Social Learning and Distributed Hypothesis Testing

Anusha Lalitha
Electrical & Computer Engineering
University of California, San Diego
Email: alalitha@ucsd.edu

Anand Sarwate
Electrical & Computer Engineering
Rutgers University
Email: asarwate@ece.rutgers.edu

Tara Javidi
Electrical & Computer Engineering
University of California, San Diego
Email: tara@ece.ucsd.edu

***Abstract*—(To be considered for an IEEE Jack Keil Wolf ISIT Student Paper Award). This paper considers the problem of distributed hypothesis testing and social learning. Suppose individual nodes in a network receive noisy (private) observations whose distribution is parameterized by one of M parameters (hypotheses). The distributions are known locally at the nodes, but the true parameter/hypothesis is not known. If the local observations are insufficient to recover the underlying parameter (for example, low dimensional measurements of a higher-dimensional parameter), individuals must share and learn from each other in order to accurately infer the true parameter.**

Inspired by recent non-Bayesian social learning algorithms, the updating of opinions of each node is broken down into two steps: a local-Bayesian update, which incorporates the noisy observations and the known parametric distribution of the noise, and a new non-Bayesian update rule which merges the nodes' opinions. It is shown that each node's opinion/belief about any hypothesis whose truth is inconsistent with the overall network-wide observations vanishes to zero exponentially fast. In other words, each node's opinion converges to the true underlying parameter exponentially fast. This new method of merging opinions allows for a concise proof of the convergence and a closed form characterization of rate of convergence. Furthermore, the exponential rate of learning is shown to be both a function of the nodes' collective ability to discriminate among the hypotheses set as well as the social structure of the network.

I. Introduction

How do we learn what television shows are popular or which pastry shop has the best cupcake? Most people do not watch all their available options, nor do they carefully read the Nielsen ratings, but instead infer the answers through a combination of channels: their own experience, their friends' opinion, or social media. Social communication helps individuals form opinions and gain knowledge about a variety of unknown parameters. In other words, people transcend the limitations of their local view by incorporating the "wisdom of the crowd" to construct a "crowd within" themselves, benefitting from the richness and diversity of others' experience. This paper proposes and analyzes a very simple model of social learning in the context of distributed hypothesis testing.

Learning in a distributed setting is more than a phenomenon of social networks; it is also an engineering challenge for networked system designers. For instance, in today's data networks, many applications need estimates of certain parameters: file-sharing systems need to know the distribution of (unique) documents shared by their users, internet-scale information retrieval systems need to deduce the criticality of various data items, and monitoring networks need to compute aggregates in a duplicate-insensitive manner. Finding a scalable, efficient, and accurate method of computing such metrics (e.g. number of documents in the network, sizes of database relations, distributions of data values) is of critical value to a broad set of network applications.

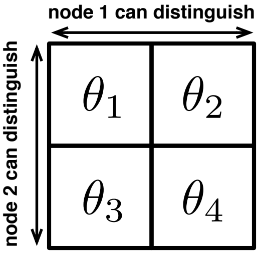

Fig. 1. Example of a parameter space in which no node can identify the true parameter. There are 4 parameters, $\{\theta_1, \theta_2, \theta_3, \theta_4\}$, and 2 nodes. The node 1 has $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3)$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4)$, and the node 2 has $f_2(\cdot;\theta_1) = f_2(\cdot;\theta_2)$ and $f_2(\cdot;\theta_3) = f_2(\cdot;\theta_4)$.

We study a model in which a network of individuals can sample local observations governed by an unknown global parameter θ^*. The observations are given by conditional distribution $\{f_i(\cdot;\theta)\}$ (or local observation channels, from a communication perspective). When these local channels are not sufficient to recover the underlying parameter locally, individuals must share and learn from each other in order to accurately estimate the parameter. Even though each individual cannot identify the parameter through local observations alone, the parameter may be collectively identifiable. A simple two-node example is illustrated in Figure 1 – one node can only identify the column in which the parameter lies, and the other can only identify the row.

