Social Learning and Distributed Hypothesis Testing

Anusha Lalitha, Tara Javidi, *Senior Member, IEEE,* and Anand Sarwate, *Member, IEEE*

***Abstract*—This paper considers a problem of distributed hypothesis testing over a network. Individual nodes in a network receive noisy local (private) observations whose distribution is parameterized by a discrete parameter (hypothesis). The marginals of the joint observation distribution conditioned on each hypothesis are known locally at the nodes, but the true parameter/hypothesis is not known. An update rule is analyzed in which nodes first perform a Bayesian update of their belief (distribution estimate) of each hypothesis based on their local observations, communicate these updates to their neighbors, and then perform a "non-Bayesian" linear consensus using the log-beliefs of their neighbors. Under mild assumptions, we show that the belief of any node on a wrong hypothesis converges to zero exponentially fast. We characterize the exponential rate of learning, which we call the network divergence, in terms of the nodes' influence of the network and the divergences between the observations' distributions. For a broad class of observation statistics which includes distributions with unbounded support such as Gaussian mixtures, we show that rate of rejection of wrong hypothesis satisfies a large deviation principle i.e., the probability of sample paths on which the rate of rejection of wrong hypothesis deviates from the mean rate vanishes exponentially fast and we characterize the rate function in terms of the nodes' influence of the network and the local observation models.**

I. Introduction

Learning in distributed settings 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 scalable, efficient, and accurate methods for computing such metrics (e.g. number of documents in the network, sizes of database relations, distributions of data values) is of critical value in a wide array of network applications.

We consider a network of nodes that sample local observations (over time) governed by an unknown true hypothesis θ^* taking values in a finite discrete set Θ. We model the i-th node's distribution (or local channel, or likelihood function) of the observations conditioned on the true hypothesis by $f_i(\cdot;\theta^*)$ from a collection $\{f_i(\cdot;\theta) : \theta \in \Theta\}$. Nodes neither have access to each others' observations nor the joint distribution of observations across all nodes in the network. Every node in the network aims to learn the unknown true hypothesis θ^*. A simple two-node example is illustrated in Figure 1 – one node can only learn the column in which the true hypothesis lies, and the other can only learn the row. In this example, the local observations of a given node are not sufficient to recover the underlying hypothesis in isolation. In this paper we study a learning rule that enables the nodes to learn the unknown true hypothesis based on message passing between one hop neighbors (local communication) in the network. In particular, each node performs a local Bayesian update and send its belief vectors (message) to its neighbors. After receiving the messages from the neighbors each node performs a consensus averaging on a reweighting of the *log beliefs*. Our result shows that under our learning rule each node can reject the wrong hypothesis exponentially fast.

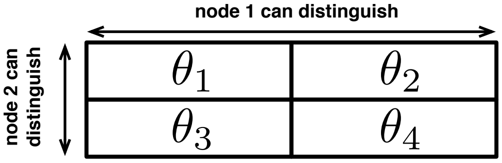

Fig. 1. Example of a parameter space in which no node can identify the true parameter. There are 4 parameters, $\{\theta_1, \theta_2, \theta_3, \theta_4\}$, and 2 nodes. The node 1 has $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3)$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4)$, and the node 2 has $f_2(\cdot;\theta_1)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

We show that the rate of rejection of wrong hypothesis is the weighted sum of Kullback-Leibler (KL) divergences between likelihood function of the true parameter and the likelihood function of the wrong hypothesis, where the sum is over the nodes in the network and the weights are the nodes' influences as dictated by the learning rule. Furthermore, we show that the probability of sample paths on which the rate of rejection deviates from the mean rate vanishes exponentially fast. For any strongly connected network and bounded ratios of log-likelihood functions, we obtain a lower bound on this exponential rate. Furthermore, for any aperiodic network we characterize the exact exponent with which probability of sample paths on which the rate of rejection deviates from the mean rate vanishes (i.e., obtain a large deviation principle) for a broader class of observation statistics which includes distributions with unbounded support such as Gaussian mixtures and Gamma distribution. The large deviation rate function is shown to be a function of observation model and the nodes' influences on the network as dictated by the learning rule.

Outline of the Paper. The rest of the paper is organized as

Some results in this paper were presented in part at conferences [1]–[3].

A. Lalitha and T. Javidi are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA. (e-mail: alalitha@ucsd.edu; tjavidi@ucsd.edu).

A. Sarwate is with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, 94 Brett Road, Piscataway, NJ 08854 , USA. (e-mail: asarwate@ece.rutgers.edu).

follows. We provide the model in Section II which defines the nodes' observation model and network. This section also contains the learning rule and assumptions on model. We then provide results on rate of convergence and their proofs in Section III. We apply our learning rule to various examples in Section IV and discuss some practical issues in Section IV-C. We conclude with a summary in Section V.

A. Related Work

The literature on distributed learning, estimation and detection can divided into two broad sets. One set deals with the fusion of information observed by a group nodes at a fusion center where the communication links (between the nodes and fusion center) are either rate limited [4]–[12] or subject to channel imperfections such as fading and packet drops [13]–[15]. Our work belongs to the second set, which models the communication network as a directed graph whose vertices/nodes are agents and an edge from node i to j indicates that i may send a message to j with perfect fidelity (the link is a noiseless channel of infinite capacity). These "protocol" models study how message passing in a network can be used to achieve a pre-specified computational task such as distributed learning [16], [17], general function evaluation [18],or stochastic approximations [19]. Message passing protocols may be synchronous or asynchronous (such as the "gossip" model [20]–[24]). This graphical model of the communication, instead of assuming a detailed physical-layer formalization, implicitly assumes a PHY/MAC-layer abstraction where sufficiently high data rates are available to send the belief vectors with desired precision when nodes are within each others' communication range. A missing edge indicates the corresponding link has zero capacity.

Due to the large body of work in distributed detection, estimation and merging of opinions, we provide a long yet detailed summary of all the related works and their relation to our setup. Readers familiar with these works can skip to Section II without loss of continuity.

Several works [25]–[29] consider an update rule which uses local Bayesian updating combined with a linear consensus strategy on the beliefs [30] that enables all nodes in the network identify the true hypothesis. Jadbabaie et al. [25] characterize the "learning rate" of the algorithm in terms of the total variational error across the network and provide an almost sure upper bound on this quantity in terms of the KL-divergences and influence vector of agents. In Corollary 2 we analytically show that the proposed learning rule in this paper provides a strict improvement over linear consensus strategies [25]. Simultaneous and independent works by Shahrampour et al. [31] and Nedić et al. [32] consider a similar learning rule (with a change of order in the update steps). They obtain similar convergence and concentration results under the assumption of bounded ratios of likelihood functions. Nedić et al. [32] analyze the learning rule for time-varying graphs. Theorem 3 strengthens these results for static networks by providing a large deviation analysis for a broader class of likelihood functions which includes Gaussian mixtures.

Rad and Tahbaz-Salehi [28] study distributed parameter estimation using a Bayesian update rule and average consensus on the log-likelihoods similar to (2)–(3). They show that the maximum of each node's belief distribution converges in probability to the true parameter under certain analytic assumptions (such as log-concavity) on the likelihood functions of the observations. Our results show almost sure convergence and concentration of the nodes' beliefs when the parameter space is discrete and the log-likelihood function is concave. Kar et al. in [33] consider the problem of distributed estimation of an unknown underlying parameter where the nodes make noisy observations that are non-linear functions of an unknown global parameter. They form local estimates using a quantized message-passing scheme over randomly-failing communication links, and show the local estimators are consistent and asymptotically normal. Note that for any general likelihood model and static strongly connected network, our Theorem 1 strengthens the results of distributed estimation (where the error vanishes inversely with the square root of total number of observations) by showing exponentially fast convergence of the beliefs. Furthermore, Theorem 2 and 3 strengthen this by characterizing the rate of convergence.

Similar non-Bayesian update rules have been in the context of one-shot merging of opinions [29] and beliefs in [34] and [35]. Olfati-Saber et al. [29] studied an algorithm for distributed one-shot hypothesis testing using belief propagation (BP), where nodes perform average consensus on the log-likelihoods under a single observation per node. The nodes can achieve a consensus on the product of their local likelihoods. A 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. Saligrama et al. [34] and Alanyali et al. [35], consider a similar setup of belief propagation (after observing single event) for the problem of distributed identification of the MAP estimate (which coincides with the true hypothesis for sufficiently large number of observations) for certain balanced graphs. Each node passes messages which are composed by taking a product of the recent messages then taking a weighted average over all hypotheses. Alanyali et al. [35] propose modified BP algorithms that achieve MAP consensus for arbitrary graphs. Though the structure of the message composition of the BP algorithm based message passing is similar to our proposed learning rule, we consider a dynamic setting in which observations are made infinitely often. Our rule incorporates new observation every time a node updates its belief to learn the true hypothesis. Other works study collective MAP estimation when nodes communicate discrete decisions based on Bayesian updates [36], [37] Harel et el. in [36] study a two-node model where agents exchange decisions rather than beliefs and show that unidirectional transmission increases the speed of convergence over bidirectional exchange of local decisions. Mueller-Frank [37] generalized this result to a setting in which nodes similarly exchange local strategies and local actions to make inferences.

Several recently-proposed models study distributed sequential binary hypothesis testing detecting between different means with Gaussian [38] and non-Gaussian observation models [39]. Jakovetic et al. [39] consider a distributed hypothesis test for i.i.d observations over time and across nodes where nodes exchange weighted sum of a local estimate from previous time instant and ratio of likelihood functions of the latest local observation with the neighbors. When the network is densely connected (for instance, a doubly stochastic weight matrix), after sufficiently long time nodes gather all the observations throughout network. By appropriately choosing a local threshold for local Neyman-Pearson test, they show that the performance of centralized Neyman-Pearson test can be achieved locally. In contrast, our M-ary learning rule applies for observations that are correlated across nodes and exchanges more compact messages i.e., the beliefs (two finite precision real values for binary hypothesis test) as opposed to messages composed of the raw observations (in the case of $\mathbb{R}^d$ Gaussian observations with $d \gg 2$, d finite precision real values for binary hypothesis test). Sahu and Kar [38] consider a variant of this test for the special case of Gaussians with shifted mean and show that it minimizes the expected stopping times under each hypothesis for given detection errors.

