastic.} \end{cases}$$

From [20, Chapter 4.2], we know that for any $z_0 \in \mathbb{Z}_+$,

$$\mathbb{P}\left(\lim_{l \to \infty} \widehat{W}(l) = N \mid \widehat{W}(0) = z_0\right) \geq 1 - \left(\frac{1 - p^*}{p^*}\right)^{-z_0}.$$

Intuitively, the original random walk $W(l)$ has a higher tendency to move one step up (if possible) than that of the standard random walk (8). Thus, starting at the same position, the original random walk has a higher chance to be absorbed at position N than that of the standard random walk. Thus,

$$\mathbb{P}\left(\lim_{l \to \infty} W(l) = N \mid W(0) = z_0\right) \geq \mathbb{P}\left(\lim_{l \to \infty} \widehat{W}(l) = N \mid \widehat{W}(0) = z_0\right)$$

$$\geq 1 - \left(\frac{1 - p^*}{p^*}\right)^{-z_0}.$$

We conclude Theorem 4.1 by choosing $z_0 \geq c_0 N$.

A.1 Proof of Lemma A.1

The proof of Lemma A.1 uses the following proposition.

PROPOSITION 1. *Suppose that G_N is doubly-stochastic. Then for any subset $\mathcal{S}$ of $[N]$, it holds that*

$$\sum_{i \in \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} = \sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}\}}.$$

Remark A.2. Proposition 1 says that if G_N is doubly-stochastic, for any set $\mathcal{S} \subseteq [N]$ in graph G_N, the "flow" out of set $\mathcal{S}$ equals the "flow" into set $\mathcal{S}$.

Proof of Proposition 1. When $\mathcal{S} = \emptyset$ or $\mathcal{S} = [N]$, it is easy to see that

$$\sum_{i \in \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} = 0 = \sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}\}}.$$

For the more general subset of $\mathcal{S}$ we have

$$\sum_{i \in \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} = \sum_{i=1}^{N} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} - \sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \left(1 - \mathbf{1}_{\{j \in \mathcal{S}\}} \right)$$

$$= \sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}\}} + \sum_{i=1}^{N} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} - \sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} 1.$$

To finish the proof, it remains to show that

$$\sum_{i=1}^{N} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} = \sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} 1. \tag{9}$$

The RHS of (9) can be written as

$$\sum_{i \notin \mathcal{S}} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} 1 = \sum_{i \notin \mathcal{S}} 1 \stackrel{(a)}{=} \sum_{i \notin \mathcal{S}} \sum_{j \in V_i^N} \frac{1}{|V_j^N|}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{1}_{\{i \notin \mathcal{S}\}} \mathbf{1}_{\{(i,j) \in \mathcal{E}\}} \frac{1}{|V_j^N|}. \tag{10}$$

where (a) follows from the fact that G_N is doubly-stochastic. The LHS of (9) can be written as

$$\sum_{i=1}^{N} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \notin \mathcal{S}\}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{1}_{\{(i,j) \in \mathcal{E}^N\}} \frac{1}{|V_i^N|} \mathbf{1}_{\{j \notin \mathcal{S}\}}$$

$$= \sum_{j=1}^{N} \sum_{i=1}^{N} \mathbf{1}_{\{(j,i) \in \mathcal{E}^N\}} \frac{1}{|V_j^N|} \mathbf{1}_{\{i \notin \mathcal{S}\}},$$

proving (9).

□

Now we are ready to prove Lemma A.1.

Proof of Lemma A.1. At any time t, if there is a jump in $Z_1^N(t)$, it must be true that

$$Z_1^N(t-) \neq N, \text{ and } Z_1^N(t-) + Z_0^N(t-) > 0. \tag{11}$$

Note that it is possible that $Z_0^N(t-) = N$. Let $\mathcal{S}(t-)$, $\mathcal{A}(t-)$ and $\mathcal{B}(t-)$ be a partition of set $[N]$ such that:

- Let $\mathcal{S}(t-)$ denote the set of agents that currently prefer arm a_1, i.e.,

$$\mathcal{S}(t-) := \left\{ i \in [N] : X_{i,1}^N(t-) = 1 \right\}.$$

- Let $\mathcal{A}(t-)$ denote the set of agents that currently have no preference over the K arms, i.e.,

$$\mathcal{A}(t-) := \left\{ i \in [N] : X_{i,0}^N(t-) = 1 \right\}.$$

- Let $\mathcal{B}(t-) := [N] - \mathcal{S}(t-) - \mathcal{A}(t-)$, i.e.,

$$\mathcal{B}(t-) := \left\{ i \in [N] : \sum_{k=2}^{K} X_{i,k}^N(t-) = 1 \right\}.$$

Note that $|\mathcal{S}(t-)| = Z_1^N(t-)$, $|\mathcal{A}(t-)| = Z_0^N(t-)$, and $|\mathcal{B}(t-)| = \sum_{k=2}^{K} Z_k^N(t-)$.

The upwards drift of $Z_1^N(t-)$ can be written as

$$\lambda p_1 \frac{\mu}{K} |\mathcal{A}(t-)| + (1-\mu)\lambda p_1 \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} + \lambda p_1 \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}. \quad (12)$$

Similarly, the downwards drift of $Z_1^N(t-)$ can be written as

$$\lambda \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \sum_{k=2}^{K} \mathbf{1}_{\left\{X_{j,k}^N(t-)=1\right\}} p_k \leq \lambda p_2 \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \sum_{k=2}^{K} \mathbf{1}_{\left\{X_{j,k}^N(t-)=1\right\}}$$

$$= \lambda p_2 \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}}. \quad (13)$$

We consider two cases: (1) there are no links between $\mathcal{S}(t-)$ and $\mathcal{B}(t-)$, and (2) there is a link between $\mathcal{S}(t-)$ and $\mathcal{B}(t-)$.