Jadbabaie et. al. [1] have recently shown that in such a setting, a combination of local Bayesian updating combined with a linear consensus strategy on the beliefs [2] can lead to all nodes in the network identifying the true parameter. Moreover, they show an exponential rate of convergence of the estimate to the true parameter. In this paper we propose a different strategy based on a reweighted local averaging of the *log beliefs* of nodes. We show that the rate of convergence of this strategy is related to the sum of weighted Kullback-Leibler divergences between likelihood of the true parameter and the likelihood of any other parameter. The sum is over the nodes in the network, and the weights are a function of a weight or influence vector associated to the algorithm. We also show that the rate of convergence under our newly proposed strategy can be significantly better than the convergence rate of the consensus strategy [1].

Olfati-Saber et. al. [3] studied an algorithm for distributed one-shot hypothesis testing using belief propagation – nodes perform average consensus on the log likelihoods under a single observation per node model. The nodes can achieve a consensus on the product of their local likelihoods – that is, they can compute the likelihood of their joint observation under each hypothesis. By contrast, we consider a dynamic setting in which observations are made as messages and are passed in the network. A side benefit of our approach is that nodes do not need to know each other's likelihood functions, or indeed even the space from which their observations are drawn.

Notation: A random variable is denoted by an upper case letter (e.g. X) and its realization is denoted by a lower case letter (e.g. x). Similarly, a random vector and its realization are denoted by bold face symbols (e.g. $\mathbf{X}$ and $\mathbf{x}$). The i-th element of vector $\mathbf{v}$ is denoted by v_i. The set of integers $\{1, 2, \ldots, n\}$ is denoted by $[n]$. Throughout the paper, for a given set E, we use notation $\mathbb{P}(E)$ to be set of probability distribution on E. Let $\mathsf{Ber}(p)$ denote the Bernoulli distribution with parameter p. Also, the *Kullback–Leibler (KL) divergence* between two probability distributions P_Z and P'_Z is defined as $D(P_Z||P'_Z) := \sum_{z\in\mathcal{Z}} P_Z(z) \log \frac{P_Z(z)}{P'_Z(z)}$ with the convention $0 \log \frac{a}{0} = 0$ and $b \log \frac{b}{0} = \infty$ for $a, b \in [0, 1]$ with $b \neq 0$.

II. Problem Statement

We consider a network with n nodes. In particular, one can define social neighborhood of i as $\mathcal{N}(i) := \{j : W_{ij} > 0\}$, where W denotes the $n \times n$ matrix of fixed and known weights $W_{ij} \in [0, 1]$ representing the social interaction of nodes in the network. the We study a simple model for parametric inference of a global parameter $\theta \in \mathbf{\Theta}$, where Θ is a finite set and θ^* is time-invariant, static, and unknown. At every time instant $t = 1, 2, \ldots$ every node draws a noisy observation $X_i^{(t)} \in \mathcal{X}_i$, where $\mathcal{X}_i$ is defined as the *observation space* of node i. Every node's observation sequence (in time) are conditionally independent and identically distributed (i.i.d.); in other words, each node has a corresponding set of probability distributions $\{f_i(\cdot; \theta) : \theta \in \mathbf{\Theta})\}$, where each $f_i(\cdot; \theta)$ is a distribution on a space $\mathcal{X}_i$. $f_i(\cdot; \theta)$ and describes the conditional distribution of node i's observations given the true parameter is θ.

The true parameter θ^* is fixed, the nodes draw observations, exchange messages, and update estimates. At each time step, the each node draws an observation $X_i^{(t)}$ according to the distribution $f_i(\cdot; \theta^*)$ and then computes a message to send based on its existing estimate and the observation. The nodes then exchange these messages with their neighbors and update their estimate of θ^* based on their previous estimate, observation, and messages from their neighbors. Detailed rules for forming the belief and determining the estimate are discussed in the next subsection.

