Nonasymptotic Convergence Rates for Cooperative Learning Over Time-Varying Directed Graphs

Angelia Nedić, Alex Olshevsky and César A. Uribe

Abstract

We study the problem of distributed hypothesis testing with a network of agents where some agents repeatedly gain access to information about the correct hypothesis. The group objective is to globally agree on a joint hypothesis that best describes the observed data at all the nodes. We assume that the agents can interact with their neighbors in an unknown sequence of time-varying directed graphs. Following the pioneering work of Jadbabaie, Molavi, Sandroni, and Tahbaz-Salehi, we propose local learning dynamics which combine Bayesian updates at each node with a local aggregation rule of private agent signals. We show that these learning dynamics drive all agents to the set of hypotheses which best explain the data collected at all nodes as long as the sequence of interconnection graphs is uniformly strongly connected. Our main result establishes a non-asymptotic, explicit, geometric convergence rate for the learning dynamic.

I. INTRODUCTION

Recent years have seen a considerable amount of work on the analysis of distributed algorithms, which have found applications in opinion dynamics, network learning and inference, cooperative robotics, communication networks, and social as well as sensor networks. It is the latter settings of social and sensor networks which is the focus of the current paper. Interactions among people produce exchange of ideas, opinions, observations and experiences, on which new ideas, opinions, and observations are generated. Analyzing the dynamics which model such processes has the potential both to gain insight into human behavior as well as to produce algorithms which will be useful in the sensor networking context. Although many results on these themes have appeared in recent years, the study of distributed decision making and computation can be traced back to the classic papers [1], [2], [3] from the 70s and 80s.

The authors are with the Coordinated Science Laboratory, University of Illinois, 1308 West Main Street, Urbana, IL 61801, USA, {angelia,aolshev2,cauribe2}@illinois.edu. This research is supported partially by the National Science Foundation under grant no. CCF-1017564 and by the Office of Naval Research under grant no. N00014-12-1-0998.
Here we consider an agent network where agents repeatedly receive information from their neighbors and private signals from an external source, which provide noisy information about a hypothesis of interest. The agents would like to collectively agree on a hypothesis that best explains the data. In a centralized setup, this can be achieved by the repeated applications of Bayes’ rule. In a distributed setting, decentralized protocols for learning and hypothesis testing problems have been proposed relatively recently.

Initial results on Bayesian learning for social networks are described in [4], wherea local update rule is designed such that it matches the Bayes’ Theorem. That is, given a prior and new observations, the agent is able to compute likelihood functions in order to generate a new posterior, see [5]. Nevertheless, a fully Bayesian approach might not be possible in general since full knowledge of neither the network structure nor other agents likelihood functions might be available [6]. Fortunately, non-Bayesian methods have been shown to be successful in learning as well. For example, in [7], the authors propose a modification of Bayes’ rule that accounts for over-reactions or under-reactions to new information.

In a distributed setting, the groundbreaking paper of [8] describes a way for agents in a network to repeatedly aggregate local Bayes estimates via arithmetic averages to result in global learning. Following [8], other methods to aggregate Bayes estimates in a network have been explored, for example via geometric averages in [9], [10]. Other extensions explore the same results for time-varying topologies [11], [12], [13]. In [8], the almost sure convergence of a non-Bayesian rule based on arithmetic mean is shown for a fixed topology graph. In [9], geometric means are used for fixed topologies as well, however the consensus and learning steps are separated. The work in [11] extends the results of [8] to time-varying undirected graphs. In [14], local exponential rates of convergence for undirected gossip-like graphs are studied. In [10], a non-Bayesian learning algorithm is proposed where a local Bayes update is followed by a consensus step. It is shown that asymptotically the process converges at a geometric rate and an upper bound for the exponent is provided.

In this paper we propose a new non-Bayesian learning rule. Our first result shows consistency: we show that over time, the protocol learns the hypothesis or set of hypotheses which explain the data collected by all the nodes best. Our main result provides a geometric, non-asymptotic, and explicit characterization of the rate of convergence which immediately leads to finite-time bounds which scale intelligibly with the number of nodes.

This paper is organized as follows. In Section II we describe the model that we study and the proposed update rule. In Section III we analyze the consistency of the information aggregation and estimation models, while in Section IV we establish non-asymptotic convergence rates of the agent
beliefs. Conclusions and future work are given in Section VI.

Notation: We use the upper case letters to represent random variables (e.g. $X_k$), and the corresponding lower case letters for their realizations (e.g. $x_k$). We write $[A_k]_{ij}$ to denote the entry of the matrix $A_k$ in the $i$-th row and $j$-th column. We write $A'$ for the transpose of a matrix $A$ and $x'$ for the transpose of a vector $x$. We use $I$ for the identity matrix. Bold letters represent vectors which are assumed to be column vectors unless specified otherwise. The $i$'th entry of a vector will be denoted by a superscript $i$, i.e., $x_k = [x_k^1, \ldots, x_k^n]^T$. We write $1_n$ to denote the all-ones vector of size $n$. For a sequence of matrices $\{A_i\}$, we let $A_{t_f:t_i} \triangleq A_{t_f} \cdots A_{t_i+1} A_{t_i}$ for all $t_f \geq t_i \geq 0$. We abbreviate terminology "almost surely" by a.s. and "independent identically distributed" by i.i.d.

II. Problem Setup and Main Results

We consider a group of $n$ agents each of which observes a random variable at each time step $k = 1, 2, 3, \ldots$. We will use $S_k^i$ to denote the random variable observed by agent $i$ at time step $k$. We denote the set of outcomes of the random variable $S_k^i$ by $S^i$, and we assume that this set is finite, i.e., $S^i = \{s^i_1, s^i_2, \ldots, s^i_m\}$ for all $i = 1, \ldots, n$. Furthermore, we assume that all $S_k^i$ are i.i.d. and drawn according to some probability distribution $f^i(\cdot)$. For convenience, we stack up all the $S_k^i$ into a vector $S_k$. Then, $S_k$ is an i.i.d. vector taking values in $\mathcal{S} = \prod_{i=1}^n S^i$. We denote the distribution of $S_k$ by $f(\cdot)$. We assume there is a finite set $\Theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$ with $m$ elements, and for each agent $i$ and each $\theta \in \Theta$, there is a probability distribution $l_i(\cdot|\theta) : S^i \rightarrow [0, 1]$. Intuitively, we think of $\Theta$ as a set of hypotheses and $l_i(\cdot|\theta)$ as the probability distribution seen by agent $i$ if hypothesis $\theta$ were true. The goal of the agents is to agree on an element of $\Theta$ that fits all the observations in the network best (in a technical sense to be described soon).