II. The Model

Notation: We use boldface for vectors and denote the i-th element of vector $\mathbf{v}$ by v_i. We let $[n]$ denote $\{1, 2, \ldots, n\}$, $\mathcal{P}(A)$ the set of all probability distributions on a set A, $|A|$ denotes the number of elements in set A, $\mathsf{Ber}(p)$ the Bernoulli distribution with parameter p, and $D(P_Z||P_Z')$ the Kullback–Leibler (KL) divergence between two probability distributions $P_Z, P_Z' \in \mathcal{P}(\mathcal{Z})$. Time is discrete and denoted by $t \in \{0, 1, 2, \ldots\}$. If $a \in A$, then $\mathbf{1}_a(.) \in \mathcal{P}(A)$ denotes the probability distribution which assigns probability one to a and zero probability to the rest of the elements in A. For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, let $\mathbf{x} \leq \mathbf{y}$ denote $x_i \leq y_i$ for each i-th element of vector $\mathbf{x}$ and $\mathbf{y}$ and let $\langle \mathbf{x}, \mathbf{y} \rangle$ denote $\sum_{i=1}^d x_i y_i$. Let $\mathbf{1}$ denote the vector of where each element is 1. For any subset $F \subset \mathbb{R}^{M-1}$, let F^o be the interior of F and $\bar{F}$ the closure. For $\epsilon > 0$ let $F_{\epsilon^+} = \{\mathbf{x} + \delta\mathbf{1}, \forall\, 0 < \delta \leq \epsilon \text{ and } \mathbf{x} \in F\}$, $F_{\epsilon^-} = \{\mathbf{x} - \delta\mathbf{1}, \forall\, 0 < \delta \leq \epsilon \text{ and } \mathbf{x} \in F\}$.

A. Nodes' Observation Model

Consider a group of n individual nodes. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ denote a finite set of M parameters which we call *hypotheses*: each θ_i denotes a hypothesis. At each time instant t, every node $i \in [n]$ makes an observation $X_i^{(t)} \in \mathcal{X}_i$, where $\mathcal{X}_i$ denotes the observation space of node i. The joint observation profile at any time t across the network, $\{X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)}\}$, is denoted by $\mathbf{X}^{(\mathbf{t})} \in \boldsymbol{\mathcal{X}}$, where $\boldsymbol{\mathcal{X}} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_n$. The joint likelihood function for all $X \in \mathcal{X}$ given θ_k is the true hypothesis is denoted as $f\left(X; \theta_k\right)$. We assume that the observations are statistically governed by a fixed global "true hypothesis" $\theta^* \in \Theta$ which is unknown to the nodes. Without loss of generality we assume that $\theta^* = \theta_M$. Furthermore, we assume that no node in network knows the joint likelihood functions $\{f\left(\cdot; \theta_k\right)\}_{k=1}^M$ but every node $i \in [n]$ knows the *local likelihood functions* $\{f_i\left(\cdot; \theta_k\right)\}_{k=1}^M$, where $f_i\left(\cdot; \theta_k\right)$ denotes the i-th marginal of $f\left(\cdot; \theta_k\right)$. Each node's observation sequence (in time) is conditionally independent and identically distributed (i.i.d) but the observations might be correlated across the nodes at any given time.

In this setting, nodes attempt to learn the "true hypothesis" θ_M using their knowledge of $\{f_i\left(\cdot; \theta_k\right)\}_{k=1}^M$. In isolation, if $f_i\left(\cdot; \theta_k\right) \neq f_i\left(\cdot; \theta_M\right)$ for some $k \in [M-1]$, node i can rule out hypothesis θ_k in favor of θ_M exponentially fast with an exponent which is equal to $D\left(f_i\left(\cdot; \theta_M\right) \| f_i\left(\cdot; \theta_k\right)\right)$ [40, Section 11.7]. Hence, for a given node the KL-divergence between the distribution of the observations conditioned over the hypotheses is a useful measure of the distinguishability of the hypotheses. Now, define

$$\begin{aligned}\bar{\Theta}_i &= \{k \in [M] : f_i\left(\cdot; \theta_k\right) = f_i\left(\cdot; \theta_M\right)\} \\ &= \{k \in [M] : D\left(f_i\left(\cdot; \theta_M\right) \| f_i\left(\cdot; \theta_k\right)\right) \neq 0\}.\end{aligned}$$

In other words, let $\bar{\Theta}_i$ be the set of all hypotheses that are *locally indistinguishable* to node i. In this work, we are interested in the case where $|\bar{\Theta}_i| > 1$ for some node i, but the true hypothesis θ_M is *globally identifiable* (see (1)).

Assumption 1. *For every pair $k \neq j$, there is at least one node $i \in [n]$ for which the KL-divergence $D\left(f_i\left(\cdot; \theta_k\right) \| f_i\left(\cdot; \theta_j\right)\right)$ is strictly positive.*

In this case, we ask whether nodes can collectively go beyond the limitations of their local observations and learn θ_M. Since

$$\{\theta_M\} = \bar{\Theta}_1 \cap \bar{\Theta}_2 \cap \ldots \cap \bar{\Theta}_n, \tag{1}$$

it is straightforward to see that Assumption 1 is a sufficient condition for the global identifiability of θ_M when only marginal distributions are known at the nodes. Also, note that this assumption does not require the existence of a single node that can distinguish θ_M from all other hypotheses θ_k, where $k \in [M-1]$. We only require that for every pair $k \neq j$, there is at least one node $i \in [n]$ for which $f_i\left(\cdot; \theta_k\right) \neq f_i\left(\cdot; \theta_j\right)$.

Finally, we define a probability triple $\left(\Omega, \mathcal{F}, \mathsf{P}^{\theta_M}\right)$, where $\Omega = \{\omega : \omega = (\mathbf{X}^{(\mathbf{0})}, \mathbf{X}^{(\mathbf{1})}, \ldots), \forall \mathbf{X}^{(\mathbf{t})} \in \boldsymbol{\mathcal{X}}, \forall t\}$, $\mathcal{F}$ is the $\sigma-$ algebra generated by the observations and P^{θ_M} is the probability measure induced by paths in Ω, i.e., $\mathsf{P}^{\theta_M} = \prod_{t=0}^{\infty} f\left(\cdot; \theta_M\right)$. We use $\mathbb{E}^{\theta_M}[\cdot]$ to denote the expectation operator associated with measure P^{θ_M}. For simplicity we drop θ_M to denote P^{θ_M} by P and denote $\mathbb{E}^{\theta_M}[\cdot]$ by $\mathbb{E}[\cdot]$.

B. Network

We model the communication network between nodes via a directed graph with vertex set $[n]$. We define the neighborhood of node i, denoted by $\mathcal{N}(i)$, as the set of all nodes which have an edge starting from themselves to node i. This means if node $j \in \mathcal{N}(i)$, it can send the information to node i along this edge. In other words, the neighborhood of node i denotes the set of all sources of information available to it. Moreover, we assume that the nodes have knowledge of their neighbors $\mathcal{N}(i)$ only and they have no knowledge of the rest of the network [41].

Assumption 2. *The underlying graph of the network is strongly connected, i.e. for every $i, j \in [n]$ there exists a directed path starting from node i and ending at node j.*

We consider the case where the nodes are connected to every other node in the network by at least one multi-hop path, i.e. a strongly connected graph allows the information gathered to be disseminated at every node throughout the network. Such a network enables learning even when some nodes in the network may not be able to distinguish the true hypothesis on their own, i.e. the case where $|\bar{\Theta}_i| > 1$ for some nodes.

C. The Learning Rule

In this section we provide a learning rule for the nodes to learn θ_M by collaborating with each other through the local communication alone.

We begin by defining the variables required in order to define the learning rule. At every time instant t each node i maintains a private belief vector $\mathbf{q_i^{(t)}} \in \mathcal{P}(\Theta)$ and a public belief vector $\mathbf{b_i^{(t)}} \in \mathcal{P}(\Theta)$, which are probability distributions on Θ. The social interaction of the nodes is characterized by a stochastic matrix W. More specifically, weight $W_{ij} \in [0, 1]$ is assigned to the edge from node j to node i such that $W_{ij} > 0$ if and only if $j \in \mathcal{N}(i)$ and $W_{ii} = 1 - \sum_{j=1}^{n} W_{ij}$. The weight W_{ij} denotes the (relative) confidence node i has on the information it receives from node j.

The steps of learning are given below. Suppose each node i starts with an initial private belief vector $\mathbf{q_i^{(0)}}$. At each time $t = 1, 2, \ldots$ the following events happen:

1) Each node i draws a conditionally i.i.d observation $X_i^{(t)} \sim f_i(\cdot; \theta_M)$.
2) Each node i performs a local Bayesian update on $\mathbf{q_i^{(t-1)}}$ to form $\mathbf{b_i^{(t)}}$ using the following rule. For each $k \in [M]$,

$$b_i^{(t)}(\theta_k) = \frac{f_i\left(X_i^{(t)}; \theta_k\right) q_i^{(t-1)}(\theta_k)}{\sum_{a \in [M]} f_i\left(X_i^{(t)}; \theta_a\right) q_i^{(t-1)}(\theta_a)}. \quad (2)$$

3) Each node i sends the message $\mathbf{Y_i^{(t)}} = \mathbf{b_i^{(t)}}$ to all nodes j for which $i \in \mathcal{N}(j)$. Similarly receives messages from its neighbors $\mathcal{N}(i)$.
4) Each node i updates its private belief of every θ_k, by averaging the log beliefs it received from its neighbors. For each $k \in [M]$,

$$q_i^{(t)}(\theta_k) = \frac{\exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta_k)\right)}{\sum_{a \in [M]} \exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta_a)\right)}. \quad (3)$$

Note that the private belief vector $\mathbf{q_i^{(t)}}$ remain locally with the nodes while their public belief vectors $\mathbf{b_i^{(t)}}$ are exchanged with the neighbors. The objective of learning rule is to ensure that the private belief vector $\mathbf{q_i^{(t)}}$ of each node $i \in [n]$ converges to $\mathbf{1}_M(\cdot)$.