A. Bayesian updating with log-belief consensus

Here we present the algorithm according to which the beliefs are updated and estimates are obtained. Let each node i start with a estimate in the $\mathbf{q_i^{(0)}} \in \mathbb{P}(\mathbf{\Theta})$, i.e. a probability distribution on $\mathbf{\Theta}$. At each time $t = 1, 2, \ldots$ the following events happen:

1) Each node i draws an observation $X_i^{(t)} \sim f_i(\cdot; \theta^*)$.
2) Each node i forms a Bayesian update of its belief $\mathbf{b_i^{(t)}}$, using the following rule. For each $\theta \in \mathbf{\Theta}$,

$$b_i^{(t)}(\theta) = \frac{f_i\left(X_i^{(t)}; \theta\right) q_i^{(t-1)}(\theta)}{\sum_{\theta'\in\Theta} f_i\left(X_i^{(t)}; \theta'\right) q_i^{(t-1)}(\theta')}. \quad (1)$$

3) Each node i sends the message $\mathbf{Y_i^{(t)}} = \mathbf{b_i^{(t)}}$ to its neighbors.
4) Each node i forms an estimate of θ, $q_i^{(t)}(\theta)$, via linear consensus on the log beliefs of itself and its neighbors as below. For any $\theta \in \mathbf{\Theta}$,

$$q_i^{(t)}(\theta) = \frac{\exp\left(\sum_{j=1}^n W_{ji} \log b_j^{(t)}(\theta)\right)}{\sum_{\theta'\in\Theta} \exp\left(\sum_{j=1}^n W_{ji} \log b_j^{(t)}(\theta')\right)}. \quad (2)$$

B. Mathematical Assumptions

The following assumptions are required to establish our main result in the later section.

Assumption 1. *The matrix W is stochastic and irreducible.*

This assumption ensures that all the nodes are connected to every other node in the network by at least one multi-hop path. In other words, in the long run, every node influences and is influenced by every other node in the network. This assumption is natural because it enables social learning even in the case that only one node in the network can distinguish the true parameter. It is intuitive that the rate of social learning enabled by such a node depends on the social influence of node i, given the matrix W, where social influence is often measured by the left eigenvector of W corresponding to the eigenvalue 1. This is rigorously characterized in Section III and numerically illustrated in Section IV.

The following fact follows from Assumption 2 and will be used in our analysis.

Fact 1 (Section 2.5 in [4])**.** *Let $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ be the left eigenvector of stochastic matrix W associated with the eigenvalue 1. Whenever W is a positive irreducible stochastic matrix, all components of $\mathbf{v}$ are strictly positive.*

Assumption 2. *For $k \in [n]$, $X \in \mathcal{X}_k$, and for any given $\theta_i, \theta_j \in \Theta$ such that $\theta_i \neq \theta_j$, $\frac{f_k(\cdot;\theta_i)}{f_k(\cdot;\theta_j)}$ is bounded,* i.e.*, there exists a positive constant C_{k,θ_i,θ_j} such that,*

$$\sup_{X\in\mathcal{X}_i} \left(\frac{f_k(X; \theta_i)}{f_k(X; \theta_j)}\right) \leq C_{k,\theta_i,\theta_j} \quad (3)$$

This is a purely technical assumption which simplifies our analysis and proof. We will discuss this assumption and relaxing it in Section V.

Assumption 3. *For every pair $\theta \neq \theta^*$, the KL-divergence $D(f_i(\cdot; \theta^*) \| f_i(\cdot; \theta))$ is positive for at least one $i \in [n]$.*

This final assumption guarantees that for each "wrong" hypothesis $\theta \neq \theta^*$, there exists at least one node in the network that can statistically distinguish between θ and θ^*. Note that this does not require the existence of a a single node that can distinguish θ^* from all other hypotheses. See the example in Figure 1.

III. Analysis

In this section provide the main results of the paper.