Agents communicate with their neighbors in some communication network. At each time instant $k$, we denote the graph as $G_k = \{V, E_k\}$ composed of a node set $V = \{1, 2, \ldots, n\}$ and a set of directional links $E_k$. i.e., each graph $G_k$ is directed. We assume that the agents can send messages to their out-neighbors in $G_k$ at time $k$.

We will refer to probability distributions over $\Theta$ as beliefs. We assume that agent $i$ begins with an initial belief $\mu^i_0$, which we also refer to as its prior distribution or prior belief. We will study dynamics wherein, at time $k$, each agent $i$ updates its previous belief $\mu^i_k$ to a new belief $\mu^i_{k+1}$ as follows:

$$\mu^i_{k+1}(\theta) = \frac{\prod_{j=1}^n \mu^j_k(\theta)^{[A_k]_{ij}} \cdot l_i(s^i_{k+1}|\theta)}{\sum_{p=1}^m \prod_{j=1}^n \mu^j_k(\theta_p)^{[A_k]_{ij}} \cdot l_i(s^i_{k+1}|\theta_p)},$$

October 9, 2014

DRAFT
with \([A_k]_{ij} > 0\) when \(i\) receives information from \(j\) at time \(k\), and else \([A_k]_{ij} = 0\). The “weight matrices” \(A_k\) satisfy some technical connectivity conditions which have been previously used in convergence analysis of distributed averaging and other consensus algorithms [15], [16], [17].

**Assumption 1** The graph sequence \(\{G_k\}\) and a matrix sequence \(\{A_k\}\) are such that:

1. \(A_k\) is row-stochastic with \([A_k]_{ij} > 0\) if \((j, i) \in E_k\).
2. \(A_k\) has positive diagonal entries, \([A_k]_{ii} > 0\).
3. If \([A_k]_{ij} > 0\) then \([A_k]_{ij} > \eta\) for some positive constant \(\eta\).
4. \(\{G_k\}\) is \(B\)-strongly connected, i.e., there is an integer \(B \geq 1\) such that the graph \(\{V, \bigcup_{i=kB}^{(k+1)B-1} E_i\}\) is strongly connected for all \(k \geq 0\).

Assumption 1 corresponds to information exchange which occurs when nodes broadcast their beliefs to out-neighbors: if \((j, i)\) belongs to the graph at time \(k\), then node \(i\) uses \(j\)’s belief in its update, but not necessarily vice versa.

As a measure for the explanatory quality of the hypotheses in the set \(\Theta\) we use the Kullback-Leibler divergence between two discrete probability distributions \(p\) and \(q\):

\[
D_{KL}(p\|q) = \sum_{i=1}^{n} p_i \log \left( \frac{p_i}{q_i} \right).
\]

Concretely, the quality of hypothesis \(\theta_j\) for agent \(i\) is measured by the Kullback-Leibler divergence \(D_{KL}(f^i(\cdot)\|l^i(\cdot|\theta_j))\) between the true distribution of the signals \(S^i_k\) and the probability distribution \(l_i(\cdot|\theta_j)\) as seen by agent \(i\) if hypothesis \(\theta_j\) were correct. We use the following assumption on the agents’ best hypotheses.

**Assumption 2** Define \(\Theta_i = \arg\min_{\theta \in \Theta} D_{KL}(f^i(\cdot)\|l^i(\cdot|\theta))\) for each \(i\), and let \(\Theta^* = \bigcap_{i=1}^{n} \Theta_i\), then \(\Theta^* \neq \emptyset\).

Assumption 2 is satisfied if there is some “true state of the world” \(\hat{\theta} \in \Theta\) such that each agent \(i\) sees distributions generated according to \(\hat{\theta}\), i.e., \(f^i(\cdot) = l_i(\cdot|\hat{\theta})\). However, this need not be the case for Assumption 2 to hold. Indeed, the assumption is considerably weaker as it merely requires that the set of hypotheses, which provide the “best fits” for each agent, have at least a single element in common.

We will further require the following assumptions on the initial distribution and the likelihood functions. The first of these is sometimes referred to as the Zero Probability Property [18].

**Assumption 3** For all agents \(i = 1, \ldots, n\),
1) The prior beliefs on all \( \theta^* \in \Theta^* \) are positive, i.e. \( \mu_i^0(\theta^*) > 0 \) for all \( \theta^* \in \Theta^* \).

2) There exists an \( \alpha > 0 \) such that \( l_i(s^i|\theta) > \alpha \) for all \( s^i \in S^i \) and \( \theta \in \Theta \).

Assumption 3.1 can be relaxed to a requirement that all prior beliefs are positive for some \( \theta^* \in \Theta^* \). Both of these conditions are equally complex to be satisfied. They can be satisfied by letting each agent have a uniform prior belief, which is reasonable in the absence of any initial information about the goodness of the hypotheses.

We can now state our first result, which asserts that the dynamics in (1) succeed in driving all agents to believe the hypotheses in the optimal set \( \Theta^* \).

**Theorem 1** Under Assumptions 1, 2, and 3, the update rule of Eq. (1) has the following properties:

\[
\lim_{k \to \infty} \mu_i^k(\theta) = 0 \quad \forall \theta \notin \Theta^*, \ i = 1, \ldots, n.
\]

The result states that the agents’ beliefs will concentrate on the set \( \Theta^* \) asymptotically as \( k \to \infty \).

Our main result is a non-asymptotic explicit convergence rate, given in the following theorem.