Given the weight matrix W, the network can be thought of as a weighted strongly connected network. Assumption 2, implies that weight matrix W is irreducible. In this context we recall the following fact.

Fact 1 (Section 2.5 of Hoel et. al. [42])**.** *Let W be the transition matrix of a Markov chain. If W is irreducible then the stationary distribution of the Markov chain denoted by $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the normalized left eigenvector of W associated with eigenvalue 1 and it is given as*

$$v_i = \sum_{j=1}^{n} v_j W_{ji}. \quad (4)$$

Furthermore, all components of $\mathbf{v}$ are strictly positive. If the Markov chain is aperiodic, then

$$\lim_{t \to \infty} W^t(i, j) = v_j, \quad i, j \in [n]. \quad (5)$$

If the chain is periodic with period d, then for each pair of states $i, j \in [n]$, there exists an integer $r \in [d]$, such that $W^t(i, j) = 0$ unless $t = md + r$ for some nonnegative integer m, and

$$\lim_{m \to \infty} W^{md+r}(i, j) = v_j d. \quad (6)$$

In the social learning literature, the eigenvector $\mathbf{v}$ also known as the eigenvector centrality; it is a measure of social influence of a node in the network. In particular we will see that v_i determines the contribution of node i in the collective network learning rate.

Definition 1 (Network Divergence)**.** For all $k \in [M-1]$, the network divergence between θ_M and θ_k, denoted by $K(\theta_M, \theta_k)$, is defined as

$$K(\theta_M, \theta_k) \triangleq \sum_{i=1}^{n} v_i D\left(f_i(\cdot; \theta_M) \| f_i(\cdot; \theta_k)\right), \quad (7)$$

where $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the normalized left eigenvector of W associated with eigenvalue 1.

Fact 1 together with Assumption 1 guarantees that $K(\theta_M, \theta_k)$ is strictly positive for every $k \in [M-1]$.

Due to the form of our learning rule, if the initial belief of any $\theta_k, k \in [M]$, for some node is zero then beliefs of that θ_k remain zero in subsequent time intervals. Hence, we require the following assumption.

Assumption 3. *For all $i \in [n]$, the initial private belief $q_i^{(0)}(\theta_k) > 0$ for every $k \in [M]$.*

III. Main Results

A. The Criteria for Learning

Before we present our main results, we discuss the metrics we use to evaluate the performance of a learning rule in the given distributed setup.

Definition 2 (Rate of Rejection of Wrong Hypothesis)**.** For any node $i \in [n]$ and $k \in [M-1]$, define the following

$$\rho_i^{(t)}(\theta_k) \triangleq -\frac{1}{t}\log q_i^{(t)}(\theta_k). \tag{8}$$

The rate of rejection of θ_k in favor of θ_M at node i is defined as

$$\rho_i(\theta_k) \triangleq \liminf_{t\to\infty} \rho_i^{(t)}(\theta_k). \tag{9}$$

Now, let

$$\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \triangleq \left[q_i^{(t)}(\theta_1), q_i^{(t)}(\theta_2), \ldots, q_i^{(t)}(\theta_{M-1})\right]^T. \tag{10}$$

Then

$$\boldsymbol{\rho}_i^{(t)} \triangleq -\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \tag{11}$$

and the rate of rejection at node i is defined as

$$\boldsymbol{\rho}_i \triangleq \liminf_{t\to\infty} \boldsymbol{\rho}_i^{(t)}. \tag{12}$$

If $\rho_i(\theta_k) > 0$ for all $k \in [M-1]$, under a given learning rule the belief vector of node i not only converges to the true hypothesis, it converges exponentially fast. Another way to measure the performance of a learning rule is the rate at which the belief of true hypothesis converges to one.

Definition 3 (Rate of Convergence to True Hypothesis)**.** For any $i \in [n]$ and $k \in [M-1]$, define the rate of convergence μ_i to θ_M by

$$\mu_i \triangleq \liminf_{t\to\infty} -\frac{1}{t}\log(1 - q_i^{(t)}(\theta_M)). \tag{13}$$

Definition 4 (Rate of Social Learning)**.** The total variational error across the network when the underlying true hypothesis is θ_k (where we allow the true hypothesis to vary, i.e. $\theta^* = \theta_k$ for any $k \in [M]$ instead of assuming that it is fixed at $\theta^* = \theta_M$) is given as

$$e^{(t)}(k) = \frac{1}{2}\sum_{i=1}^{n} ||q_i^{(t)}(\cdot) - \mathbf{1}_k(\cdot)|| = \sum_{i=1}^{n}\sum_{j\neq k} q_i^{(t)}(\theta_j). \tag{14}$$

This equals the total probability that all nodes in the network assign to "wrong hypotheses". Now, define

$$e^{(t)} \triangleq \max_{k\in[M]} e^{(t)}(k). \tag{15}$$

The rate of social learning is defined as the rate at which total variational error, $e^{(t)}$, converges to zero and mathematically it is defined as

$$\rho_L \triangleq \liminf_{t\to\infty} -\frac{1}{t}\log e^{(t)}. \tag{16}$$

This measure of performance for the learning rule has been used in the social learning literature [27]. For a given network and a given observation model for nodes, ρ_L gives the least rate of learning guaranteed in the network and therefore provides a worst case guarantee. It is straightforward to see that with a characterization for $\rho_i(\theta_k)$ for all $k \in [M-1]$ we obtain a lower bound on rate of convergence to true hypothesis, μ_i, and on the rate of social learning, ρ_L, under a given learning rule.

B. Learning: Convergence to True Hypothesis

Theorem 1 (Rate of Rejecting Wrong Hypotheses, $\boldsymbol{\rho}_i$)**.** *Let θ_M be the true hypothesis. Under the Assumptions 1–3, for every node in the network, the private belief (and hence the public belief) under the proposed learning rule converges to true hypothesis exponentially fast with probability one. Furthermore, the rate of rejecting hypothesis θ_k in favor of θ_M is given by the network divergence between θ_M and θ_k. Specifically, we have*

$$\lim_{t\to\infty} \mathbf{q}_{\mathbf{i}}^{(\mathbf{t})} = \mathbf{1}_M \quad \mathsf{P}\text{-}a.s. \tag{17}$$

and

$$\boldsymbol{\rho}_i = -\lim_{t\to\infty}\frac{1}{t}\log \tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} = \mathbf{K} \quad \mathsf{P}\text{-}a.s. \tag{18}$$

where

$$\mathbf{K} = \left[K(\theta_M,\theta_1), K(\theta_M,\theta_2), \ldots, K(\theta_M,\theta_{M-1})\right]^T. \tag{19}$$

The proof of Theorem 1 is provided in Appendix A. Theorem 1 establishes that the beliefs of wrong hypotheses, θ_k for $k \in [M-1]$, vanish exponentially fast and it characterizes the exponent with which a node rejects θ_k in favor of θ_M. The rate of rejection is a function of the node's ability to distinguish between the hypotheses, which is given by the KL-divergences and structure of the weighted network, weighted by the eigenvector centrality of the nodes. Hence, every node influences the rate in two ways. Firstly, if the node has higher eigenvector centrality (i.e. the node is centrality located), it has larger influence over the beliefs of other nodes as a result has a greater influence over the rate of exponential decay as well. Secondly, if the node has high KL-divergence (i.e highly informative observations that can distinguish between θ_k and θ_M), then again it increases the rate. If an influential node has highly informative observations then it boosts the rate of rejecting θ_k by improving the rate. We will illustrate this through numerical examples in Section IV-A.

We obtain lower bound on the rate of convergence to the true hypothesis and rate of learning as corollaries to Theorem 1.

Corollary 1 (Lower Bound on Rate of Convergence to θ_M)**.** *Let θ_M be the true hypothesis. Under the Assumptions 1–3, for every $i \in [n]$, the rate of convergence to θ_M can be lower-bounded as*

$$\mu_i \geq \min_{k \in [M-1]} K(\theta_M, \theta_k) \quad \mathsf{P}\text{-}a.s. \tag{20}$$

Corollary 2 (Lower Bound on Rate of Learning)**.** *Let θ_M be the true hypothesis. Under the Assumptions 1–3, the rate of learning ρ_L across the network is lower-bounded by,*

$$\rho_L \geq \min_{i,j \in [M]} K(\theta_i, \theta_j) \quad \mathsf{P}\text{-}a.s.$$

Remark 1. Jadbabaie et. al. proposed a learning rule in [25], which differs from the proposed rule at the private belief vector $\mathbf{q_i^{(t)}}$ formation step. Instead of averaging the log beliefs, nodes average the beliefs received as messages from their neighbors. In [27], Jadbabaie et. al. provide an upper bound on the rate of learning ρ_L obtained using their algorithm. They show

$$\rho_L \leq \alpha \min_{i,j \in [M]} K(\theta_i, \theta_j) \quad \mathsf{P}\text{-a.s.} \tag{21}$$

where α is a constant strictly less than one. Corollary 2 shows that lower bound on ρ_L using the proposed algorithm is greater than the upper bound provided in (21).