Case 1: Suppose that there are no links between $\mathcal{S}(t-)$ and $\mathcal{B}(t-)$. In this case, it holds that

$$\sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}} = 0. \quad (14)$$

By (13), the downwards drift of $Z_1^N(t-)$ is zero. The upwards drift of $Z_1^N(t-)$ can be written as

$$\lambda p_1 \frac{\mu}{K} |\mathcal{A}(t-)| + (1-\mu)\lambda p_1 \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}.$$

Suppose that $\mathcal{A}(t-) = \emptyset$. By (14) and the fact that G_N is connected, we know that either $|\mathcal{S}(t-)| = 0$ or $|\mathcal{S}(t-)| = N$, contradicting (11). Thus, $\mathcal{A}(t-) \neq \emptyset$. So, the upwards drift of $Z_1^N(t-)$ is nonzero. Therefore, by [20, Proposition 4.10], we have

$$\mathbb{P}\left(Z_1^N(t) = Z_1^N(t-) + 1 \mid Z_1^N(t) \neq Z_1^N(t-) \right) = 1.$$

Case 2: Suppose that there exists a link between $\mathcal{S}(t-)$ and $\mathcal{B}(t-)$. Thus,

$$\sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}} > 0. \quad (15)$$

We consider two sub-cases:

Case 2-1:

$$\sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} \geq \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}};$$

Case 2-2:

$$\sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} < \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}}.$$

Note that when G_N is a regular graph, case 2-1 always holds. To see this, let G_N be D-regular, where $D \geq 1$. We have

$$\sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} = \frac{1}{D} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{1}_{\{(i,j) \in \mathcal{E}^N\}} \mathbf{1}_{\{i \in \mathcal{B}(t-)\}} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}$$

$$= \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}}.$$

Henceforth, we consider general doubly-stochastic graphs.

Case 2-1: Suppose that

$$\sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} \geq \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}}. \tag{16}$$

The downward drift in (13) can be bounded as

$$\lambda \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \sum_{k=2}^{K} \mathbf{1}_{\left\{X_{j,k}^N(t-)=1\right\}} p_k \leq \lambda p_2 \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}.$$

Thus, the total rate of moving away from state $Z_1^N(t-)$ is upper bounded as

$$\lambda p_1 \frac{\mu}{K} |\mathcal{A}(t-)| + (1-\mu) \lambda p_1 \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} + \lambda(p_1 + p_2) \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}. \tag{17}$$

Thus, we have

$$\mathbb{P}\left(Z_1^N(t) = Z_1^N(t-) + 1 \;\middle|\; Z_1^N(t) \neq Z_1^N(t-)\right) \geq \frac{\text{Eq. (12)}}{\text{Eq. (17)}} \geq \frac{p_1}{p_1 + p_2}.$$

Case 2-2: Suppose that

$$\sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} < \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}}. \tag{18}$$

By Proposition 1, we have

$$\sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}} + \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{A}(t-)\}}$$

$$= \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} + \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}.$$

Rearrange the terms, we get

$$\sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}} - \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}$$

$$= \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} - \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{A}(t-)\}}. \tag{19}$$

Thus, we further bound the downward drift in (13) as

$$
\lambda p_2 \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}} = \lambda p_2 \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}
$$

$$
+ \lambda p_2 \left(\sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{B}(t-)\}} - \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} \right)
$$

$$
\overset{(a)}{=} \lambda p_2 \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}
$$

$$
+ \lambda p_2 \left(\sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} - \sum_{i \in \mathcal{S}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{A}(t-)\}} \right)
$$

$$
\leq \lambda p_2 \left(\sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} + \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} \right),
$$

where equality (a) follows from (19). Thus, the total rate of moving away from state is $Z_1^N(t-)$ is upper bounded as

$$
\lambda p_1 \frac{\mu}{K} |\mathcal{A}(t-)| + \lambda \left((1 - \mu) p_1 + p_2 \right) \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} + \lambda(p_1 + p_2) \sum_{i \in \mathcal{B}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}}.
$$
(20)

So we get

$$
\mathbb{P} \left(Z_1^N(t) = Z_1^N(t-) + 1 \,\Big|\, Z_1^N(t) \neq Z_1^N(t-) \right) \geq \frac{\text{Eq. (12)}}{\text{Eq. (20)}}.
$$

Note that

$$
\frac{\mu}{K} |\mathcal{A}(t-)| = \frac{\mu}{K} \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} 1 \geq \frac{\mu}{K} \sum_{i \in \mathcal{A}(t-)} \frac{1}{|V_i^N|} \sum_{j \in V_i^N} \mathbf{1}_{\{j \in \mathcal{S}(t-)\}} \geq 0.
$$
(21)

Thus,

$$
\mathbb{P} \left(Z_1^N(t) = Z_1^N(t-) + 1 \,\Big|\, Z_1^N(t) \neq Z_1^N(t-) \right) \geq \frac{\frac{\mu}{K} p_1 + (1 - \mu) p_1}{\frac{\mu}{K} p_1 + (1 - \mu) p_1 + p_2}.
$$