**Theorem 2** Let Assumptions 1, 2, and 3 hold. Also, let \( \rho \in (0, 1) \) be a given error percentile (or confidence value). Then, the update rule of Eq. (1) has the following property: for any \( \theta \notin \Theta^* \), there is an integer \( N(\rho) \) such that, with probability \( 1 - \rho \), for all \( k \geq N(\rho) \) there holds

\[
\mu_i^k(\theta) \leq \exp \left( -\frac{k}{2} \gamma_2 + \gamma_1 \right) \quad \forall i = 1, \ldots, n,
\]

where

\[
N(\rho) \triangleq \left\lceil \frac{8 (\log(\alpha))^2 \log \left( \frac{1}{\rho} \right)}{\gamma_2^2} + 1 \right\rceil,
\]

\[
\gamma_1 \triangleq \max_{\theta^* \in \Theta^*} \left\{ \max_{1 \leq i \leq n} \log \frac{\mu_i^0(\theta^*)}{\mu_i^0(\theta^*)} + \frac{C}{1 - \lambda} \|H(\theta, \theta^*)\|_1 \right\},
\]

\[
\gamma_2 \triangleq \frac{\delta}{n} \min_{\theta^* \in \Theta^*} \|H(\theta, \theta^*)\|_1.
\]

\[
[H(\theta, \theta^*)]_i = D_{KL}(f^i(\cdot) || l_i(\cdot | \theta)) - D_{KL}(f^i(\cdot) || l_i(\cdot | \theta^*))
\]

with \( \alpha \) from Assumption 3.2. The constants \( C, \delta \) and \( \lambda \) satisfy the following relations:

(1) For general \( B \)-connected graph sequences \( \{G_k\} \),

\[
C = 2, \quad \lambda \leq \left( 1 - \eta^{nB} \right)^{\frac{1}{n}}, \quad \delta \geq \frac{1}{\eta^{nB}}.
\]
(2) If every matrix $A_k$ is doubly stochastic,

$$C = \sqrt{2}, \quad \lambda = \left(1 - \frac{\eta}{n^2}\right)^{\frac{1}{2}}, \quad \delta = 1.$$ 

(3) If each $G_k$ is an undirected graph and each $A_k$ is the lazy Metropolis matrix, i.e. the stochastic matrix which satisfies

$$[A_k]_{ij} = \frac{1}{2 \max(d(i), d(j))} \quad \text{for all } \{i, j\} \in G_k,$$

then

$$C = \sqrt{2}, \quad \lambda = 1 - \frac{1}{\Theta(n^2)}, \quad \delta = 1.$$ 

In contrast to the previous literature, these convergence rates are not only geometric but also non-asymptotic and explicit in the sense of immediately leading to bounds which scale intelligible in terms of the number of nodes. For example, in the case of doubly stochastic matrices, Theorem 2 immediately implies that, after a transient time, which scales cubically in the number $n$ of nodes, the network will achieve exponential decay to the correct answer with the exponent $-\frac{1}{2} \min_{\theta^* \in \Theta^*} \| H(\theta, \theta^*) \|_1/n$.

Now, consider the case when Assumption 3.1 is relaxed to the following requirement: The prior beliefs on some $\theta^* \in \Theta^*$ are positive (i.e. $\mu^0_i(\theta^*) > 0$ for some $\theta^* \in \Theta^*$ and all $i$). Then, it can be seen that the Theorem 2 is valid with $\max_{\theta^* \in \Theta^*}$ and $\min_{\theta^* \in \Theta^*}$ replaced, respectively, by $\max_{\theta^* \in \tilde{\Theta}^*}$ and $\max_{\theta^* \in \tilde{\Theta}^*}$, where $\tilde{\Theta}^* \subseteq \Theta^*$ is the set of all $\theta^* \in \Theta^*$ for which all the agents priors $\mu^i_0$ are positive.

A. Motivation

We now describe the motivation for the update rule of Eq. (1). Standard Bayes’ rule can be described as the solution of a constrained optimization problem [19], [20], [21]. The cost function to be minimized is composed of two terms: one being the Maximum Likelihood Estimation (MLE) of a state given the observed data and the other being a regularization function minimized by the current prior [20], i.e.,

$$\mu_{k+1}(\theta) = \arg \min_{\pi \in \mathcal{P}(\Theta)} \left\{ D_{KL}(\pi, \mu_k) - \mathbb{E}_\pi \left[ \log (l(s_{k+1}|\theta)) \right] \right\}$$

$$= \frac{\mu_k(\theta) l(s_{k+1}|\theta)}{\sum_{p=1}^m \mu_k(\theta_p) l(s_{k+1}|\theta_p)},$$

where $s_{k+1}$ is the most recent observation, $l(\cdot|\theta)$ is the likelihood function for hypothesis $\theta$, $\mathbb{E}_\pi$ is the expected value with respect to the probability distribution $\pi$, and $\mathcal{P}(\Theta)$ is the set of all probability distributions on the set $\Theta$.

One can modify the optimization problem associated with a Bayesian update to take into account the network structure. This is done by changing the KL divergence term from a single prior belief to a convex
The solution for the new optimization problem is precisely given by the proposed update rule (1). This update rule can be seen as a two step procedure. First, prior beliefs of the neighbor set are combined according to an opinion aggregation function. Second, the resulting aggregate distribution is updated using Bayes’ rule.

B. Generalized Distributed non-Bayesian Learning

Opinion pooling or opinion aggregation has been studied before in [18], [22], [23], [24]. It is considered a traditional problem in economics, where several experts have beliefs about a hypothesis and one needs to aggregate their beliefs into a single probability distribution. Different opinion aggregation functions result from using different divergence metric for probability distributions (see [25]). In the same way, different opinion pool operators define different non-Bayesian distributed learning rules. A general form of opinion pooling was introduced in [22], termed \textit{g-Quasi-Linear Opinion pools} (g-QLOP), defined as follows:

$$\tau_g(\ldots, \mu_{jk}^l(\theta), \ldots) = \frac{g^{-1}\left(\sum_{j=1}^{n} a_{ij} g(\mu_k^i(\theta))\right)}{\sum_{p=1}^{m} g^{-1}\left(\sum_{j=1}^{n} a_{ij} g(\mu_k^j(\theta_p))\right)}.$$ 

with \(\tau : \prod_{i=1}^{n} \mathbb{P}(\Theta) \rightarrow \mathbb{P}(\Theta)\). The g-QLOP produces the same results as \textit{LinPool} when \(g(x) = x\) and \textit{LogPool} when \(g(x) = \log x\). Note that consensus will be reached if and only if \(g(x) = x\) or \(g(x) = \log x\), see Theorem 4 in [22].

The proposed update rule (1) uses the Logarithmic Opinion Pool, where

$$\tau(\ldots, \mu_k^j(\theta), \ldots) = \frac{\prod_{j=1}^{n} \mu_k^j(\theta)[A_k]_{ij}}{\sum_{p=1}^{m} \prod_{j=1}^{n} \mu_k^j(\theta_p)[A_k]_{ij}}.$$ 

Logarithmic Pools are Externally Bayesian [18], [26], so the order in which the new evidence is included does not influence the consensus. That is, from a learning point of view, if the function is Externally Bayesian, the innovation and diffusion parts can be interchanged. Therefore, the order in which the opinion aggregation and the Bayesian update are performed does not change the update rule.