C. Concentration under Bounded Log-likelihood ratios

Under mild assumptions, Theorem 1 shows that the belief about a wrong hypothesis θ_k for $k \in [M-1]$ converges to zero exponentially fast at rate equal to the network divergence, $K(\theta_M, \theta_k)$, between θ_M and θ_k with probability one. We strength this result for periodic networks with period d under the following assumption.

Assumption 4. *There exists a positive constant L such that*

$$\max_{i \in [n]} \max_{j,k \in [M]} \sup_{X \in \mathcal{X}_i} \left| \log \frac{f_i\left(X; \theta_j\right)}{f_i\left(X; \theta_k\right)} \right| \leq L. \tag{22}$$

Theorem 2 (Concentration of Rate of Rejecting Wrong Hypotheses, $\rho_i^{(t)}(\theta_k)$)**.** *Let θ_M be the true hypothesis. Under Assumptions 1–4, for periodic networks with period d, for every node $i \in [n]$, $k \in [M-1]$, and for all $\epsilon > 0$ we have*

$$\lim_{t \to \infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M, \theta_k) - \epsilon\right) \leq -\frac{\epsilon^2}{2L^2 d}. \tag{23}$$

For $0 < \epsilon \leq L - K(\theta_M, \theta_k)$, we have

$$\begin{aligned} &\lim_{t \to \infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \epsilon\right) \\ &\leq -\frac{1}{2L^2 d} \min \left\{\epsilon^2, \min_{j \in [M-1]} K(\theta_M, \theta_j)^2\right\}. \end{aligned} \tag{24}$$

For $\epsilon \geq L - K(\theta_M, \theta_k)$ we have

$$\begin{aligned} &\lim_{t \to \infty} \frac{1}{t} \log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M, \theta_k) + \epsilon\right) \\ &\qquad \leq -\min_{k \in [M-1]} \left\{\frac{K(\theta_M, \theta_k)^2}{2L^2 d}\right\}. \end{aligned} \tag{25}$$

Corollary 3 (Rate of convergence to True Hypothesis)**.** *Let θ_M be the true hypothesis. Under Assumptions 1–4, for every $i \in [n]$, we have*

$$\mu_i = \min_{k \in [M-1]} K(\theta_M, \theta_k) \quad \mathsf{P}\text{-}a.s.$$

Proofs of Theorem 2 and Corollary 3 are provided in Appendix B. From Theorem 1 we know that $\rho_i^{(t)}(\theta_k)$ converges to $K(\theta_M, \theta_k)$ almost surely. Theorem 2 strengthens Theorem 1 by showing that the probability of sample paths where $\rho_i^{(t)}(\theta_k)$ deviates by some fixed ϵ from $K(\theta_M, \theta_k)$ vanishes exponentially fast. This implies that $\rho_i^{(t)}(\theta_k)$ converges to $K(\theta_M, \theta_k)$ exponentially fast in probability. Theorem 2 also characterizes a lower bound on the exponent when the probability of such events vanishes and shows that periodicity of the network reduces the exponent.

D. Large Deviation Analysis

We require a technical assumption that relaxes the assumption of bounded ratios of the likelihood functions in prior work [1], [2], [31], [43].

Assumption 5. *For every pair $\theta_i \neq \theta_j$ and every node $k \in [n]$, the random variable $\left|\log \frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}\right|$ has finite log moment generating function under distribution $f_k\left(\cdot; \theta_j\right)$.*

Next, we give examples of families of distributions which satisfy Assumption 5 but violate Assumption 4.

Remark 2. Distributions $f(X;\theta_i)$ and $f(X;\theta_j)$ for $i \neq j$ which the following properties for some positive constants C and β, satisfy Assumption 5

$$\mathsf{P}_i\left(\frac{f(X;\theta_j)}{f(X;\theta_i)} \geq x\right) \leq \frac{C}{x^\beta}, \quad \mathsf{P}_i\left(\frac{f(X;\theta_i)}{f(X;\theta_j)} \geq x\right) \leq \frac{C}{x^\beta}. \tag{26}$$

Note that (26) is a sufficient condition but not a necessary condition. Examples 1–2 below do not satisfy (26) yet satisfy Assumption 5.

Example 1 (Gaussian Mixtures)**.** Let $f(X;\theta_1) = \mathcal{N}(\mu_1, \sigma)$ and $f(X;\theta_2) = \mathcal{N}(\mu_2, \sigma)$. Then

$$g_1(x) := \left|\log \frac{f(x;\theta_1)}{f(x;\theta_2)}\right| \leq c_1|x| + c_2, \tag{27}$$

where $c_1 = \left|\frac{\mu_1 - \mu_2}{\sigma^2}\right|$ and $c_2 = \left|\frac{\mu_1^2 - \mu_2^2}{2\sigma^2}\right|$. Hence, for $i \in \{1, 2\}$ and for $\lambda \geq 0$ we have

$$\mathbb{E}_i\left[e^{\lambda g_1(X)}\right] \leq e^{c_2 \lambda} \mathbb{E}_i\left[e^{c_1 \lambda |X|}\right] < \infty. \tag{28}$$

More generally for $i \in \{1, 2\}$, and $p \in [0, 1]$, let

$$\begin{aligned} f(x;\theta_i) = &\frac{p}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x - \alpha_i)^2}{2\sigma^2}\right) \\ &+ \frac{1-p}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x - \beta_i)^2}{2\sigma^2}\right). \end{aligned} \tag{29}$$

Then the log moment generating function of $\left|\log \frac{f(X;\theta_1)}{f(X;\theta_2)}\right|$ is finite for all $\lambda \geq 0$.

Example 2 (Gamma distribution)**.** Let $f(X;\theta_1) = \frac{\beta^{\alpha_1}}{\Gamma(\alpha_1)} x^{\alpha_1 - 1} e^{-\beta x}$ and $f(X;\theta_2) = \frac{\beta^{\alpha_2}}{\Gamma(\alpha_2)} x^{\alpha_2 - 1} e^{-\beta x}$, then

$$g_2(x) := \left|\log \frac{f(x;\theta_1)}{f(x;\theta_2)}\right| \leq c_1 |\log x| + c_2, \tag{30}$$

where $c_1 = |\alpha_1 - \alpha_2|$ and $c_2 = \left|(\alpha_1 - \alpha_2)\log\beta + \log\frac{\Gamma(\alpha_2)}{\Gamma(\alpha_1)}\right|$. Hence, for $i \in \{1,2\}$ and for $\lambda \geq 0$ we have

$$\mathbb{E}_i\left[e^{\lambda g_2(X)}\right] \leq e^{c_2\lambda}\mathbb{E}_i\left[e^{c_1\lambda|\log X|}\right] < \infty. \tag{31}$$

The above examples show that Assumption 5 is satisfied for distributions which have unbounded support. In order to analyze the concentration of $\boldsymbol{\rho}_i^{(t)}$ under Assumption 5 we replace Assumption 2 with the following assumption.

Assumption 2′. The underlying graph of the network is strongly connected and aperiodic.

Next we provide few more definitions. Let

$$\mathbf{Y}^{(t)}(\theta_k) \triangleq \langle \mathbf{v}, \mathbf{L}^{(t)}(\theta_k)\rangle, \tag{32}$$

where $\mathbf{L}^{(t)}(\theta_k)$ is the vector of log likelihood ratios given by

$$\mathbf{L}^{(t)}(\theta_k) = \left[\log\frac{f_1\left(X_1^{(t)};\theta_k\right)}{f_1\left(X_1^{(t)};\theta_M\right)}, \ldots, \log\frac{f_n\left(X_n^{(t)};\theta_k\right)}{f_n\left(X_n^{(t)};\theta_M\right)}\right]^T. \tag{33}$$

Definition 5 (Moment Generating Function)**.** For every $\lambda_k \in \mathbb{R}$, let $\Lambda_k(\lambda_k)$ denote the log moment generating function of $\mathbf{Y}^{(t)}(\theta_k)$ by

$$\Lambda_k(\lambda_k) \triangleq \log\mathbb{E}[e^{\lambda_k \mathbf{Y}^{(t)}(\theta_k)}] = \log\mathbb{E}[e^{\lambda_k\langle\mathbf{v},\mathbf{L}(\theta_k)\rangle}] \tag{34}$$

For every $\boldsymbol{\lambda} \in \mathbb{R}^{M-1}$, let $\Lambda(\boldsymbol{\lambda})$ denote the log moment generating function of $\mathbf{Y}$ by

$$\Lambda(\boldsymbol{\lambda}) \triangleq \log\mathbb{E}[e^{\langle\boldsymbol{\lambda},\mathbf{Y}\rangle}]. \tag{35}$$

Note that each entry of vector $\mathbf{Y}^{(t)}$ is a function of joint observation vector $\mathbf{X}^{(t)}$ whose distribution is governed by $f(\cdot;\theta_M)$.