Theorem 1. *Let θ^* be the true parameter and for every $i \in [n]$, the initial estimate $q_i^{(0)}(\theta^*) > 0$. Then under Assumptions 1-3, each node's estimate converges to the true parameter θ^*. Mathematically, for all $\theta \neq \theta^*$ and for every* *node* $i \in [n]$,

$$\limsup_{t\to\infty} \frac{1}{t} \log q_i^{(t)}(\theta) \leq -\frac{\overline{W}_i}{\underline{v}} K(\theta,\theta^*), \tag{4}$$

where,

$$\underline{v} = \min_i v_i, \tag{5}$$

$$\overline{W}_i = \max_j W_{ji}, \tag{6}$$

and

$$K(\theta,\theta^*) = \sum_{i=1}^{n} v_i D\left(f_i\left(\cdot;\theta^*\right) \| \, f_i\left(\cdot;\theta\right)\right). \tag{7}$$

Proof: We begin by obtaining a recursion in terms of the logarithm of the ratio of the estimates $q_i^{(t)}$ at different values of θ.

$$\begin{aligned}
&\log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} && (8)\\
&\quad= \sum_{j=1}^{n} W_{ji} \log \frac{b_j^{(t)}(\theta)}{b_j^{(t)}(\theta^*)}\\
&\quad= \sum_{j=1}^{n} W_{ji} \left(\log \frac{f_j\left(X_j^{(t)};\theta\right)}{f_j\left(X_j^{(t)};\theta^*\right)} + \log \frac{q_j^{(t-1)}(\theta)}{q_j^{(t-1)}(\theta^*)} \right) && (9)
\end{aligned}$$

where the first and the third equalities follow from (2) and (1), respectively.

Let us define

$$L_{\theta,\theta^*}^{(t)} := \sum_{i=1}^{n} v_i \log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)}. \tag{10}$$

From (9), we have that

$$L_{\theta,\theta^*}^{(t)} = \sum_{j=1}^{n} v_j \log \frac{f_j\left(X_j^{(t)};\theta\right)}{f_j\left(X_j^{(t)};\theta^*\right)} + L_{\theta,\theta^*}^{(t-1)} \tag{11}$$

Let $\mathcal{F}_t$ be the σ-algebra induced by $\left\{X_i^{(\tau)} : i \in [n], \tau \leq t\right\}$. Taking expectation of both sides of equation (11) and using the definition of KL divergence, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} v_i \log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \middle| \mathcal{F}_{t-1} \right]\\
&\quad= -\sum_{i=1}^{n} v_i D\left(f_i\left(\cdot;\theta^*\right) \| \, f_i\left(\cdot;\theta\right)\right) + \sum_{i=1}^{n} v_i \log \frac{q_i^{(t-1)}(\theta)}{q_i^{(t-1)}(\theta^*)}.
\end{aligned}$$

In other words,

$$\mathbb{E}\left[L_{\theta,\theta^*}^{(t)} - L_{\theta,\theta^*}^{(t-1)} \middle| \mathcal{F}_{t-1} \right] = -K(\theta,\theta^*). \tag{12}$$

Assumptions 1 and 3 imply that $K(\theta,\theta^*)$ is a positive constant. This, together with (12) means that for every $\theta \neq \theta^*$, $L_{\theta,\theta^*}^{(t)}$ forms a supermartingale with respect to filtration $\mathcal{F}_t$ with a negative drift equal to $-K(\theta,\theta^*)$. Furthermore,

$$\left| L_{\theta,\theta^*}^{(t)} - L_{\theta,\theta^*}^{(t-1)} \right| < A_{\theta,\theta^*} := \max_k C_{k,\theta,\theta^*}. \tag{13}$$

From Lemma 1 in the appendix, we have that for any $\epsilon > 0$

$$P\left(L_{\theta,\theta^*}^{(t)} \geq -(K(\theta,\theta^*) - \epsilon)t \right) \leq \exp\left(\frac{-\epsilon^2 t}{2A_{\theta,\theta^*}^2} \right). \tag{14}$$

Let $\{\Omega_t\}$ be a sequence of events defined as

$$\Omega_t = \left\{ \omega : \frac{1}{t} L_{\theta,\theta^*}^{(t)}(\omega) \geq -(K(\theta,\theta^*) - \epsilon) \right\}. \tag{15}$$