Consider now a Linear Opinion pool, where

$$\tau(\ldots, \mu_k^j(\theta), \ldots) = \sum_{j=1}^{n} [A_k]_{ij} \mu_k^j(\theta).$$
If the opinion aggregation is done first, then the resulting update rule is

$$\mu_{k+1}^i(\theta) = \frac{\sum_{j=1}^{n} [A_k]_{ij} \mu_k^j(\theta) l_i (s_{k+1}^i | \theta)}{\sum_{p=1}^{m} \sum_{j=1}^{n} [A_k]_{ij} \mu_k^j(\theta) l_i (s_{k+1}^i | \theta_p)}.$$ 

On the other hand if the Bayesian Update is done first, then the resulting update rule is

$$\mu_{k+1}^i(\theta) = \frac{\sum_{j=1}^{n} [A_k]_{ij} \mu_k^j(\theta) l_j (s_{k+1}^j | \theta)}{\sum_{p=1}^{m} \mu_k^j(\theta) l_j (s_{k+1}^j | \theta_p)}.$$ (2)

The Linear Pool-based update rule is similar to the update rule proposed in [8]. The main difference is in the fact that in (2), a convex combination of the posteriors received from the neighbor set is used to generate the new individual posterior, while in [8], the update rule is a convex combination of the individual posterior and the neighbors’ priors. Our update rule (1) is a Logarithmic-Pool analog of the rule in [8].

III. CONSISTENCY

This section provides the proof for Theorem 1 which provides a statement about the consistency of the consensus-like distributed estimator of Eq. (1) (see [27], [28]). Our analysis will require some auxiliary results. First, we will recall some results from [29] about the convergence of a product of row stochastic matrices.

**Lemma 1** [29], [30] Under Assumption 1 for a graph sequence \(\{G_k\}\) and each \(t \geq 0\), there is a stochastic vector \(\phi_t\) (meaning its entries are nonnegative and sum to one) such that for all \(i, j\) and \(k \geq t\),

$$| [A_{k:t}]_{ij} - \phi_t^i | \leq C \lambda^{k-t} \quad \forall k \geq t \geq 0$$

where \(C > 0\) and \(\lambda \in (0, 1)\) satisfy the relations described in Theorem 2.

**Proof:** The proof may be found in [29], with the exception of the bounds on \(C, \lambda\) for the lazy Metropolis chains which we omit here due to space constraints.

**Lemma 2** [29] Let the graph sequence \(\{G_k\}\) satisfy Assumption 1. Define

$$\delta \triangleq \inf_{k \geq 0} \left( \min_{1 \leq i \leq n} \left[ 1 - \frac{1}{n} A_{k:0} \right]_{ii} \right).$$ (3)

Then, \(\delta \geq \eta n B\), and if all \(A_k\) are doubly stochastic, then \(\delta = 1\). Furthermore, the sequence \(\phi_t\) from Lemma 1 satisfies \(\phi_t^i \geq \delta / n\) for all \(t \geq 0, j = 1, \ldots, n\).
Next, we need a technical lemma regarding the weighted average of random variables with a finite variance.

**Lemma 3** Assume that the graph sequence \( \{G_k\} \) satisfies Assumption \[.\] Also, let Assumptions \[2\] and \[\] hold. Then, we have for any \( \theta \notin \Theta^* \) and \( \theta^* \in \Theta^* \),

\[
\lim_{k \to \infty} \frac{1}{k} \sum_{t=1}^{k} A_{k:t} \mathcal{L}_t^\theta + \frac{1}{k} \sum_{t=1}^{k} \mathbb{1}_n \phi'_t \mathbf{H} (\theta, \theta^*) = 0 \quad \text{a.s.}
\]

where \( \mathcal{L}_t^\theta \) is the random vector with coordinates given by

\[
\left[ \mathcal{L}_t^\theta \right]_i = \log \frac{l_i(S_i^t | \theta)}{l_i(S_i^t | \theta^*)} \quad \forall i = 1, \ldots, n,
\]

while the vector \( \mathbf{H} (\theta, \theta^*) \) has coordinates given by \( \left[ \mathbf{H} (\theta, \theta^*) \right]_i = D_{KL}(f^i(\cdot) || l_i(\cdot | \theta)) - D_{KL}(f^i(\cdot) || l_i(\cdot | \theta^*)) \).

**Proof:** Adding and subtracting \( \frac{1}{k} \sum_{t=1}^{k} \mathbb{1}_n \phi'_t \mathbf{L}_t^\theta \) yields

\[
\frac{1}{k} \sum_{t=1}^{k} A_{k:t} \mathcal{L}_t^\theta + \frac{1}{k} \sum_{t=1}^{k} \mathbb{1}_n \phi'_t \mathbf{H} (\theta, \theta^*) = \frac{1}{k} \sum_{t=1}^{k} (A_{k:t} - \mathbb{1}_n \phi'_t) \mathcal{L}_t^\theta + \frac{1}{k} \sum_{t=1}^{k} \mathbb{1}_n \phi'_t \left( \mathcal{L}_t^\theta + \mathbf{H} (\theta, \theta^*) \right). \tag{4}
\]

By Lemma \[1\] \( \lim_{k \to \infty} A_{k:t} = \mathbb{1} \phi'_t \) for all \( t \geq 0 \). Moreover, each of the entries of \( \mathcal{L}_t^\theta \) are upper bounded by Assumption \[2\] Thus, the first term on the right hand side of (4) goes to zero as we take the limit over \( k \to \infty \). Regarding the second term in (4), by the definition of the KL divergence measure, we have that

\[
\mathbb{E} \left[ \log \frac{l_i(S_i^t | \theta)}{l_i(S_i^t | \theta^*)} \right] = \sum_{j=1}^{m_i} f^i(s_j^i) \log \frac{l_i(s_j^i | \theta)}{l_i(s_j^i | \theta^*)}
\]

\[
= \sum_{j=1}^{m_i} f^i(s_j^i) \log \left( \frac{l_i(s_j^i | \theta)}{l_i(s_j^i | \theta^*)} \frac{f^i(s_j^i | \theta)}{f^i(s_j^i | \theta^*)} \right)
\]

\[
= D_{KL}(f^i(\cdot) || l_i(\cdot | \theta^*)) - D_{KL}(f^i(\cdot) || l_i(\cdot | \theta)),
\]

or equivalently

\[
\mathbb{E} \left[ \mathcal{L}_t^\theta \right] = -\mathbf{H} (\theta, \theta^*).
\]