Definition 6 (Large Deviation Rate Function)**.** For all $x \in \mathbb{R}$, let $I_k(x)$ denote the Fenchel-Legendre transform of $\Lambda_k(\cdot)$

$$I_k(x) \triangleq \sup_{\lambda_k\in\mathbb{R}}\left\{\lambda x - \Lambda_k(\lambda_k)\right\}. \tag{36}$$

For all $\mathbf{x} \in \mathbb{R}^{M-1}$, let $I(\mathbf{x})$ denote the Fenchel-Legendre transform of $\Lambda(\cdot)$

$$I(\mathbf{x}) \triangleq \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{M-1}}\left\{\langle\boldsymbol{\lambda},\mathbf{x}\rangle - \Lambda(\boldsymbol{\lambda})\right\}. \tag{37}$$

Theorem 3 (Large Deviations of $\boldsymbol{\rho}_i^{(t)}$)**.** *Let θ_M be the true hypothesis. Under Assumptions 1, 2′, 3, 5, the rate of rejection $\boldsymbol{\rho}_i^{(t)}$ satisfies an Large Deviation Principle with rate function $J(\cdot)$, i.e., for any set $F \subset \mathbb{R}^{M-1}$ we have*

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\boldsymbol{\rho}_i^{(t)} \in F\right) \geq -\inf_{\mathbf{y}\in F^o} J(\mathbf{y}), \tag{38}$$

and

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\boldsymbol{\rho}_i^{(t)} \in F\right) \leq -\inf_{\mathbf{y}\in\bar{F}} J(\mathbf{y}), \tag{39}$$

where large deviation rate function $J(\cdot)$ is defined as

$$J(\mathbf{y}) \triangleq \inf_{\mathbf{x}\in\mathbb{R}^{M-1}:g(\mathbf{x})=\mathbf{y}} I(\mathbf{x}),\ \forall\,\mathbf{y}\in\mathbb{R}^{M-1}, \tag{40}$$

where $g : \mathbb{R}^{M-1} \to \mathbb{R}^{M-1}$ is a continuous mapping given by

$$g(\mathbf{x}) \triangleq \left[g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_{M-1}(\mathbf{x})\right]^T, \tag{41}$$

and

$$g_k(\mathbf{x}) \triangleq x_k - \max\{0, x_1, x_2, \ldots, x_{M-1}\}. \tag{42}$$

The proof of Theorem 3 is provided in Appendix C. Theorem 3 characterizes the asymptotic rate of concentration of $\boldsymbol{\rho}_i^{(t)}$ in any set $F \subset \mathbb{R}^{M-1}$. In other words, it characterizes the rate at which the probability of deviations in each $\rho_i^{(t)}(\theta_k)$ from the rate of rejection $K(\theta_M,\theta_k)$ for every $\theta_k \neq \theta_M$ vanish simultaneously. It characterizes the asymptotic rate as a function of the observation model of each node (not just the bound L on the ratios of log-likelihood function) and as a function of eigenvector centrality $\mathbf{v}$. The following corollary specializes this result to obtain the individual rate of rejecting a wrong hypothesis at every node. It can be obtained by repeating the proof of Theorem 3 for each hypothesis alone.

Corollary 4. *Let θ_M be the true hypothesis. Under Assumptions 1, 2′, 3, 5, for $0 < \epsilon \leq K(\theta_M,\theta_k)$, $k \in [M-1]$, we have*

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M,\theta_k) - \epsilon\right) = -I_k\left(K(\theta_M,\theta_k) - \epsilon\right). \tag{43}$$

For $\epsilon > 0$, we have

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M,\theta_k) + \epsilon\right) = -I_k\left(K(\theta_M,\theta_k) + \epsilon\right). \tag{44}$$

Using Theorem 3 and Hoeffding's Lemma, we obtain the following corollary.

Corollary 5. *Suppose Assumption 4 is satisfied for some finite $L \in \mathbb{R}$. For ϵ as specified in Theorem 2, we recover the exponents of Theorem 2 under aperiodic networks, given by*

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \geq K(\theta_M,\theta_k) + \epsilon\right) \leq -\frac{\epsilon^2}{2L^2}, \tag{45}$$

and

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \leq K(\theta_M,\theta_k) - \epsilon\right) \leq -\frac{\epsilon^2}{2L^2}. \tag{46}$$

Remark 3. Under Assumption 4, Corollary 5 shows that lower bound on the asymptotic rate of concentration of $\boldsymbol{\rho}_i^{(t)}$ as characterized by Theorem 2 is loose in comparision to that obtained from Theorem 3. Nedic et al. [32] and Shahrampour et al. [31] provide non-asymptotic lower bounds on the rate of concentration of $\boldsymbol{\rho}_i^{(t)}$ whose asymptotic form coincides with the lower bound on rate characterized by Theorem 2 for aperiodic networks. This implies that under Assumption 4 Theorem 3 provides a tighter asymptotic rate than their results in [31], [32]. Hence, Theorem 3 strengthens Theorem 2 by extending the large deviation to larger class of distributions and providing a tighter bound that captures the complete effect of nodes' influence in the network and the local observation statistics.

IV. Examples

In this section through numerical examples we illustrate how nodes learn using the proposed learning rule and examine the factors which affect the rate of rejection of wrong hypotheses and its rate of concentration.

A. *Factors influencing Convergence*

Example 3. Consider a group of two nodes as shown in Figure 1, where the set of hypotheses is $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and true hypothesis $\theta^* = \theta_4$. Observations at each node at time t, $X_i^{(t)}$, take values in $\mathbb{R}^{100}$ and have a Gaussian distribution. For node 1, $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3) = \mathcal{N}(\boldsymbol{\mu}_{11}, \boldsymbol{\Sigma})$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4) = \mathcal{N}(\boldsymbol{\mu}_{12}, \boldsymbol{\Sigma})$, and for node 2, $f_2(\cdot;\theta_1) = f_2(\cdot;\theta_2) = \mathcal{N}(\boldsymbol{\mu}_{21}, \boldsymbol{\Sigma})$ and $f_2(\cdot;\theta_3) = f_2(\cdot;\theta_4) = \mathcal{N}(\boldsymbol{\mu}_{22}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{11}, \boldsymbol{\mu}_{12}, \boldsymbol{\mu}_{21}, \boldsymbol{\mu}_{22} \in \mathbb{R}^{100}$ and $\boldsymbol{\Sigma}$ is a positive semi-definite matrix of size 100-by-100. Here, node 1 can identify the column containing θ_4, and node 2 can identify the row. In other words, $\bar{\Theta}_1 = \{\theta_2, \theta_4\}$ and $\bar{\Theta}_2 = \{\theta_3, \theta_4\}$. Also, $\theta_4 = \bar{\Theta}_1 \cap \bar{\Theta}_2$, hence θ_4 is globally identifiable.

1) Strong Connectivity: Nodes are connected to each other in a network and the weight matrix is given by

$$W = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}. \tag{47}$$

Figure 2 shows the evolution of beliefs with time for node 2 on a single sample path. We see that using the proposed learning rule, belief of θ_4 goes to one while the beliefs of wrong hypotheses go to zero. This example shows that each node through collaboration is able to learn θ_4. Figure 3 shows the rate of rejection of wrong hypotheses. We see that the rate of rejection θ_k for $k \in \{1, 2, 3\}$ closely follows the asymptotic rate $K(\theta_4, \theta_k)$.

Suppose the nodes are connected to each other in a network whose weight matrix is given by

$$W = \begin{pmatrix} 1 & 0 \\ 0.5 & 0.5 \end{pmatrix}. \tag{48}$$

Since there is no path from node 2 to node 1, the network is not strongly connected. Node 2 as seen in Figure 4 does not converge to θ_4. Even though node 1 cannot distinguish the elements of $\bar{\Theta}_1$ from θ_4, it rejects the hypotheses in $\{\theta_1, \theta_3\}$ in favor of θ_4. This forces node 2 also to reject the set $\{\theta_1, \theta_3\}$. For node 1, θ_2 and θ_4 are observationally equivalent, hence their respective beliefs equal half. But node 2 oscillates between θ_2 and θ_4 and is unable to learn θ_4. Hence, when the network is not strongly connected both nodes fail to learn.

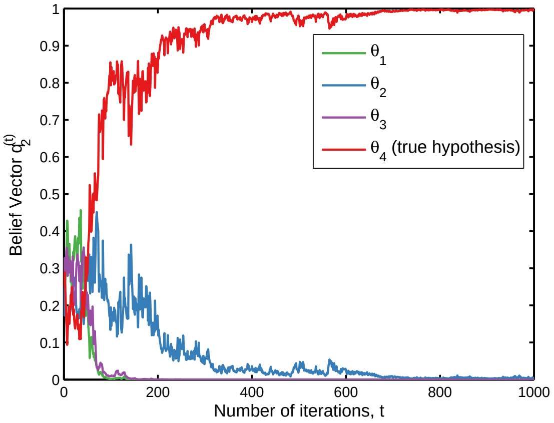

Fig. 2. For the set of nodes described in Figure 1, this figure shows the evolution of beliefs for one instance using the proposed learning rule. Belief of the true hypothesis θ_4 of node 2 converges to 1 and beliefs of all other hypotheses go to zero.

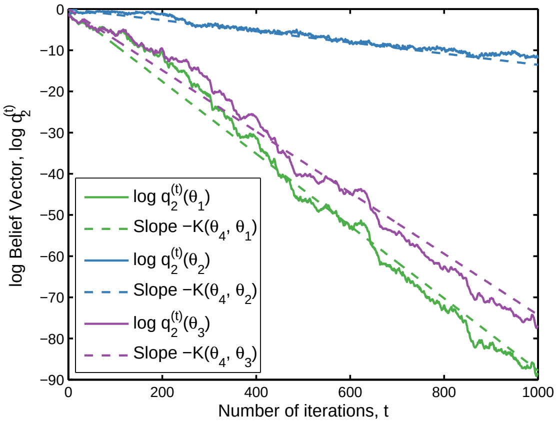

Fig. 3. Figure shows the exponential decay of beliefs of θ_1, θ_2 and θ_3 of node 2 using the learning rule.