Putting all the cases together, we conclude Lemma A.1. □

A.2 Proof of Lemma 4.3

For a given time t_c and for each $i = 1, \cdots, N$, let

$$
L_i(t_c) = \mathbf{1}_{\left\{ X_{i,0}^N(0) = 1 \ \& \ \text{agent } i \text{ wakes up only once in } [0, t_c] \ \& \ M_i(t_c) = 1 \right\}},
$$
(22)

where $M_i(t_c)$ is the memory of agent i at time t_c. Since an agent wakes up whenever its Poisson clock ticks and the Poisson clocks are independent among agents, it holds that $L_i(t_c), \forall i = 1, \cdots, N$ are independent. In addition, by symmetry, $L_i(t_c), \forall i = 1, \cdots, N$ are identically distributed. We have

$$
\mathbb{E} \left[L_i(t_c) \right] \geq (t_c \lambda) \exp(-t_c \lambda) \frac{\mu}{K} c_0 p_1.
$$
(23)

Choosing $t_c = \frac{1}{\lambda}$, by Hoeffding's inequality, we have

$$\mathbb{P}\left(\sum_{i=1}^{N} L_i\left(\frac{1}{\lambda}\right) \leq (1-C)\frac{\mu c_0 p_1}{eK}N\right) \leq \exp\left(-2\left(\frac{C\mu c_0 p_1}{eK}\right)^2 N\right).$$

In addition, we know $Z_1^N\left(\frac{1}{\lambda}\right) \geq \sum_{i=1}^{N} L_i\left(\frac{1}{\lambda}\right)$. Thus,

$$\mathbb{P}\left(Z_1^N\left(\frac{1}{\lambda}\right) \leq (1-C)\frac{\mu c_0 p_1}{eK}N\right) \leq \exp\left(-2\left(\frac{C\mu c_0 p_1}{eK}\right)^2 N\right).$$

B PROOF OF THEOREM 4.6

The proof follows the same line of argument as in [11] and [6]. Since

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\sum_{k=0}^{K}\left(X_{i,k}^N(t) - X_{i,k}(t)\right)^2\right] \leq \sum_{k=0}^{K}\mathbb{E}\left[\sup_{0\leq t\leq T}\left(X_{i,k}^N(t) - X_{i,k}(t)\right)^2\right],$$

we just need to show that the right hand side converges to zero. For $k \geq 0$, using that $X_{i,k}^N(0) = X_{i,k}(0)$, we have

$$\sup_{0\leq t\leq T}\left(X_{i,k}^N(t) - X_{i,k}(t)\right)^2$$

$$\leq 2\sup_{0\leq t\leq T}\left(\int_{(0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{i,k}^N(s^-)=0\}}\mathbf{1}_{\{y\in C_{i,k}^{N,+}(s^-)\}} - \mathbf{1}_{\{X_{i,k}(s^-)=0\}}\mathbf{1}_{\{y\in C_{i,k}^{+}(s^-)\}}\right)\mathcal{N}_i(dy,ds)\right)^2$$

$$+ 2\sup_{0\leq t\leq T}\left(\int_{(0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{i,k}^N(s^-)=1\}}\mathbf{1}_{\{y\in C_{i,k}^{N,-}(s^-)\}} - \mathbf{1}_{\{X_{i,k}(s^-)=1\}}\mathbf{1}_{\{y\in C_{i,k}^{-}(s^-)\}}\right)\mathcal{N}_i(dy,ds)\right)^2$$

$$\leq 4\sup_{0\leq t\leq T}\left(\int_{(0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{i,k}^N(s^-)=0\}} - \mathbf{1}_{\{X_{i,k}(s^-)=0\}}\right)\mathbf{1}_{\{y\in C_{i,k}^{N,+}(s^-)\}}\mathcal{N}_i(dy,ds)\right)^2$$

$$+ 4\sup_{0\leq t\leq T}\left(\int_{(0,K)\times[0,t]}\mathbf{1}_{\{X_{i,k}(s^-)=0\}}\left(\mathbf{1}_{\{y\in C_{i,k}^{N,+}(s^-)\}} - \mathbf{1}_{\{y\in C_{i,k}^{+}(s^-)\}}\right)\mathcal{N}_i(dy,ds)\right)^2$$

$$+ 4\sup_{0\leq t\leq T}\left(\int_{(0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{i,k}^N(s^-)=1\}} - \mathbf{1}_{\{X_{i,k}(s^-)=1\}}\right)\mathbf{1}_{\{y\in C_{i,k}^{N,-}(s^-)\}}\mathcal{N}_i(dy,ds)\right)^2$$

$$+ 4\sup_{0\leq t\leq T}\left(\int_{(0,K)\times[0,t]}\mathbf{1}_{\{X_{i,k}(s^-)=1\}}\left(\mathbf{1}_{\{y\in C_{i,k}^{N,-}(s^-)\}} - \mathbf{1}_{\{y\in C_{i,k}^{-}(s^-)\}}\right)\mathcal{N}_i(dy,ds)\right)^2. \tag{24}$$

We bound the terms above individually. For the first one, we have

$$\sup_{0 \le t \le T} \left(\int\limits_{[0,K) \times [0,t]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s) \right\}} \mathcal{N}_i(dy, ds) \right)^2$$

$$= \sup_{0 \le t \le T} \left(\int\limits_{[0,K) \times [0,t]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s) \right\}} \left(\mathcal{N}_i(dy, ds) - \lambda dy ds + \lambda dy ds \right) \right)^2$$

$$\le 2 \sup_{0 \le t \le T} \left(\int\limits_{[0,K) \times [0,t]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s^-) \right\}} \left(\mathcal{N}_i(dy, ds) - \lambda dy ds \right) \right)^2$$

$$+ 2\lambda^2 \sup_{0 \le t \le T} \left(\int\limits_{[0,K) \times [0,t]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s) \right\}} dy ds \right)^2. \quad (25)$$