Kolmogorov’s strong law of large numbers states that if \( \{X_t\} \) is a sequence of independent random variables with variances such that \( \sum_{k=1}^{\infty} \frac{\text{Var}(X_k)}{k^2} < \infty \), then \( \frac{1}{n} \sum_{k=1}^{n} X_k - \frac{1}{n} \sum_{k=1}^{n} \mathbb{E} [X_k] \to 0 \) a.s. Let \( X_t = \phi'_t \mathcal{L}_t^\theta \). Then, by using Assumptions \[1\] and \[2\], it can be seen that \( \sup_{t \geq 0} \text{Var} (X_t) < \infty \).

The result follows by Lemma \[1\] and Kolmogorov’s strong law of large numbers. \( \blacksquare \)

With Lemma \[3\] in place, we are ready to prove Theorem \[1\]. The proof of Theorem \[1\] (and also Theorem \[2\]) makes use of the following quantities: for all \( i = 1, \ldots, n \) and \( k \geq 0 \),

\[
\varphi_k^i (\theta) \triangleq \log \frac{\mu_k^i (\theta)}{\mu_k^i (\theta^*)} \quad \text{for all } \theta \in \Theta,
\]
defined for any $\theta^* \in \Theta^*$ (dependence on $\theta^*$ is suppressed).

**Proof:** (Theorem 2) Dividing both sides of (1) by $\mu_{k+1}^\theta (\theta^*)$ and by taking logs we obtain:

$$\log \frac{\mu_{k+1}^\theta (\theta)}{\mu_{k+1}^\theta (\theta^*)} = \log \frac{\prod_{j=1}^n \mu_j^i (\theta)[\delta_{k+1}], l_i (s_{k+1}^i | \theta)}{\prod_{j=1}^n \mu_j^i (\theta^*)[\delta_{k+1}], l_i (s_{k+1}^i | \theta^*)}.$$ 

Using the definition of $\varphi_{k+1}^i (\theta)$, we can write the preceding relation equivalently:

$$\varphi_{k+1}^i (\theta) = \sum_{j=1}^n [A_{kj}]_{ij} \varphi_j^i (\theta) + \log \frac{l_i (s_{k+1}^i | \theta)}{l_i (s_{k+1}^i | \theta^*)}.$$ 

Stacking up the values $\varphi_{k+1}^i (\theta)$ over agents $i = 1, \ldots, n$, into a single vector $\varphi_{k+1} (\theta)$, we can compactly write the preceding relations, as follows:

$$\varphi_{k+1} (\theta) = A_k \varphi_k (\theta) + L_{k+1}^\theta,$$

which implant that for all $k \geq 0$,

$$\varphi_{k+1} (\theta) = A_{k:0} \varphi_0 (\theta) + \sum_{t=1}^k A_{k:t} L_t^\theta + L_{k+1}^\theta.$$ 

We add and subtract $\sum_{t=1}^k 1_n \phi^t_i H (\theta, \theta^*)$ in (7) to obtain

$$\varphi_{k+1} (\theta) = A_{k:0} \varphi_0 (\theta) + \sum_{t=1}^k \left( A_{k:t} L_t^\theta + 1_n \phi^t_i H (\theta, \theta^*) \right) + L_{k+1}^\theta - \sum_{t=1}^k 1_n \phi^t_i H (\theta, \theta^*).$$ 

By using the lower bounds on $\phi_i$ described in Lemma 2 and the fact that $H(\theta, \theta^*) \geq 0$, we obtain

$$\varphi_{k+1} (\theta) \leq A_{k:0} \varphi_0 (\theta) + \sum_{t=1}^k \left( A_{k:t} L_t^\theta + 1_n \phi^t_i H (\theta, \theta^*) \right) + L_{k+1}^\theta - \frac{\delta}{n} ||H(\theta, \theta^*)||_1 1_n.$$ 

Therefore, we have

$$\lim_{k \to \infty} \frac{1}{k} \varphi_{k+1} (\theta) \leq \lim_{k \to \infty} \frac{1}{k} A_{k:0} \varphi_0 (\theta) - \frac{\delta}{n} ||H(\theta, \theta^*)||_1 1_n + \lim_{k \to \infty} \frac{1}{k} L_{k+1}^\theta + \lim_{k \to \infty} \frac{1}{k} \sum_{t=1}^k \left( A_{k:t} L_t^\theta + 1_n \phi^t_i H (\theta, \theta^*) \right).$$ 

The first term of the right hand side of the preceding relation converges to zero deterministically. The third term goes to zero as well since $L_t^\theta$ is bounded, and the fourth term converges to zero almost surely by Lemma 3. Consequently,

$$\lim_{k \to \infty} \frac{1}{k} \varphi_{k+1} (\theta) \leq - \frac{\delta}{n} ||H(\theta, \theta^*)||_1 1_n \quad a.s.$$ 

Now if $\theta \notin \Theta^*$, then $H(\theta, \theta^*) > 0$ and, thus, $\varphi_k (\theta) \to -\infty$ almost surely. This implies $\mu_k (\theta) \to 0$ almost surely. 

IV. Rates of Convergence

In this section, we prove Theorem 2 which states an explicit rate of convergence for cooperative agent learning process. Before proving the theorem, we will estate an auxiliary lemma that provides a bound for the expectation of the random variables \( \varphi_k^i(\theta) \) as defined in (5).