From (14) we obtain that $\sum_{t=1}^{\infty} P(\Omega_t) < \infty$. This together with Borel-Cantelli Lemma implies that

$$P\left(\limsup_{t\to\infty} \Omega_t \right) = 0. \tag{16}$$

where

$$\limsup_{t\to\infty} \Omega_t = \bigcap_{t=1}^{\infty} \bigcup_{n=t}^{\infty} \{\omega : \frac{1}{n} L_{\theta,\theta^*}^{(n)}(\omega) \geq -(K(\theta,\theta^*) - \epsilon)\}. \tag{17}$$

Rewriting this, we have that for any arbitrary $\epsilon > 0$

$$P\left(\limsup_{t\to\infty} \frac{1}{t} L_{\theta,\theta^*}^{(t)} \geq -(K(\theta,\theta^*) - \epsilon) \right) = 0 \tag{18}$$

In other words,

$$\liminf_{t\to\infty} \frac{1}{t} L_{\theta,\theta^*}^{(t)} \leq \limsup_{t\to\infty} \frac{1}{t} L_{\theta,\theta^*}^{(t)} \leq -K(\theta,\theta^*) \quad a.s. \tag{19}$$

This means that for every $\epsilon' > 0$, there exists t_0 such that for all $t \geq t_0$, $\frac{1}{t} L_{\theta,\theta^*}^{(t)} \leq -K(\theta,\theta^*) + \epsilon'$, *i.e.*

$$\underline{v} \sum_{i=1}^{n} \log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \leq \sum_{i=1}^{n} v_i \log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \leq (-K(\theta,\theta^*) + \epsilon')t. \tag{20}$$

On the other hand, from (9) and Assumption (2), we have

$$\begin{aligned}
\log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} &\leq \sum_{j=1}^{n} W_{ji} \left(\log C_{j,\theta,\theta^*} + \log \frac{q_j^{(t-1)}(\theta)}{q_j^{(t-1)}(\theta^*)} \right)\\
&\leq \sum_{j=1}^{n} W_{ji} \log C_{j,\theta,\theta^*} + \overline{W}_i \log \frac{q_j^{(t-1)}(\theta)}{q_j^{(t-1)}(\theta^*)}\\
&\leq \sum_{j=1}^{n} W_{ji} \log C_{j,\theta,\theta^*} + \frac{\overline{W}_i}{\underline{v}} (-K(\theta,\theta^*) + \epsilon')t.
\end{aligned} \tag{21}$$

where the last inequality follows from (20).

Since $q_i^{(t)}(\theta^*) \leq 1$, we can rewrite the above equation as

$$q_i^{(t)}(\theta) \leq \exp\left(-\frac{\overline{W}_i}{\underline{v}} (K(\theta,\theta^*) + \epsilon')t + \delta(i,\theta,\theta^*) \right) \tag{22}$$

where $\delta(i,\theta,\theta^*) = \sum_{j=1}^{n} W_{ji} \log C_{j,\theta,\theta^*}$ and $\epsilon' > 0$ is any arbitrary positive scalar.

Hence, for every i and $\theta \neq \theta^*$, we have that $q_i^{(t)}(\theta)$ goes to zero exponentially fast, *i.e.*

$$\limsup_{t\to\infty} \frac{1}{t} \log q_i^{(t)}(\theta) \leq -\frac{\overline{W}_i}{\underline{v}} K(\theta,\theta^*) + \epsilon'. \tag{23}$$

Since the choice of ϵ' is arbitrary, we have the assertion of the lemma. ■

IV. Examples

To illustrate our proposed scheme and how the rate of convergence depends on various parameters, we look at a few simple examples. The first example, shown in Figure 1 earlier in the paper, is a two-node network in which neither node can identify the parameter locally. The second example is shown in Figure 2 and is a star network with 3 satellite nodes around a single central node.

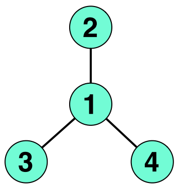

Fig. 2. A star network in which a central node (node 1) is connected to other satellite nodes. The network topology and distribution of nodes' local power to identify the hypotheses affects the rate of convergence.