Note that the process

$$\int\limits_{[0,K) \times [0,t]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s^-) \right\}} \left(\mathcal{N}_i(dy, ds) - \lambda dy ds \right)$$

is a martingale (part 2 of Lemma 1.12 in [16]). Then, Doob's inequality yields

$$\mathbb{E} \left[\sup_{0 \le t \le T} \left(\int\limits_{[0,K) \times [0,t]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s^-) \right\}} \left(\mathcal{N}_i(dy, ds) - \lambda dy ds \right) \right)^2 \right]$$

$$\le 4\mathbb{E} \left[\left(\int\limits_{[0,K) \times [0,T]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s^-) \right\}} \left(\mathcal{N}_i(dy, ds) - \lambda dy ds \right) \right)^2 \right]$$

Furthermore, we have (part 3 of Lemma 1.12 in [16])

$$\mathbb{E} \left[\left(\int\limits_{[0,K) \times [0,T]} \left(\mathbf{1}_{\left\{ X_{i,k}^N(s^-)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s^-)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s^-) \right\}} \left(\mathcal{N}_i(dy, ds) - \lambda dy ds \right) \right)^2 \right]$$

$$= \lambda \mathbb{E} \left[\int\limits_{[0,K) \times [0,T]} \left(\left(\mathbf{1}_{\left\{ X_{i,k}^N(s)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s)=0 \right\}} \right) \mathbf{1}_{\left\{ y \in C_{i,k}^{N,+}(s) \right\}} \right)^2 dy ds \right]$$

$$\le \lambda \mathbb{E} \left[\int_0^T \left(\mathbf{1}_{\left\{ X_{i,k}^N(s)=0 \right\}} - \mathbf{1}_{\left\{ X_{i,k}(s)=0 \right\}} \right)^2 ds \right]$$

$$= \lambda \mathbb{E} \left[\int_0^T \left(X_{i,k}^N(s) - X_{i,k}(s) \right)^2 ds \right],$$

where the inequality comes from the fact that $\int_0^K \mathbf{1}_{\left\{y \in C_{i,k}^{N,+}(s)\right\}} dy \leq 1$ for all $s \geq 0$. By Tonelli's theorem we have

$$\mathbb{E}\left[\int_0^T \left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2 ds\right] = \int_0^T \mathbb{E}\left[\left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2\right] ds,$$

and thus the first term in Equation (25) is upper bounded by

$$8\lambda \int_0^T \mathbb{E}\left[\left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2\right] ds. \tag{26}$$

On the other hand, since $\int_0^K \mathbf{1}_{\left\{y \in C_{i,k}^{N,+}(s)\right\}} dy \leq 1$ for all $s \geq 0$, we have

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} \left(\int_{[0,K)\times[0,t]} \left(\mathbf{1}_{\left\{X_{i,k}^N(s)=0\right\}} - \mathbf{1}_{\left\{X_{i,k}(s)=0\right\}}\right) \mathbf{1}_{\left\{y \in C_{i,k}^{N,+}(s)\right\}} dy ds\right)^2\right]$$

$$\leq \mathbb{E}\left[\sup_{0 \leq t \leq T} \left(\int_{[0,K)\times[0,t]} \left|\mathbf{1}_{\left\{X_{i,k}^N(s)=0\right\}} - \mathbf{1}_{\left\{X_{i,k}(s)=0\right\}}\right| \mathbf{1}_{\left\{y \in C_{i,k}^{N,+}(s)\right\}} dy ds\right)^2\right]$$

$$\leq \mathbb{E}\left[\sup_{0 \leq t \leq T} \left(\int_0^t \left|\mathbf{1}_{\left\{X_{i,k}^N(s)=0\right\}} - \mathbf{1}_{\left\{X_{i,k}(s)=0\right\}}\right| ds\right)^2\right].$$

Furthermore, Jensen's inequality yields

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} \left(\int_0^t \left|\mathbf{1}_{\left\{X_{i,k}^N(s)=0\right\}} - \mathbf{1}_{\left\{X_{i,k}(s)=0\right\}}\right| ds\right)^2\right] \leq \mathbb{E}\left[\sup_{0 \leq t \leq T} \int_0^t \left(\mathbf{1}_{\left\{X_{i,k}^N(s)=0\right\}} - \mathbf{1}_{\left\{X_{i,k}(s)=0\right\}}\right)^2 ds\right]$$

$$\leq \mathbb{E}\left[\int_0^T \left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2 ds\right]$$

$$= \int_0^T \mathbb{E}\left[\left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2\right] ds,$$

where we used Tonelli's theorem in the last equality. Thus, the second term in Equation (25) is upper bounded by

$$2\lambda^2 \int_0^T \mathbb{E}\left[\left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2\right] ds.$$

Combining this with Equation (26), we obtain that the first term in Equation (24) is upper bounded by

$$8(4+\lambda)\lambda \int_0^T \mathbb{E}\left[\left(X_{i,k}^N(s) - X_{i,k}(s)\right)^2\right] ds$$

$$\leq 8(4+\lambda)\lambda \int_0^T \mathbb{E}\left[\sup_{0 \leq u \leq s} \left(X_{i,k}^N(u) - X_{i,k}(u)\right)^2\right] ds$$

$$\leq 8(4+\lambda)\lambda \int_0^T \max_{i \in [N]} \sum_{j=0}^K \mathbb{E}\left[\sup_{0 \leq u \leq s} \left(X_{i,j}^N(u) - X_{i,j}(u)\right)^2\right] ds. \tag{27}$$

Moreover, the same argument yields the same upper bound for the third term in Equation (24).