**Lemma 4** Let \( \theta^* \in \Theta^* \) be arbitrary, and consider \( \varphi_k^i(\theta) \) as defined in (5). Then, for any \( \theta \notin \Theta^* \) we have

\[
\mathbb{E} [\varphi_{k+1}^i(\theta)] \leq \gamma_1 - k\gamma_2 \quad \text{for all } i \text{ and } k \geq 0,
\]

where

\[
\gamma_1 \triangleq \max_{\theta^* \in \Theta^*} \left\{ \| \varphi_0(\theta) \|_\infty + \frac{C}{1 - \lambda} \| H(\theta, \theta^*) \|_1 \right\},
\]

\[
\gamma_2 \triangleq \frac{\delta}{n} \min_{\theta^* \in \Theta^*} \| H(\theta, \theta^*) \|_1.
\]

**Proof:** Grouping \( \varphi_{k+1}^i(\theta) \) for all agents into a single vector \( \varphi_{k+1}(\theta) \), we obtain equation (6), from which by taking the expected values and using \( \mathbb{E} [\mathcal{L}^\theta_{k+1}] = -H(\theta, \theta^*) \), we obtain

\[
\mathbb{E} [\varphi_{k+1}(\theta)] = A_k \mathbb{E} [\varphi_k(\theta)] - H(\theta, \theta^*).
\]

Therefore, by recursion we can see that for all \( k \geq 0 \),

\[
\mathbb{E} [\varphi_{k+1}(\theta)] = A_{k:0} \varphi_0(\theta) - \sum_{t=1}^k A_{k:t} H(\theta, \theta^*) - H(\theta, \theta^*).
\]

By adding and subtracting \( \sum_{t=1}^k I_n \phi_t^j H(\theta, \theta^*) \), we obtain

\[
\mathbb{E} [\varphi_{k+1}(\theta)] = A_{k:0} \varphi_0(\theta) + \sum_{t=1}^k (I_n \phi_t^j - A_{k:t}) H(\theta, \theta^*) - \sum_{t=1}^k I_n \phi_t^j H(\theta, \theta^*) - H(\theta, \theta^*).
\]

By bounding the entries for the first two terms on the right hand side of the preceding relation, and using the fact that \( A_{k:0} \) is stochastic matrix, we find that

\[
\mathbb{E} [\varphi_{k+1}(\theta)] \leq \| \varphi_0(\theta) \|_\infty I_n + \sum_{t=1}^k \max_{1 \leq i,j \leq n} | \phi_t^j - [A_{k:t}]_{ij} | \| H(\theta, \theta^*) \|_1 I_n - \sum_{t=1}^k I_n \phi_t^j H(\theta, \theta^*) - H(\theta, \theta^*).
\]

Next, we use the upper bound on terms \( | \phi_t^j - [A_{k:t}]_{ij} | \) from Lemma 1 and the lower bound for the entries in \( \phi_t \) as given in Lemma 2 and we arrive at the following relation:

\[
\mathbb{E} [\varphi_{k+1}(\theta)] \leq \| \varphi_0(\theta) \|_\infty I_n + \sum_{t=1}^k C \lambda^{k-t} \| H(\theta, \theta^*) \|_1 I_n - \frac{k\delta}{n} H(\theta, \theta^*) - H(\theta, \theta^*).
\]
In view of $H(\theta, \theta^*) > 0$, we can remove the last term of the right hand side in the preceding relation. Therefore, for all $k \geq 0$,
\[
\mathbb{E} \left[ \varphi_{k+1}(\theta) \right] \leq \|\varphi_0(\theta)\|_\infty 1_n + C \left( \sum_{t=1}^{k} \lambda^{k-t} \right) \|H(\theta, \theta^*)\|_1 1_n - k \frac{\delta}{n} \|H(\theta, \theta^*)\|_1 1_n
\]
\[
\leq \|\varphi_0(\theta)\|_\infty 1_n + \frac{C}{1 - \lambda} \|H(\theta, \theta^*)\|_1 1_n - k \frac{\delta}{n} \|H(\theta, \theta^*)\|_1 1_n.
\]

The result follows by letting
\[
\gamma_1 = \max_{\theta^* \in \Theta_*} \left\{ \|\varphi_0(\theta)\|_\infty + \frac{C}{1 - \lambda} \|H(\theta, \theta^*)\|_1 \right\},
\]
\[
\gamma_2 = \delta \min_{\theta^* \in \Theta_*} \|H(\theta, \theta^*)\|_1,
\]
and recalling the definition of $\varphi_0(\theta)$.

In the proof of Theorem 2 we will make use of McDiarmid’s inequality [31], which provides some bounds for concentration of probabilities. This inequality will allow us to establish bounds on the probability that the beliefs exceed a given value $\epsilon$. McDiarmid’s inequality is provided below.

**Theorem 3 (McDiarmid’s inequality [31])** Let $\{X_t\}_{t=1}^k = (X_1, \ldots, X_k)$ be a sequence of independent random variables with $X_t \in \mathcal{X}$. If a function $g : \{X_t\}_{t=1}^k \rightarrow \mathbb{R}$ has bounded differences, i.e., for all $t$,
\[
\sup_{X_t \in \mathcal{X}} g(\ldots, X_t, \ldots) - \inf_{Y_t \in \mathcal{X}} g(\ldots, Y_t, \ldots) \leq c_t
\]
then for any $\epsilon > 0$ and all $k \geq 1$,
\[
\mathbb{P} \left( g \left( \{X_t\}_{t=1}^k \right) - \mathbb{E} \left[ g \left( \{X_t\}_{t=1}^k \right) \right] \geq \epsilon \right) \leq \exp \left( -\frac{2\epsilon^2}{\sum_{t=1}^{k} c_t^2} \right)
\]

(9)

Now, we are ready to prove Theorem 2.

**Proof: (Theorem 2)** First we will express the belief $\mu_{k+1}^i(\theta)$ in terms of the variable $\varphi_{k+1}^i(\theta)$. This will allow us to use the McDiarmid’s inequality to obtain the concentration bounds. By the dynamics of the beliefs expressed in Eq. (1) and Assumption 3.1, since $\mu_{k+1}^i(\theta^*) \in (0, 1]$ for any $\theta^* \in \Theta^*$, we have
\[
\mu_{k+1}^i(\theta) \leq \frac{\mu_{k+1}^i(\theta)}{\mu_{k+1}^i(\theta^*)} = \exp \varphi_{k+1}^i(\theta).
\]

Therefore,
\[
\mathbb{P} \left( \mu_{k+1}^i(\theta) \geq \exp \left( -\frac{k}{2} \gamma_2 + \gamma_1 \right) \right) \leq \mathbb{P} \left( \exp \left( \varphi_{k+1}^i(\theta) \right) \right) \geq \exp \left( -\frac{k}{2} \gamma_2 + \gamma_1 \right)
\]
\[= \mathbb{P}\left( \varphi_{k+1}^i(\theta) \geq -\frac{k}{2}\gamma_2 + \gamma_1 \right)\]

\[= \mathbb{P}\left( \varphi_{k+1}^i(\theta) - \mathbb{E}\left[ \varphi_{k+1}^i(\theta) \right] \geq -\frac{k}{2}\gamma_2 + \gamma_1 - \mathbb{E}\left[ \varphi_{k+1}^i(\theta) \right] \right)\]

\[= \mathbb{P}\left( \varphi_{k+1}^i(\theta) - \mathbb{E}\left[ \varphi_{k+1}^i(\theta) \right] \geq \frac{k}{2}\gamma_2 \right),\]

where the last equality follows from Lemma 4.