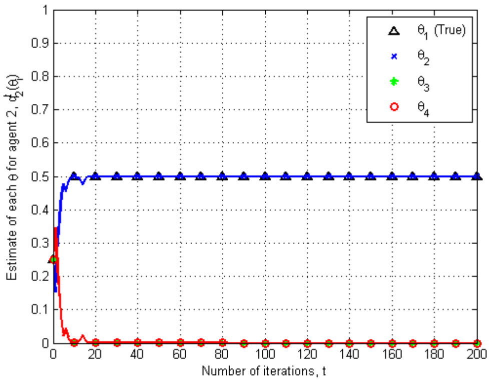

Fig. 3. Estimates of node 2 in the example of Figure 1 vs. iterations when no communication is allowed. Neither node can converge on the true parameter $\theta^* = \theta_1$ without communication.

A. Communication is necessary for convergence

In the two-node network of Figure 1, the parameter space $\boldsymbol{\Theta} = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. Node 1 has $f_1(\cdot; \theta_1) = f_1(\cdot; \theta_3)$ and $f_1(\cdot; \theta_2) = f_1(\cdot; \theta_4)$, and the node 2 has $f_2(\cdot; \theta_1) = f_2(\cdot; \theta_2)$ and $f_2(\cdot; \theta_3) = f_2(\cdot; \theta_4)$. Thus node 1 can identify the column containing θ^*, and node 2 the row.

For this example, we will assume that for node 1, $f_1(\cdot; \theta_1) = f_1(\cdot; \theta_3) \sim \text{Ber}(\frac{3}{4})$ and $f_1(\cdot; \theta_2) = f_1(\cdot; \theta_4) \sim \text{Ber}(\frac{1}{3})$, and for node 2, $f_2(\cdot; \theta_1) = f_2(\cdot; \theta_2) \sim \text{Ber}(\frac{2}{3})$ and $f_2(\cdot; \theta_3) = f_2(\cdot; \theta_4) \sim \text{Ber}(\frac{1}{4})$. Let the true parameter be $\theta^* = \theta_1$, and assume that the nodes all begin with a uniform prior on the parameter space.

First consider the case where there is no communication, i.e. the matrix $W = I$. Because neither node can identify the true parameter on the basis of their own observations, their estimates do not converge to the true parameter θ^*. For example, for node 1 the estimates converge to $(0.5, 0, 0.5, 0)$ and for node 2 they converge to $(0.5, 0.5, 0, 0)$. This is illustrated in Figure 3, which shows the non-convergence.

Suppose now that we set the weight matrix to $W = \begin{pmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{pmatrix}$, so the nodes do communicate after each iteration. Figure 4 shows the rapid convergence of the node estimates to the true parameter as a function of the number of iterations. It is clear that communicating helps significantly in this setting where the parameter is not identifiable locally.

Figure 5 compares the empirical convergence rate of our method with that of Jadbabaie et al. [1], which uses linear consensus on the updated beliefs, as opposed to the reweighed average of log-beliefs that we propose here. In this simple two-node example we see that the our new method converges significantly faster than the previous algorithm.

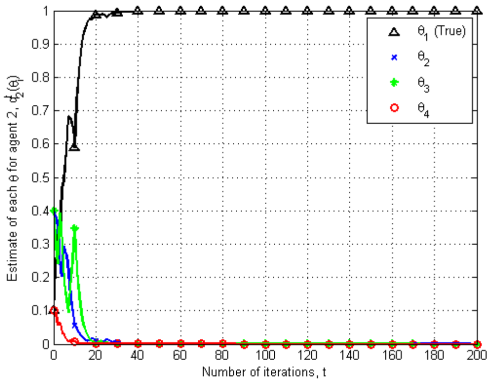

Fig. 4. Estimates of node 2 in the example of Figure 1 vs. iterations when communication is allowed. Node 2 very rapidly distinguishes between θ_1 and θ_2 using the information communicated by node 1.