In this setup we apply the learning rule considered in [25], where in the consensus step public beliefs are updated by averaging the beliefs received from the neighbors instead of averaging the logarithm of the beliefs. As seen in Figure 5, rate of rejecting learning using the proposed learning rule is greater than the upper bound on learning rule in [25]. Note that the precision of the belief vectors in the simulations is 8 bytes (64 bits) per hypothesis. This implies the nodes each send 32 bytes per unit time, which is less than the case when nodes exchange local Gaussian observations which may

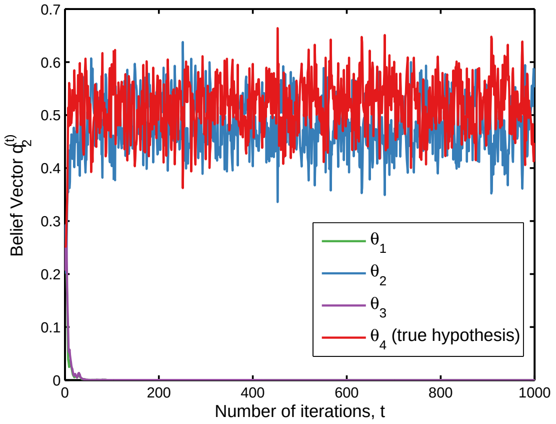

Fig. 4. Figure shows the beliefs of node 2 shown in Figure 1. When the network is not strongly connected node 2 cannot learn θ_4.

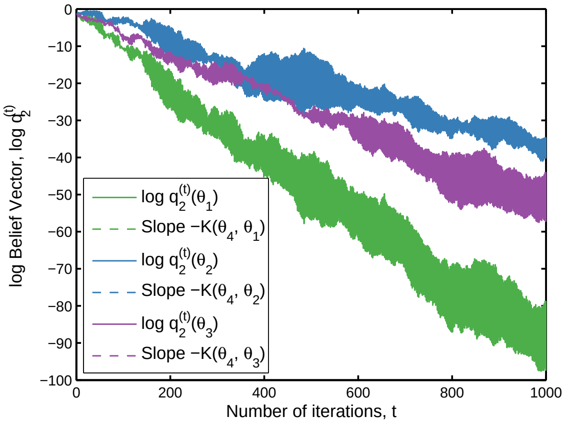

Fig. 6. Figure shows the exponential decay of beliefs of θ_1, θ_2, and θ_3 of node 2 connected to node 1 in a periodic network with period 2.

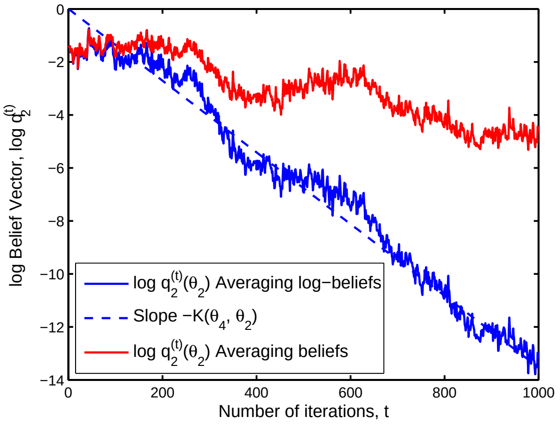

Fig. 5. Figure shows that the rate of rejection of θ_2 using the proposed learning rule (averaging the log beliefs) is greater than the rate of rejection of θ_2 obtained using the learning rule in [25] (averaging the beliefs).

require data rate as high as 800 bytes per observation.

2) Periodicity: Now suppose the nodes are connected to each other in periodic network with period 2 and the weight matrix given by

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{49}$$

From Figure 6, we see that the belief on wrong hypotheses converges to zero but beliefs oscillate significantly about the expected value of rate of rejection as compared to the case of an aperiodic network considered in (47).

Even though nodes do not have a positive self-weight (W_{ii}), the new information (through observations) entering at every node reaches its neighbors and gets dispersed in throughout the network; eventually reaches every node. Hence, nodes learn even when the network is periodic as long as it remains strongly connected.

3) Eigenvector Centrality and Extent of distinguishability: From Theorem 1, we know that a larger weighted sum of the KL divergences, *i.e.* a larger network divergence, $K(\theta_M, \theta_k)$, yields a better rate of rejecting hypothesis θ_k. We look at a numerical example to show this.

Example 4. Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5\}$ and $\theta^* = \theta_4$. Consider a set of 25 nodes which are arranged in 5×5 array to form a grid. We obtain a grid network by connecting every node to its adjacent nodes. We define the weight matrix as,

$$W_{ij} = \begin{cases} \frac{1}{|\mathcal{N}(i)|}, & \text{if } j \in \mathcal{N}(i) \\ 0, & \text{otherwise} \end{cases} \tag{50}$$

Consider an extreme scenario where only one node can distinguish true hypothesis θ_1 from the rest and to the remaining nodes in the network all hypotheses are observationally equivalent *i.e.* $\bar{\Theta}_i = \Theta$ for 24 nodes and $\bar{\Theta}_i = \{\theta_1\}$ for only one node. We call that one node which can distinguish the true hypothesis from other hypotheses as the "informed node" and the rest of the nodes called the "non-informed nodes".

For the weight matrix in (50), the eigenvector centrality of node i is proportional to $\mathcal{N}(i)$, which means in this case, more number of neighbors implies higher social influence. This implies that the corner nodes (namely node 1, node 5, node 20 and node 25 at the four corners of the grid) have least eigenvector centrality among all nodes. Hence, they are least influential. The nodes on four edges have a greater influence than the corner nodes. Most influential nodes are the ones with four connections, such as node 13 which is located in third row and third column of the grid. It is also the central location of the grid.

Figure 7 shows the variation in the rate of rejection of θ_2 of node 5 as the location of informed node changes. We see that if the informed node is at the center of the grid then the rate of rejection is fastest and the rate is slowest when the informed node is placed at a corner. In other words, rate of convergence is highest when the most influential node in the network has high distinguishability.

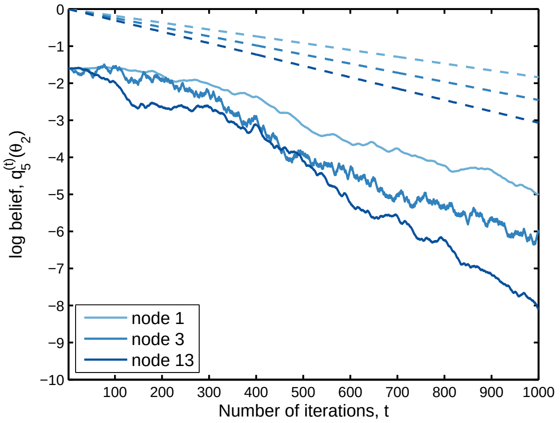

Fig. 7. Figure illustrates the manner in which rate of rejection of θ_2 at node 5 is influenced by varying the location of an informed node. As seen here when the informed node is more central *i.e.* at node 13, rate of rejection is fastest and when the informed node is at the corner node 1, rate of rejection is slowest.

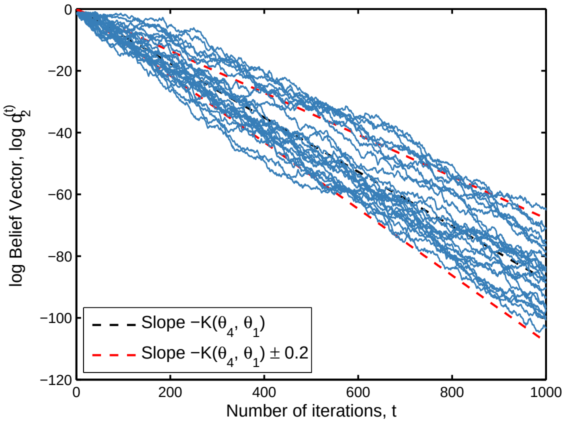

Fig. 8. Figure shows the decay of belief of θ_1 (wrong hypothesis) of node 2 for 25 instances. We see that the number of sample paths on which the rate of rejecting θ_1 deviates more than $\eta = 0.1$ reduces as the number of iterations increase.

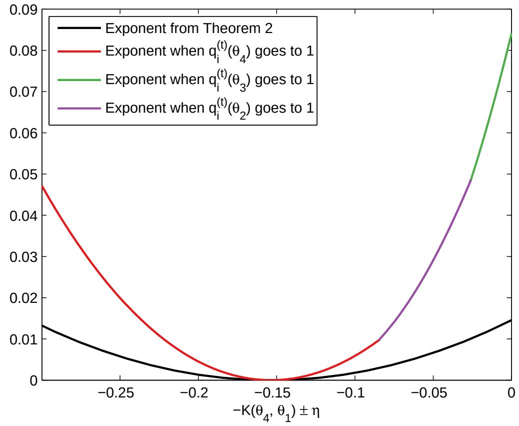

Fig. 9. Figure shows the asymptotic exponent with which the probability of events where rate of rejecting θ_1 deviates by η from $K(\theta_4, \theta_1)$; θ_4 is the true hypothesis. The black curve shows the asymptotic exponent as characterized by Theorem 2. The colored curve shows the exact asymptotic exponent as characterized by Theorem 3, where the exponent depends on the hypothesis to which the learning rule is converging. This shows that small deviations from $K(\theta_4, \theta_1)$ occur when the learning rule is converging to θ_4 and larger deviations occur when the learning rule is converging to a wrong hypothesis.

B. *Factors influencing Concentration*

Now to examine the results from Theorem 2 and Theorem 3, we go back to Example 3, where two nodes are in a strongly connected aperiodic network given by (47). Observation model for each node is defined as follows. For node 1, $f_1(\cdot;\theta_1) = f_1(\cdot;\theta_3) \sim \text{Ber}(\frac{4}{5})$ and $f_1(\cdot;\theta_2) = f_1(\cdot;\theta_4) \sim \text{Ber}(\frac{1}{4})$, and for node 2, $f_2(\cdot;\theta_1) = f_2(\cdot;\theta_2) \sim \text{Ber}(\frac{1}{3})$ and $f_2(\cdot;\theta_3) = f_2(\cdot;\theta_4) \sim \text{Ber}(\frac{1}{4})$. Figure 8 shows the exponential decay of θ_1 for 25 instances. We see that the number of sample paths that deviate more than $\epsilon = 0.1$ from $K(\theta_4, \theta_1)$ decrease with number of iterations. Theorem 2 characterizes the asymptotic rate at which the probability of such sample paths vanishes when the log-likelihoods are bounded. This asymptotic rate is given as a function of L and period of the network. From Corollary 5 the rate given by Theorem 2 is loose for aperiodic networks. A tighter bound which utilizes the complete observation model is given by Theorem 3. Figure 9 shows the gap between the rates.