We now view \( \varphi_{k+1}^i(\theta) \) a function of the random vectors \( s_1, \ldots, s_k, s_{k+1} \) (see Eq. (7)), where \( s_t = (s_1^t, \ldots, s_n^t) \in S \) for all \( t \). Thus, for all \( t \) with \( 1 \leq t \leq k \), we have

\[
\max_{s_t \in S} \varphi_{k+1}^i(\theta) - \min_{s_t \in S} \varphi_{k+1}^i(\theta) = \max_{s_t \in S} \sum_{j=1}^n [A_{k:t}]_{ij} \mathcal{L}_j^0 - \min_{s_t \in S} \sum_{j=1}^n [A_{k:t}]_{ij} \mathcal{L}_j^0
\]

\[
= \max_{s_t \in S} \sum_{j=1}^n [A_{k:t}]_{ij} \log \frac{l_j(s_t^j|\theta)}{l_j(s_t^j|\theta^*)} - \min_{s_t \in S} \sum_{j=1}^n [A_{k:t}]_{ij} \log \frac{l_j(s_t^j|\theta)}{l_j(s_t^j|\theta^*)}
\]

\[
\leq \log \frac{1}{\alpha} + \log \frac{1}{\alpha}
\]

\[
= 2 \log \frac{1}{\alpha}.
\]

Similarly, from Eq. (7) we can see that

\[
\max_{s_{k+1} \in S} \varphi_{k+1}^i(\theta) - \min_{s_{k+1} \in S} \varphi_{k+1}^i(\theta) \leq 2 \log \frac{1}{\alpha}.
\]

It follows that \( \varphi_{k+1}^i(\theta) \) has bounded variations and by McDiarmid’s inequality \( \Phi \) we obtain the following concentration inequality,

\[
\mathbb{P}\left( \varphi_{k+1}^i(\theta) - \mathbb{E}\left[ \varphi_{k+1}^i(\theta) \right] \geq \frac{k}{2}\gamma_2 \right) \leq \exp\left( \frac{-\frac{1}{2}(k\gamma_2)^2}{\sum_{t=1}^{k+1} (2 \log \frac{1}{\alpha})^2} \right)
\]

\[
= \exp\left( \frac{-\frac{1}{8}(k\gamma_2)^2}{(k+1) (\log \frac{1}{\alpha})^2} \right)
\]

\[
\leq \exp\left( \frac{-\frac{1}{8}(k-1)\gamma_2^2}{(\log \alpha)^2} \right)
\]

Therefore, for a given confidence level \( \rho \), in order to have \( \mathbb{P}\left( \mu_k^i(\theta) \geq \exp\left( -\frac{1}{2}k\gamma_2 + \gamma_1 \right) \right) \leq \rho \) we require that

\[
k \geq \frac{8 (\log (\alpha))^2 \log \frac{1}{\rho}}{\gamma_2^2} + 1
\]
V. Simulation Results

In this section we will show the simulation results for a group of agents connected over the time-varying directed graph depicted in Figure 1.

with the weighting matrices,

\[
A_{2k} = \begin{bmatrix}
0.2 & 0.8 & 0 & 0 & 0 & 0 \\
The rest of the matrix components are as follows:
0.7 & 0.1 & 0.1 & 0 & 0 & 0.1 \\
0 & 0.1 & 0.1 & 0.8 & 0 & 0 \\
0 & 0 & 0.8 & 0.1 & 0.1 & 0 \\
0 & 0 & 0.1 & 0.1 & 0.8 & 0 \\
0 & 0.1 & 0 & 0.8 & 0.1 & .
\end{bmatrix}
\] (10)

\[
A_{2k+1} = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.8 & 0.1 & 0 & 0 & 0.1 \\
0 & 0.1 & 0.1 & 0.8 & 0 & 0 \\
0 & 0 & 0.8 & 0.1 & 0.1 & 0 \\
0 & 0 & 0 & 0.1 & 0.1 & 0.8 \\
0 & 0.1 & 0 & 0.8 & 0.1 & .
\end{bmatrix}
\] (11)

Note that the graph is designed such that the edge connecting agent 1 and agent 2 is switching on and off at each time step. Agents 2-6 connecting edges are changing at each time step as well.

Every agent \(i\) receives information from a binary random variable \(S^i_k : \Omega \rightarrow \{0, 1\}\) with probability distribution \(f^i(0) = 0.1\) and \(f^i(1) = 0.9\) for all \(i\)'s. Each agent updates its beliefs according to
Eq. (1). Moreover, every agent has two possible models $\theta_1$ and $\theta_2$. Nevertheless, agents 2 to 6 have uniformly distributed observationally equivalent hypothesis for both $\theta_1$ and $\theta_2$, that is, they are not able to differentiate between the hypothesis individually. Thus $l_i(s|\theta) = 0.5$ for $i = \{2, \ldots, 6\}$, $s = \{0, 1\}$ and $\theta = \{\theta_1, \theta_2\}$. On the other hand, agent 1 has two hypothesis with the following likelihood functions: $l_1(0|\theta_1) = 0.2$ and $l_1(1|\theta_1) = 0.8$ for hypothesis $\theta_1$; and $l_1(0|\theta_2) = 0.9$ and $l_1(1|\theta_2) = 0.1$ for hypothesis $\theta_2$. This indicates that the likelihood functions have been selected such that hypothesis corresponding to $\theta_1$ will describe the data best.

Figure 2 and Figure 3 show the empirical mean over 5000 Monte Carlo simulations of the beliefs of agents 1, 2 and 3 and agents 1, 4, 5 and 6 respectively. The beliefs of agents 1 has been included in Figure 3 for comparison purposes. Results show that agent 1 is the fastest learning agent, since is the one with the correct model. Nevertheless, all other agents are converging to the correct parameter model as well, even if they do not have differentiable models.