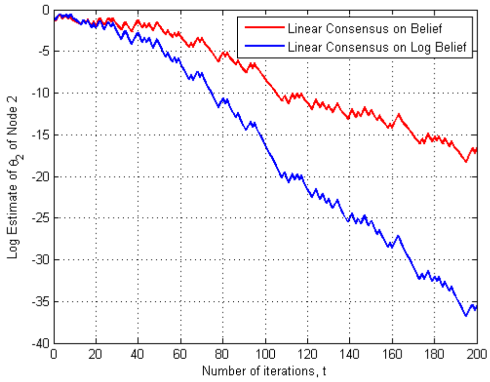

Fig. 5. A comparison of the non-Bayesian learning algorithm of Jadbabaie et al. [1] using linear consensus on beliefs (upper line) and the algorithm of this paper, which uses linear consensus on log-beliefs (lower line). The estimates converge must faster using log beliefs.

B. Topology and rate of convergence

We now turn to the star network in in Figure 2 to address the interplay between the network topology and the rate of convergence. From the analysis, the rate of convergence depends on two factors: the KL-divergences at each node, and the weighting factor from the eigenvector of the weight matrix W. We consider a binary hypothesis test with $\boldsymbol{\Theta} = \{\theta_1, \theta_2\}$, and consider the weight matrix

$$W = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{pmatrix}.$$

That is, the satellite nodes average their estimate with the central node's estimate, and the central node averages all nodes. The corresponding left eigenvector is $v = [\frac{2}{\sqrt{7}} \; \frac{1}{\sqrt{7}} \; \frac{1}{\sqrt{7}} \; \frac{1}{\sqrt{7}}]$, assigning a weight (social influence) to the first node that is twice as much as that of the other nodes.

We consider the scenario where only one node ("the informed node") can identify the parameter. We expect that if that central node is the informed node, then the rate of convergence will be much faster, since the the contribution to $K(\theta, \theta^*)$ will be zero for the nodes that cannot identify the parameter. In particular, let us consider the case where the informed node has observations $f_i(\cdot; \theta_1) \sim \text{Ber}(\frac{3}{4})$ and $f_i(\cdot; \theta_2) \sim \text{Ber}(\frac{1}{3})$, whereas uninformed nodes have $f_j(\cdot; \theta_1) = f_2(j; \theta_2) \sim \text{Ber}(\frac{1}{2})$.

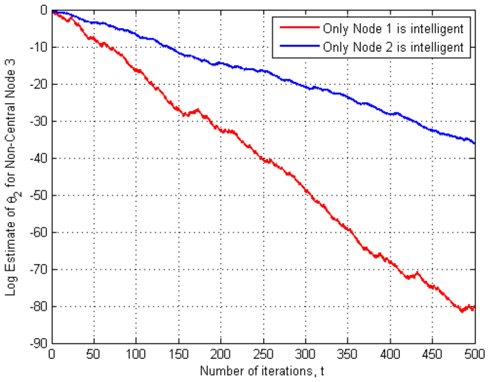

Fig. 6. The rate of convergence for two assignments in the star network of Figure 2. In both cases, all nodes but one have useless observations, where $f_i(\cdot;\theta)$ is the same for all θ. The lower line shows the case where the center node (node 1) can identify true parameter. The upper line shows the case where a satellite node can identify the true parameter.

Figure 6 shows the convergence of estimates for two scenarios. In the first, the informed node is node 1, the central node. In the second, the informed node is node 2, a satellite node. The rate of convergence is clearly better when the central node is informed. This makes sense because the net impact of informative observations of the informed node can more easily reach the other nodes in the network.

For a given influence or weight matrix W, the corresponding left eigenvector gives a measure of the influence of different nodes in the network. Our analysis supports the natural conclusion that faster convergence comes from assigning more influence to the informed nodes in the network. Conversely, the rate of convergence appears to be slower when the informed nodes are not central to the network. However, our theoretical results say that the network will still learn the parameter in this case, albeit at a slower rate.