Figure 9 in the context of Example 3 shows the rate at which the probability of sample paths deviating from rate of rejection can be thought of as operating in three different regimes. Here, each regime denotes the hypothesis to which the learning rule is converging. In order to see this consider the rate function of θ_1, i.e. $J_1(\cdot)$ from Corollary 4;

$$J_1(\mathbf{y}) = \inf_{x \in \mathbb{R}^3 : g(\mathbf{x}) = y} I(\mathbf{x}), \forall y \in \mathbb{R}.$$

The behavior of the rate function $J_1(\cdot)$ depends on the function $g_1(\mathbf{x}) = x_1 - \max\{0, x_1, x_2, x_3\}$. Whenever $g_1(\mathbf{x}) = x_1$, the rate function is $I_1(\cdot)$. This shows that whenever there is a deviation of $x - k(\theta_4, \theta_1)$ from the rate of rejection of θ_1, the sample paths that vanish with slowest exponents are those for which $\frac{1}{t} \log \frac{q_i^{(t)}(\theta_1)}{q_i^{(t)}(\theta_4)} < 0$ as $t \to \infty$. In other words, small deviations occur when the learning rule is converging to true hypothesis θ_4 and they depend on $I_1(\cdot)$ (and hence θ_1) alone. Whereas large deviations occur when the learning rule is mistakenly converging to a wrong hypothesis and hence, the rate function depends on θ_1 and the wrong hypothesis to which the learning rule is converging. Hence, we have three different regimes corresponding to the three wrong hypotheses.

C. *Learning with Communication Constraints*

Now, we consider a variant of our learning rule where the communication between the nodes is quantized to belong to a predefined finite set. Each node i starts with an initial private belief vector $\mathbf{q_i^{(0)}}$ and at each time $t = 1, 2, \ldots$ the following events happen:

1) Each node i draws a conditionally i.i.d observation $X_i^{(t)} \sim f_i\left(\cdot;\theta_M\right)$.
2) Each node i performs a local Bayesian update on $\mathbf{q_i^{(t-1)}}$ to form $\mathbf{b_i^{(t)}}$ using the following rule. For each $k \in [M]$,

$$b_i^{(t)}(\theta_k) = \frac{f_i\left(X_i^{(t)};\theta_k\right) q_i^{(t-1)}(\theta_k)}{\sum_{a\in[M]} f_i\left(X_i^{(t)};\theta_a\right) q_i^{(t-1)}(\theta_a)}. \qquad (51)$$

3) Each node i sends the message $Y_i^{(t)}(\theta_k) = \left[Db_i^{(t)}(\theta_k)\right]$, for all $k \in [M]$, to all nodes j for which $i \in \mathcal{N}(j)$, where $D \in \mathbb{Z}^+$ and

$$[x] = \begin{cases} \lfloor x \rfloor + 1, & \text{if } x > \lfloor x \rfloor + 0.5, \\ \lfloor x \rfloor, & \text{if } x \leq \lfloor x \rfloor + 0.5, \end{cases} \qquad (52)$$

where $\lfloor x \rfloor$ denotes the largest integer less than x.
4) Each node i normalizes the beliefs received from the neighbors $\mathcal{N}(i)$ as

$$\tilde{Y}_i^{(t)}(\theta_k) = \frac{Y_i^{(t)}(\theta_k)}{\sum_{a\in[M]} Y_i^{(t)}(\theta_a)} \qquad (53)$$

and updates its private belief of θ_k, for each $k \in [M]$,

$$q_i^{(t)}(\theta_k) = \frac{\exp\left(\sum_{j=1}^{n} W_{ij} \log \tilde{Y}_i^{(t)}(\theta_k)\right)}{\sum_{a\in[M]} \exp\left(\sum_{j=1}^{n} W_{ij} \tilde{Y}_i^{(t)}(\theta_a)\right)}. \qquad (54)$$

In the above learning rule, the belief on each hypothesis belongs to a set of size $D + 1$. Hence transmitting the entire belief vector, i.e., transmitting the entire message requires $M \log(D + 1)$ bits.

Note that all of our simulations so far, we have used 64-bit precision to represent the belief on each hypothesis, meaning our simulations can be interpreted as limiting the communication links to support 64 bits, or equivalently 8 bytes, per hypothesis per unit of time. Our previous numerical results show a close match with the analysis using this level of quantization. Next we show the impact of a coarser quantization.

Example 5. Consider a network of radars or ultrasound sensors whose aim is to find the location of a target. Each sensor can sense the target's location along one dimension only, whereas the target location is a point in three-dimensional space. Consider the configuration in Figure 10: there are two nodes along each of the three coordinate axes at locations $[\pm 2, 0, 0]$, $[0, \pm 2, 0]$, and $[0, 0, \pm 2]$. The communication links are given by the directed arrows. Nodes located on the x-axis can sense whether x-coordinate of the target lies in the interval $(-2, -1]$ or in the interval $(-1, 0)$ or in the interval $[0, 1)$ or in the interval $[1, 2)$. If a target is located in the interval $(-\infty, -2] \cup [2, \infty)$ on the x-axis then no node can detect it. Similarly nodes on y-axis and z-axis can each distinguish between 4 distinct non-intersecting intervals on the y-axis and the z-axis respectively. Therefore, the total number of hypotheses is $M = 4^3 = 64$.

The sensors receive signals which are three dimensional Gaussian vectors whose mean is altered in the presence of a target. In the absence of a target, the ambient signals have a Gaussian distribution with mean $[0, 0, 0]$. For the sensor node along x-axis located at $[2, 0, 0]$, if the target has x-coordinate $\theta_x \in (-2, 2)$, the mean of the sensor's observation is $[\lfloor 3 + \theta_x \rfloor, 0, 0]$. If a target is located in $(-\infty, -2] \cup [2, \infty)$ on the x-axis, then the mean of the Gaussian observations is $[0, 0, 0]$. Local marginals of the nodes along y-axis and z-axis are described similarly, i.e., as the target moves away from the node by one unit the signal mean strength goes by one unit. For targets located at a distance four units and beyond the sensor cannot detect the target. In this example, suppose θ_1 is the true hypothesis.

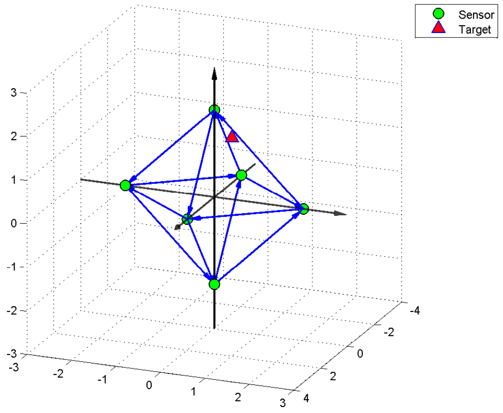

Fig. 10. Figure shows a sensor network where each node is a low cost radar that can sense along the axis it is placed and not the other. The directed edges indicate the directed communication between the nodes. Through cooperative effort the nodes aim to learn location of the target in 3 dimensions.

Consider $D = 2^{12} - 1$ which implies that belief on each hypothesis is of size 12 bits or equivalently 1.5 bytes. Figure 11 shows evolution of log beliefs of node 3 for hypotheses for θ_2, θ_5 and θ_6 for 500 instances when the link rate is limited to 1.5 bytes per hypothesis per unit time. We see that the learning rule converges to the true hypotheses on all 500 instances. Similarly, Figure 12 shows the evolution of beliefs of node 3 for hypotheses θ_2, θ_5 and θ_6 when the link rate is limited to 1 byte per hypothesis per unit time, i.e., when $D = 2^8 - 1$. We see that the learning rule converges to a wrong hypothesis θ_2. However, on the same sample path in Figure 13 we see that if the link rate is 1.5 bytes per hypothesis per unit time, the learning rule converges to true hypothesis. This happens because on every sample path our learning rule has an initial transient phase where beliefs may have large fluctuations during which the belief on true hypothesis may get close to zero. For low link

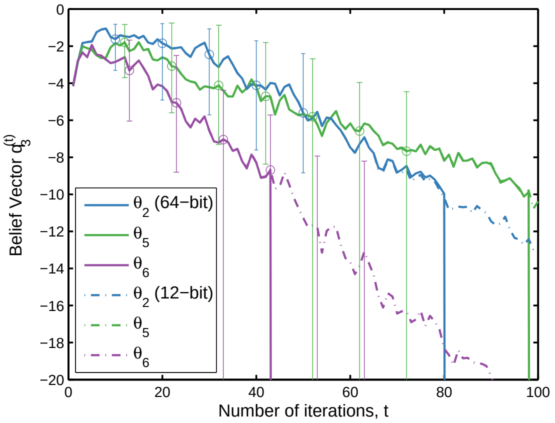

Fig. 11. The solid lines in figure show the evolution of the log beliefs of node 3 with time for hypotheses θ_2, θ_5 and θ_6 when links support a maximum of 12 bits per hypothesis per unit time. This is compared with the evolution of the log beliefs with no rate restriction case (dotted lines) which translates a maximum of 64 bits per hypothesis per unit time. Figure also shows the confidence intervals (one standard deviation above and below) around log beliefs over 500 instances of learning rule with 12 bits per hypothesis. We see the learning rule with link rate 12 bits per hypothesis converges in all the instances.