VI. CONCLUSIONS AND FUTURE WORK

We have studied the convergence and the rate of convergence for a distributed non-Bayesian learning system. We have shown almost sure consistency and have provided bounds on the global exponential rate of convergence. The novelty of our results is in the establishment of convergence rate estimates that are non-asymptotic, geometric, and explicit, in the sense that the bounds capture the quantities characterizing the graph sequence properties as well as the agent learning capabilities.
Our work suggests a number of open questions. It is natural to attempt to extend the results here to continuous spaces, for example. Another problem is to explore the situation when the distribution of observations is time-varying; ideas from social sampling can also be incorporated in this framework [32]. Moreover, the possibility of corrupted measurements or conflicting models between the agents are also of interest, especially in the setting of social networks.

REFERENCES

[1] R. J. Aumann, “Agreeing to disagree,” The annals of statistics, pp. 1236–1239, 1976.
[2] V. Borkar and P. P. Varaiya, “Asymptotic agreement in distributed estimation,” IEEE Transactions on Automatic Control, vol. 27, no. 3, pp. 650–655, 1982.
[3] J. N. Tsitsiklis and M. Athans, “Convergence and asymptotic agreement in distributed decision problems,” IEEE Transactions on Automatic Control, vol. 29, no. 1, pp. 42–50, 1984.
[4] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, “Bayesian learning in social networks,” The Review of Economic Studies, vol. 78, no. 4, pp. 1201–1236, 2011.
[5] M. Mueller-Frank, “A general framework for rational learning in social networks,” Theoretical Economics, vol. 8, no. 1, pp. 1–40, 2013.
[6] D. Gale and S. Kariv, “Bayesian learning in social networks,” Games and Economic Behavior, vol. 45, no. 2, pp. 329–346, 2003.
[7] L. G. Epstein, J. Noor, and A. Sandroni, “Non-bayesian learning,” The BE Journal of Theoretical Economics, vol. 10, no. 1, 2010.
[8] A. Jadababaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, “Non-bayesian social learning,” Games and Economic Behavior, vol. 76, no. 1, pp. 210–225, 2012.
[9] S. Bandyopadhyay and S.-J. Chung, “Distributed estimation using bayesian consensus filtering,” in American Control Conference (ACC), June 2014, pp. 634–641.
[10] A. Lalitha, A. Sarwate, and T. Javidi, “Social learning and distributed hypothesis testing,” in IEEE International Symposium on Information Theory (ISIT), 2014, pp. 551–555.
[11] Q. Liu, A. Fang, L. Wang, and X. Wang, “Social learning with time-varying weights,” Journal of Systems Science and Complexity, vol. 27, no. 3, pp. 581–593, 2014.
[12] L. Qipeng, F. Aili, W. Lin, and W. Xiaofan, “Non-bayesian learning in social networks with time-varying weights,” in 30th Chinese Control Conference (CCC), 2011, pp. 4768–4771.
[13] Q. Liu and X. Wang, “Social learning in networks with time-varying topologies,” Asian Journal of Control, 2012.
[14] S. Shahrampour and A. Jadbabaie, “Exponentially fast parameter estimation in networks using distributed dual averaging,” in IEEE 52nd Annual Conference on Decision and Control (CDC), 2013, pp. 6196–6201.
[15] D. P. Bertsekas and J. N. Tsitsiklis, Parallel and distributed computation: numerical methods. Prentice-Hall, Inc., 1989.
[16] L. Moreau, “Stability of multiagent systems with time-dependent communication links,” IEEE Transactions on Automatic Control, vol. 50, no. 2, pp. 169–182, 2005.
[17] A. Jadbabaie, J. Lin, and A. S. Morse, “Coordination of groups of mobile autonomous agents using nearest neighbor rules,” IEEE Transactions on Automatic Control, vol. 48, no. 6, pp. 988–1001, 2003.
[18] C. Genest, J. V. Zidek et al., “Combining probability distributions: A critique and an annotated bibliography,” Statistical Science, vol. 1, no. 1, pp. 114–135, 1986.
[19] A. Zellner, “Optimal information processing and bayes’s theorem,” The American Statistician, vol. 42, no. 4, pp. 278–280, 1988.
[20] S. G. Walker, “Bayesian inference via a minimization rule,” Sankhyā: The Indian Journal of Statistics, pp. 542–553, 2006.
[21] T. P. Hill and M. Dall’Aglio, “Bayesian posteriors without bayes’ theorem,” arXiv preprint arXiv:1203.0251, 2012.
[22] G. L. Gilardoni and M. K. Clayton, “On reaching a consensus using degroot’s iterative pooling,” The Annals of Statistics, pp. 391–401, 1993.
[23] R. Cooke, “Statistics in expert resolution: A theory of weights for combining expert opinion,” in Statistics in Science, ser. Boston Studies in the Philosophy of Science, R. Cooke and D. Costantini, Eds. Springer Netherlands, 1990, vol. 122, pp. 41–72.
[24] M. H. DeGroot, “Reaching a consensus,” Journal of the American Statistical Association, vol. 69, no. 345, pp. 118–121, 1974.
[25] A. G. Jayaram, A. Garg, T. S. Jayram, S. Vaithyanathan, and H. Zhu, “Generalized opinion pooling,” in In Proceedings of the 8th Intl. Symp. on Artificial Intelligence and Mathematics, 2004, pp. 79–86.
[26] A. Madansky, “Externally bayesian groups,” The Rand Corporation, Santa Monica, CA, vol. RM-4141-PR, 1964.
[27] J. L. Doob, “Application of the theory of martingales,” Le calcul des probabilités et ses applications, pp. 23–27, 1949.
[28] S. Ghosal, “A review of consistency and convergence of posterior distribution,” in Varanashi Symposium in Bayesian Inference, Banaras Hindu University, 1997.
[29] A. Nedić and A. Olshevsky, “Distributed optimization over time-varying directed graphs,” arXiv preprint arXiv:1303.2289, 2013.
[30] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, “On distributed averaging algorithms and quantization effects,” IEEE Transactions on Automatic Control, vol. 54, no. 11, pp. 2506–2517, 2009.
[31] C. McDiarmid, “On the method of bounded differences,” Surveys in combinatorics, vol. 141, no. 1, pp. 148–188, 1989.
[32] A. D. Sarwate and T. Javidi, “Distributed learning of distributions via social sampling,” *arXiv preprint arXiv:1305.4548*, 2013.