For the fourth term in Equation (24), an analogous argument yields the upper bound

$$8(4 + \lambda)\lambda \int_0^T \mathbb{E}\left[\int_{[0,K)} \left(\mathbf{1}_{\{y \in C_{i,k}^{N,-}(s)\}} - \mathbf{1}_{\{y \in C_{i,k}^{-}(s)\}}\right)^2 dy\right] ds. \tag{28}$$

Let us define the sets

$$\widetilde{C}_{i,0}^{-}(t) := \bigcup_{j=1}^{K} \left[j - 1, \; j - 1 + \frac{\mu p_j}{K} + (1 - \mu)\frac{1}{|V_i^N|}\sum_{\ell \in V_i^N} p_j \mathbf{1}_{\{X_{\ell,j}(t)=1\}}\right),$$

$$\widetilde{C}_{i,k}^{-}(t) := \bigcup_{\substack{j=1 \\ j \neq k}}^{K} \left[j - 1, \; j - 1 + \frac{1}{|V_i^N|}\sum_{\ell \in V_i^N} p_j \mathbf{1}_{\{X_{\ell,j}(t)=1\}}\right), \qquad \forall k \geq 1.$$

We have

$$\mathbb{E}\left[\int_{[0,K)} \left(\mathbf{1}_{\{y \in C_{i,k}^{N,-}(s)\}} - \mathbf{1}_{\{y \in C_{i,k}^{-}(s)\}}\right)^2 dy\right]$$

$$\leq \mathbb{E}\left[\int_{[0,K)} \left(\mathbf{1}_{\{y \in C_{i,k}^{N,-}(s)\}} - \mathbf{1}_{\{y \in \widetilde{C}_{i,k}^{-}(s)\}} + \mathbf{1}_{\{y \in \widetilde{C}_{i,k}^{-}(s)\}} - \mathbf{1}_{\{y \in C_{i,k}^{-}(s)\}}\right)^2 dy\right]$$

$$\leq 2\mathbb{E}\left[\int_{[0,K)} \left(\mathbf{1}_{\{y \in C_{i,k}^{N,-}(s)\}} - \mathbf{1}_{\{y \in \widetilde{C}_{i,k}^{-}(s)\}}\right)^2 dy + \int_{[0,K)} \left(\mathbf{1}_{\{y \in \widetilde{C}_{i,k}^{-}(s)\}} - \mathbf{1}_{\{y \in C_{i,k}^{-}(s)\}}\right)^2 dy\right]. \tag{29}$$

For the first term, we have

$$\mathbb{E}\left[\int_{[0,K)} \left(\mathbf{1}_{\{y \in C_{i,k}^{N,-}(s)\}} - \mathbf{1}_{\{y \in \widetilde{C}_{i,k}^{-}(s)\}}\right)^2 dy\right] \leq \mathbb{E}\left[\sum_{\substack{j=1 \\ j \neq k}}^{K} \frac{1}{|V_i^N|}\left|\sum_{\ell=1}^{|V_i^N|} p_j\left(\mathbf{1}_{\{X_{i_\ell,j}^N(s)=1\}} - \mathbf{1}_{\{X_{i_\ell,j}(s)=1\}}\right)\right|\right]$$

$$\leq \mathbb{E}\left[\sum_{\substack{j=1 \\ j \neq k}}^{K} \frac{1}{|V_i^N|}\sum_{\ell=1}^{|V_i^N|}\left|\mathbf{1}_{\{X_{i_\ell,j}^N(s)=1\}} - \mathbf{1}_{\{X_{i_\ell,j}(s)=1\}}\right|\right]$$

$$= \mathbb{E}\left[\sum_{\substack{j=1 \\ j \neq k}}^{K} \frac{1}{|V_i^N|}\sum_{\ell=1}^{|V_i^N|}\left(X_{i_\ell,j}^N(s) - X_{i_\ell,j}(s)\right)^2\right]$$

$$\leq \max_{i \in [N]} \mathbb{E}\left[\sum_{j=0}^{K}\left(X_{i,j}^N(s) - X_{i,j}(s)\right)^2\right]. \tag{30}$$

For the second term, we have

$$
\mathbb{E}\left[\int_{[0,K)}\left(\mathbf{1}_{\left\{y\in\widetilde{C}_{i,k}^{-}(s)\right\}}-\mathbf{1}_{\left\{y\in C_{i,k}^{-}(s)\right\}}\right)^{2}dy\right]=\mathbb{E}\left[\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\left|\sum_{\ell=1}^{|V_{i}^{N}|}p_{j}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)\right|\right]
$$

$$
=\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\mathbb{E}\left[\left|\sum_{\ell=1}^{|V_{i}^{N}|}p_{j}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)\right|\right]
$$

$$
\leq\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\mathbb{E}\left[\left|\sum_{\ell=1}^{|V_{i}^{N}|}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)\right|\right].
$$

By Jensen's inequality, we have

$$
\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\mathbb{E}\left[\left|\sum_{\ell=1}^{|V_{i}^{N}|}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)\right|\right]
$$

$$
\leq\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\left(\mathbb{E}\left[\left(\sum_{\ell=1}^{|V_{i}^{N}|}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)\right)^{2}\right]\right)^{\frac{1}{2}}.
$$