V. Discussion and Future Work

In this paper we examined a network in which nodes take observations whose distribution depends on a global parameter and communicate in order to collectively estimate the parameter. We demonstrate a protocol which alternates local Bayesian updating with a weighted-averaging step and demonstrated exponential converge of the estimates to the true parameter. Our approach is similar in structure to a previously proposed scheme [1], but the second update step is significantly different.

Our algorithm converges under mild assumptions on the observation model. The irreducibility of W in Assumption 1 is clearly necessary. To see this, consider a matrix F which is not irreducible. If there exists a pair (i, j) such that there is no path from i to j, then it is possible to assign local observation models such that only node i can distinguish the true parameter and all other nodes cannot distinguish any parameter. In this setting, node j cannot converge to the true parameter. Assumption 3 says that for each incorrect hypothesis $\theta \neq \theta^*$, there must exist one node which distinguishes θ from θ^*. If not, then no node can distinguish θ from θ^*, preventing convergence. In particular, Assumptions 1 and 3 are necessary to ensure that $K(\theta, \theta^*) > 0$.

Assumption 2 is purely technical. It guarantees that the likelihood ratio at each node between any two hypotheses is bounded from above. This assumption is commonly used in hypothesis testing and allows us to use Azuma's inequality in the proof of Theorem 1. We believe the assumption can be relaxed significantly; extending our work to a less stringent technical assumption (similar to that established by Naghshvar and Javidi [5]) is an area of future work.

An interesting question for future investigation is to find the optimal exponent for estimation in this setting. For the scheme proposed here, the optimal rate of convergence is achieved when the weighted sum of the KL divergences, *i.e.* $K(\theta, \theta^*)$, is largest. Proving a lower bound on the rate of convergence in terms of the local divergences may shed some light on the fundamental limits of hypothesis testing from distributed observations using local communication.

Appendix

Lemma 1. *Assume that the sequence* $\{\xi(t)\}$, $t = 0, 1, 2, \ldots$ *forms a supermartingale with respect to a filtration* $\{\mathcal{F}(t)\}$. *Furthermore, assume there exist positive constants* K_1 *and* K_2 *such that*

$$\mathbb{E}[\xi(t+1)|\mathcal{F}(t)] \leq \xi(t) - K_1 \tag{24}$$

$$|\xi(t+1) - \xi(t)| \leq K_2. \tag{25}$$

Then, for any $\epsilon > 0$,

$$P\left\{\xi(t) \geq \xi(0) - K_1 t + \epsilon t\right\} \leq \exp \frac{-\epsilon^2 t}{2{K_2}^2}.$$

Proof. Define $V(t) = \xi(t) + K_1 t$. From (24), we have

$$\mathbb{E}\left[\xi(t) + K_1 t - \xi(t-1) + K_1(t-1)\right| \mathcal{F}_{t-1}] \leq 0. \tag{26}$$

Hence, $V(t)$ is also a supermartingale. We also have that $|V(t) - V(t-1)|$ is bounded. From Azuma's inequality [6], we have that for all positive δ,

$$P\left(\xi(t) - \xi(0) + K_1 t \geq \delta\right) \leq \exp\left(\frac{-\delta^2}{2\sum_{i=1}^{t} {K_2}^2}\right). \tag{27}$$

Setting $\delta = \epsilon t$, we have the assertion of the lemma. ∎

References

[1] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[2] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974. [Online]. Available: http://www.jstor.org/stable/2285509

[3] R. Olfati-Saber, E. Franco, E. Frazzoli, and J. S. Shamma, "Belief consensus and distributed hypothesis testing in sensor networks," in *Workshop on Network Embedded Sensing and Control*, Notre Dame University, South Bend, IN, October 2005.

[4] C. J. S. Paul G. Hoel, Sidney C. Port, *Introduction to Stochastic Processes*. Waveland Press, 1972.

[5] M. Naghshvar and T. Javidi, "Active sequential hypothesis testing," *The Annals of Statistics*, no. 6, pp. 2703–2738, December 2013.

[6] D. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge, UK: Cambridge University Press, 2009.