Since the initial conditions $\{X_i(0)\}_{i\in[N]}$ are i.i.d., $\{X_i(s)\}_{i\in[N]}$ are also i.i.d., for all $s\geq 0$. Thus

$$
\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\left(\mathbb{E}\left[\left(\sum_{\ell=1}^{|V_{i}^{N}|}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)\right)^{2}\right]\right)^{\frac{1}{2}}
$$

$$
=\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\left(\mathbb{E}\left[\sum_{\ell=1}^{|V_{i}^{N}|}\left(\mathbf{1}_{\left\{X_{i_{\ell},j}(s)=1\right\}}-\mathbb{P}\left(X_{i,j}(s)=1\right)\right)^{2}\right]\right)^{\frac{1}{2}}
$$

$$
=\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{|V_{i}^{N}|}\left(|V_{i}^{N}|\mathbb{E}\left[\left(\mathbf{1}_{\left\{X_{1,j}(s)=1\right\}}-\mathbb{P}\left(X_{1,j}(s)=1\right)\right)^{2}\right]\right)^{\frac{1}{2}}
$$

$$
=\sum_{\substack{j=1\\j\neq k}}^{K}\frac{1}{\sqrt{|V_{i}^{N}|}}\left(\mathbb{E}\left[\left(\mathbf{1}_{\left\{X_{1,j}(s)=1\right\}}-\mathbb{P}\left(X_{1,j}(s)=1\right)\right)^{2}\right]\right)^{\frac{1}{2}}
$$

$$
\leq\frac{K-1}{\sqrt{|V_{i}^{N}|}}
$$

Combining this with equations (28)–(30) we obtain that the fourth term in Equation (24) is upper bounded by

$$16(4 + \lambda)\lambda \int_0^T \left(\max_{i \in [N]} \sum_{j=0}^K \mathbb{E}\left[\left(X_{i,j}^N(s) - X_{i,j}(s) \right)^2 \right] + \frac{K-1}{\sqrt{|V_i^N|}} \right) ds$$

$$\leq 16(4 + \lambda)\lambda \int_0^T \left(\max_{i \in [N]} \sum_{j=0}^K \mathbb{E}\left[\sup_{0 \leq u \leq s} \left(X_{i,j}^N(u) - X_{i,j}(u) \right)^2 \right] + \frac{K-1}{\sqrt{|V_i^N|}} \right) ds. \quad (31)$$

Moreover, an analogous argument yields the upper bound

$$16(4 + \lambda)\lambda \int_0^T \left(\max_{i \in [N]} \sum_{j=0}^K \mathbb{E}\left[\sup_{0 \leq u \leq s} \left(X_{i,j}^N(u) - X_{i,j}(u) \right)^2 \right] + \frac{1}{\sqrt{|V_i^N|}} \right) ds \quad (32)$$

for the second term in Equation (24).

Finally, combining equations (27), (31), and (32), and substituting in Equation (24), we have that

$$\mathbb{E}\left[\sup_{0 \leq t \leq T} \left| X_{i,k}^N(t) - X_{i,k}(t) \right|^2 \right]$$

$$\leq 48(4 + \lambda)\lambda \int_0^T \left(\max_{i \in [N]} \sum_{j=0}^K \mathbb{E}\left[\sup_{0 \leq u \leq s} \left(X_{i,j}^N(u) - X_{i,j}(u) \right)^2 \right] \right) ds + \frac{16(4 + \lambda)\lambda KT}{\sqrt{|V_i^N|}},$$

for all $k \geq 0$. It follows that

$$\max_{i \in [N]} \sum_{k=0}^K \mathbb{E}\left[\sup_{0 \leq t \leq T} \left| X_{i,k}^N(t) - X_{i,k}(t) \right|^2 \right]$$

$$\leq 48(4 + \lambda)\lambda(K + 1) \int_0^T \left(\max_{i \in [N]} \sum_{k=0}^K \mathbb{E}\left[\sup_{0 \leq u \leq s} \left(X_{i,k}^N(u) - X_{i,k}(u) \right)^2 \right] \right) ds$$

$$+ \frac{16(4 + \lambda)\lambda KT(K + 1)}{\sqrt{D_{\min}^N}}.$$

Applying Gronwall's inequality, we obtain

$$\max_{i \in [N]} \sum_{k=0}^K \mathbb{E}\left[\sup_{0 \leq t \leq T} \left| X_{i,k}^N(t) - X_{i,k}(t) \right|^2 \right] \leq \frac{16(4 + \lambda)\lambda KT(K + 1)}{\sqrt{D_{\min}^N}} \exp\left(48(4 + \lambda)\lambda(K + 1)T \right),$$

which concludes the proof.

C   PROOF OF THEOREM 4.8

Let us define the function $\eta : [0, T] \rightarrow [0, 1]^{K+1}$, such that

$$\eta_k(t) := \mathbb{E}\left[X_{1,k}(t) \right], \quad (33)$$

for all $t \in [0, T]$, and for all $k = 0, \ldots, K$, where $X_1(\cdot)$ is as in (5), with initial distribution y. Recall that

$$Y_k^N(t) = \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t),$$

for all $k = 0, \ldots, N$. Since

$$\left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - \eta_k(t) \right| \leq 1,$$

we have

$$\mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left(\frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - \eta_k(t) \right)^2 \right] \leq \mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - \eta_k(t) \right| \right].$$

Furthermore,

$$\mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - \eta_k(t) \right| \right]$$

$$= \mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - X_{i,k}(t) + X_{i,k}(t) - \eta_k(t) \right| \right]$$

$$\leq \mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - X_{i,k}(t) \right| + \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - \eta_k(t) \right| \right]$$

$$\leq \mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}^N(t) - X_{i,k}(t) \right| \right] + \mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - \eta_k(t) \right| \right]. \quad (34)$$

We first show that the second term converges to zero. Since

$$\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - y_k(t) \right| \leq 2,$$

the Dominated Convergence Theorem implies that

$$\lim_{N \to \infty} \mathbb{E} \left[\sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - \eta_k(t) \right| \right] = \mathbb{E} \left[\lim_{N \to \infty} \sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - \eta_k(t) \right| \right].$$

Furthermore, since the initial conditions of the processes $X_i(\cdot)$ are i.i.d., then the whole processes are also i.i.d. by construction. Combining this with the fact that

$$\mathbb{E} \left[X_{i,k}(t) \right] = \eta_k(t),$$

for all $k = 0, \ldots, K$, $i \in [N]$, and $t \in [0, T]$, and using standard arguments like in [17, 38], it can be shown that

$$\lim_{N \to \infty} \sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - \eta_k(t) \right| = 0, \qquad a.s.,$$

and thus

$$\mathbb{E} \left[\lim_{N \to \infty} \sup_{0 \leq t \leq T} \sum_{k=0}^{K} \left| \frac{1}{N} \sum_{i=1}^{N} X_{i,k}(t) - \eta_k(t) \right| \right] = 0.$$

For the first term in Equation (34), we have

$$\mathbb{E}\left[\sup_{0\le t\le T}\sum_{k=0}^{K}\left|\frac{1}{N}\sum_{i=1}^{N}X_{i,k}^{N}(t)-X_{i,k}(t)\right|\right]\le\mathbb{E}\left[\sup_{0\le t\le T}\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K}\left|X_{i,k}^{N}(t)-X_{i,k}(t)\right|\right]$$

$$\le\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\sup_{0\le t\le T}\sum_{k=0}^{K}\left|X_{i,k}^{N}(t)-X_{i,k}(t)\right|\right]$$

$$\le\max_{i\in[N]}\mathbb{E}\left[\sup_{0\le t\le T}\sum_{k=0}^{K}\left|X_{i,k}^{N}(t)-X_{i,k}(t)\right|\right].$$

Since

$$\left|X_{i,k}^{N}(t)-X_{i,k}(t)\right|=\left(X_{i,k}^{N}(t)-X_{i,k}(t)\right)^{2},$$

and

$$\lim_{N\to\infty}D_{\min}^{N}=\infty,$$

Theorem 4.6 implies that

$$\lim_{N\to\infty}\max_{i\in[N]}\mathbb{E}\left[\sup_{0\le t\le T}\sum_{k=0}^{K}\left|X_{i,k}^{N}(t)-X_{i,k}(t)\right|\right]=0,$$

and thus the first term in Equation (34) also converges to 0.

The rest of the proof is devoted to showing that η is the solution of the ODE defined by equations (6) and (7), with initial condition q. Since $\eta_k(0)=\mathbb{E}\left[X_{1,k}^{N}(0)\right]=q_k$, we have that

$$\eta_k(t)=\mathbb{E}\left[X_{1,k}^{N}(0)\right]+\mathbb{E}\left[\int_{[0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{1,k}(s^-)=0\}}\mathbf{1}_{\{y\in C_{1,k}^+(s^-)\}}-\mathbf{1}_{\{X_{1,k}(s^-)=1\}}\mathbf{1}_{\{y\in C_{1,k}^-(s^-)\}}\right)\mathcal{N}_1(dy,ds)\right]$$

$$=q_k+\mathbb{E}\left[\int_{[0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{1,k}(s)=0\}}\mathbf{1}_{\{y\in C_{1,k}^+(s)\}}-\mathbf{1}_{\{X_{1,k}(s)=1\}}\mathbf{1}_{\{y\in C_{1,k}^-(s)\}}\right)\lambda\,dyds\right]$$

$$+\mathbb{E}\left[\int_{[0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{1,k}(s^-)=0\}}\mathbf{1}_{\{y\in C_{1,k}^+(s^-)\}}-\mathbf{1}_{\{X_{1,k}(s^-)=1\}}\mathbf{1}_{\{y\in C_{1,k}^-(s^-)\}}\right)\left(\mathcal{N}_1(dy,ds)-\lambda\,dyds\right)\right].$$

Since the term inside the second expectation is a martingale that starts at 0 (part 2 of Lemma 1.12 in [?]), Doob's optional stopping theorem implies that its expectation is equal to 0 as well. As a result, we have that

$$\eta_k(t)=q_k+\mathbb{E}\left[\int_{[0,K)\times[0,t]}\left(\mathbf{1}_{\{X_{1,k}(s)=0\}}\mathbf{1}_{\{y\in C_{1,k}^+(s)\}}-\mathbf{1}_{\{X_{1,k}(s)=1\}}\mathbf{1}_{\{y\in C_{1,k}^-(s)\}}\right)\lambda\,dyds\right]$$

$$=q_k+\lambda\int_{[0,K)\times[0,t]}\left(\mathbb{E}\left[\mathbf{1}_{\{X_{1,k}(s)=0\}}\mathbf{1}_{\{y\in C_{1,k}^+(s)\}}\right]-\mathbb{E}\left[\mathbf{1}_{\{X_{1,k}(s)=1\}}\mathbf{1}_{\{y\in C_{1,k}^-(s)\}}\right]\right)dyds,$$

where in the last equality we used Fubini's theorem. For $k = 0$, we have

$$\eta_0(t) = q_0 - \lambda \int\limits_{[0,K)\times[0,t]} \mathbb{E}\left[\mathbf{1}_{\{X_{1,0}(s)=1\}}\right] \mathbf{1}_{\{y\in C_{1,0}^-(s)\}} dyds,$$

$$= q_0 - \lambda \int_0^t \eta_0(s) \sum_{j=1}^K p_j \left(\frac{\mu}{K} + (1-\mu)\eta_j(s)\right) ds. \tag{35}$$

For $k \geq 1$, we have

$$\eta_k(t) = q_k + \lambda \int\limits_{[0,K)\times[0,t]} \left(\mathbb{E}\left[\left(1-\mathbf{1}_{\{X_{1,k}(s)=1\}}\right)\mathbf{1}_{\{y\in C_{1,k}^+(s)\}}\right] - \mathbb{E}\left[\mathbf{1}_{\{X_{1,k}(s)=1\}}\right]\mathbf{1}_{\{y\in C_{1,k}^-(s)\}}\right) dyds,$$

$$= q_k + \lambda \int_0^t p_k\left(\frac{\mu}{K}\eta_0(s) + \left(1-\mu\eta_0(s)\right)\eta_k(s)\right) - p_k\eta_k(s)^2 - \eta_k(s) \sum_{\substack{j=1 \\ j\neq k}}^K p_j\eta_j(s)ds$$

$$= q_k + \lambda \int_0^t \left((1-\mu)p_k\eta_0(s) + \sum_{j=1}^K (p_k-p_j)\eta_j(s)\right)\eta_k(s) + \eta_0(s)\frac{\mu}{K}p_k ds, \tag{36}$$

where the last equality comes from simple algebraic manipulation of the expressions. Combining equations (35) and (36) we conclude that $\eta(\cdot)$ is the solution to the ODE defined by equations (6) and (7), with initial condition q.

D   PROOF OF THEOREM 4.10

We only prove the second case. The first one is just a special case of the proof.

We start by bounding the convergence rate of y_0. Our first characterization may be loose. However, we can use this loose bound to more refined characterization of the convergence rate of y_1. From (6), we have

$$\dot{y}_0(t) = -y_0(t)\lambda\frac{\mu}{K}\sum_{k'=1}^K p_{k'} - y_0(t)\lambda \sum_{k'=1}^K (1-\mu)p_{k'}y_{k'}(t)] \leq -y_0(t)\lambda\frac{\mu}{K}\sum_{k'=1}^K p_{k'}. \tag{37}$$

Then, Gronwall's inequality implies

$$y_0(t) \leq y_0(0) \exp\left\{-\left(\lambda\frac{\mu}{K}\sum_{k=1}^K p_k\right)t\right\}. \tag{38}$$

Although the bound in (38) is only for one entry of y, it can help us to get a convergence rate for y_1. Note that, at time

$$\bar{t}_c = \frac{\log\frac{1}{c}}{\lambda\frac{\mu}{K}\sum_{k=1}^K p_k},$$

we have

$$y_0(\bar{t}_c) \leq y_0(0)c.$$

By (7), we know that

$$y_1(\bar{t}_c) \geq \frac{y_0(0)(1-c)}{K}.$$

On the other hand, from (6), we have

$$
\begin{aligned}
\dot{y_1}(t) &= y_0(t)\lambda\frac{\mu}{K}p_1 + \lambda\left((1-\mu)p_1 y_0(t) + \sum_{k=1}^{K}(p_1 - p_k)y_k(t)\right)y_1(t) \\
&\geq y_0(t)y_1(t)\lambda\frac{\mu}{K}p_1 + \lambda\left((1-\mu)p_1 y_0(t) + (p_1 - p_2)\sum_{k=2}^{K}y_k(t)\right)y_1(t) \\
&= \lambda\left(\left(1-\mu+\frac{\mu}{K}\right)p_1 y_0(t) + (p_1 - p_2)\sum_{k=2}^{K}y_k(t)\right)y_1(t) \\
&\geq \lambda\cdot\min\left\{\left(1-\mu+\frac{\mu}{K}\right)p_1, p_1 - p_2\right\}\left(y_0(t) + \sum_{k=2}^{K}y_k(t)\right)y_1(t) \\
&= \lambda\cdot\min\left\{\left(1-\mu+\frac{\mu}{K}\right)p_1, p_1 - p_2\right\}(1-y_1(t))\,y_1(t). \tag{39}
\end{aligned}
$$

Let us define

$$
R \triangleq \lambda\min\left\{\left(1-\mu+\frac{\mu}{K}\right)p_1, p_1 - p_2\right\},
$$

and let z_1 be an auxiliary ODE equation such that

$$
\dot{z_1} = R\,(1-z_1)\,z_1, \tag{40}
$$

with

$$
z_1(0) \triangleq y_1(\bar{t}_c) \geq \frac{y_0(0)(1-c)}{K}. \tag{41}
$$

It can be shown that for all $t \in [\bar{t}_c, \infty)$,

$$
y_1(\bar{t}_c + t) \geq z_1(t). \tag{42}
$$

Thus, the convergence rate of z_1 provides a lower bound of the convergence rate of the original ODE system. Note that z_1 is an autonomous and separable. Thus, we have

$$
\begin{aligned}
z_1(t) &= 1 - \frac{1}{\frac{z_1(0)}{1-z_1(0)}\exp\left(Rt\right) + 1} \\
&\geq 1 - \frac{1}{\frac{y_0(0)(1-c)}{K-y_0(0)(1-c)}\exp\left(Rt\right) + 1},
\end{aligned}
$$

where the last inequality follows from (41). By (42), we know that $y_1(\bar{t}_c + t) \geq z_1(t)$. Therefore, we conclude that

$$
y_1(t) \geq 1 - \frac{1}{\frac{y_0(0)(1-c)}{K-y_0(0)(1-c)}\exp\left[R(t - t_c)\right] + 1},
$$

for all $t \geq \bar{t}_c$.