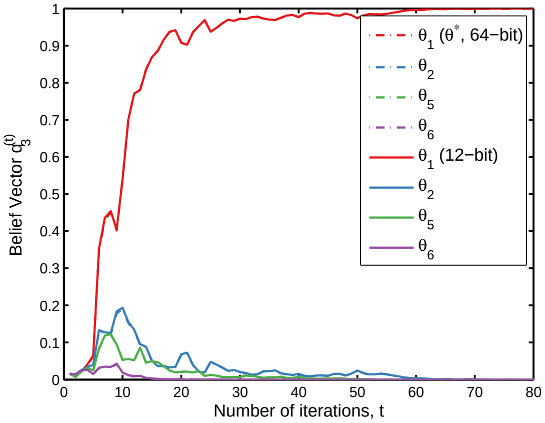

Fig. 13. The solid lines in figure show the evolution of the beliefs of node 3 with time for hypotheses θ_2, θ_5 and θ_6 when links support a maximum of 12 bits per hypothesis per unit time. This is compared with the evolution of the beliefs with no rate restriction case (dotted lines) which in our simulations translates to the case when the links support a maximum of 64 bits per hypothesis per unit time. On the same sample path in Figure 12, we see that learning rule converges to true hypothesis when the communication is restricted to 12 bits per hypothesis.

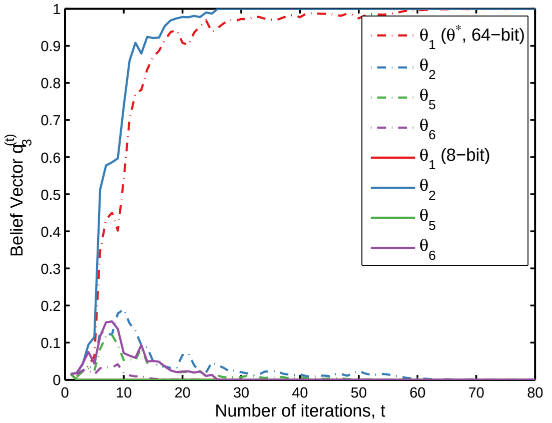

Fig. 12. The solid lines in the figure show the evolution of the log beliefs of node 3 with time for hypotheses θ_2, θ_5 and θ_6 when links support a maximum of 8 bits per hypothesis per unit time. This is compared with the evolution of the log beliefs with no rate restriction case (dotted lines) which translates a maximum of 64 bits per hypothesis per unit time. For this sample path, we see that learning rule converges to a wrong hypothesis θ_5 when the communication is restricted to 8 bits per hypothesis.

rates (small D), even when the belief on true hypothesis is strictly positive but less than $\frac{1}{2D}$, it gets quantized to zero. Recall that for our learning rule, when a belief goes to zero, propagates the zero belief to all subsequent time instants. This shows that as we increase link rate (increase value of D), the quantized learning rule is more robust to the initial fluctuations. Moreover, we observe that for both Examples 3 and 5, when link rates are greater than or equal to 1.5 bytes per hypothesis per unit time the learning rule converges for all instances and its performance coincides with the prediction of our the analysis under the assumption of perfect links.

V. Discussion

In this paper we study a learning rule through which a network of nodes make observations and communicate in order to collectively learn an unknown fixed global hypothesis that statistically governs the distribution of their observations. Our learning rule performs local Bayesian updating followed by averaging log-beliefs. We showed that our rule guarantees exponentially fast convergence to the true hypothesis almost surely. We showed the rate of rejection of any wrong hypothesis has an explicit characterization in terms of the local divergences and network topology. Furthermore, under the (mild technical) Assumption 5 on the tail of the log-likelihood ratios of observations, we provide an asymptotically tight characterization of rate of concentration for the rate of rejection of wrong hypotheses. This assumption admits a broad class of distributions with unbounded support such as Gaussian mixtures. In the next subsections we address two important aspects of our algorithm construction and network model.

A. *Lack of Knowledge of Joint Observation Distribution*

Our algorithm does not require that the the nodes in the network (a) have knowledge of the full joint distribution of the observations nor (b) share their raw local observations. These two properties of our algorithm are highly desirable in many social network settings due to privacy considerations. The performance of our algorithm seems to be overtly pessimistic compared to the performance of a fully cooperative network with identically distributed and independent observations across the nodes (where the rate of rejecting the wrong hypothesis is n times our rate $K(\theta^*, \theta)$). However interestingly, in the case of fully correlated identical observations across the network, our algorithm performs as well as a centralized aggregator would perform. In short, our work can be viewed as a first step towards addressing these questions in settings where nodes keep their local observations and marginal distributions and completely prioritizing local privacy. Nonetheless, we acknowledge that many nontrivial questions remain: (i) what is the trade-off between privacy preservation and learning rate and (ii) what are the cost/benefits of learning the joint distribution in order to optimally combine the local observations.

B. Availability of Perfect Communication Links

In this work, we have assumed that communicating public beliefs among the neighbors can occur with an infinite precision. Although this is a hard assumption to justify in resource-constrained settings, we believe that it is a reasonable abstraction for a practical "protocol-level" model of communication constraints, in which sufficiently high data rates are available to send messages when nodes are within each others' communication range, whereas no communication is possible for physically distant nodes. In Section IV-C, we have provided detailed simulations to show that the gap between the true model and the idealized protocol model is not of significant practical consequence. In particular, Examples 3 and 5 show the impact of quantizing the beliefs before exchanging them is negligible at even low link rates. However, from a theoretical perspective, a study of distributed hypothesis testing with constraints on communication is a major topic of ongoing research [10], [44].

Furthermore, through Example 5, we have also highlighted the practical gains, in terms of communication, associated with communicating the beliefs instead of the raw local observations where the observations are in a high dimensional space. In other words, the nodes that rely on our learning rule do not need to keep track of their neighbors' reported observations, but only the beliefs.

APPENDIX

A. Proof of Theorem 1

We begin with the following recursion for each node i and $k \in [M-1]$:

$$\begin{aligned}
&\log \frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} \\
&= \sum_{j=1}^{n} W_{ij} \log \frac{b_j^{(t)}(\theta_M)}{b_j^{(t)}(\theta_k)} \\
&= \sum_{j=1}^{n} W_{ij} \left(\log \frac{f_j\left(X_j^{(t)}; \theta_M\right)}{f_j\left(X_j^{(t)}; \theta_k\right)} + \log \frac{q_j^{(t-1)}(\theta_M)}{q_j^{(t-1)}(\theta_k)} \right),
\end{aligned} \tag{55}$$

where the first and the second equalities follow from (3) and (2), respectively. Now for each node j we rewrite $\log \frac{q_j^{(\cdot)}(\theta_M)}{q_j^{(\cdot)}(\theta_k)}$ in terms of node j's neighbors and their samples at the previous instants. We can expand in this way until we express everything in terms of the samples collected and the initial estimates. Noting that $W^t(i,j) = \sum_{i_{t-1}=1}^{n} \cdots \sum_{i_1=1}^{n} W_{ii_1} \ldots W_{i_{t-1}j}$, it is easy to check that (55) can be further expanded to obtain the following:

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t} \log \frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=1}^{t} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)} \\
&\quad + \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} W^t(i,j) \log \frac{q_j^{(0)}(\theta_M)}{q_j^{(0)}(\theta_k)}.
\end{aligned} \tag{56}$$

From Assumption 3, the prior $q_j^{(0)}(\theta_k)$ is strictly positive for every node j and every $k \in [M]$. Since $W^t(i,j) \leq 1$, we have

$$\lim_{t\to\infty} \frac{1}{t} \left\{ \sum_{j=1}^{n} W^t(i,j) \log \frac{q_j^{(0)}(\theta_M)}{q_j^{(0)}(\theta_k)} \right\} = 0. \tag{57}$$

Let W be periodic with period d. If W is aperiodic, then the same proof still holds by putting $d = 1$. Now, we fix node i as a reference node and for every $r \in [d]$, define

$$A_r = \{j \in [n] : W^{md+r}(i,j) > 0 \text{ for some } m \in \mathbb{N}\}.$$

In particular, $(A_1, A_2, \ldots, A_d)$ is a partition of $[n]$; these sets form cyclic classes of the Markov chain. Fact 1 implies that for every $\delta > 0$, there exists an integer N which is function of δ alone, such that for all $m \geq N$, for some fixed $r \in [d]$, if $j \in A_r$, then

$$\left| W^{md+r}(i,j) - v_j d \right| \leq \delta \tag{58}$$

and if $j \notin A_r$

$$0 \leq W^{md+r}(i,j) \leq \delta. \tag{59}$$

Using this the first term in (56) can be decomposed as follows

$$\begin{aligned}
&\lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=1}^{t} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)} \\
&= \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=1}^{Nd-1} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)} \\
&\quad + \lim_{t\to\infty} \frac{1}{t} \sum_{j=1}^{n} \sum_{\tau=Nd}^{t} W^{\tau}(i,j) \log \frac{f_j\left(X_j^{(t-\tau+1)}; \theta_M\right)}{f_j\left(X_j^{(t-\tau+1)}; \theta_k\right)}.
\end{aligned} \tag{60}$$

Using the triangle inequality and the fact that $W^{\tau}(i,j) \leq 1$ for every $\tau \in \mathbb{N}$ we have

$$\left|\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau)};\theta_M\right)}{f_j\left(X_j^{(t-\tau)};\theta_k\right)}\right|$$
$$\leq \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{Nd-1}\left|\log\frac{f_j\left(X_j^{(t-\tau)};\theta_M\right)}{f_j\left(X_j^{(t-\tau)};\theta_k\right)}\right|.$$

For every $j \in [n]$, $\log\frac{f_j(X_j;\theta_M)}{f_j(X_j;\theta_k)}$ is integrable, implying $\left|\log\frac{f_j(X_j;\theta_M)}{f_j(X_j;\theta_k)}\right|$ is almost surely finite. This implies that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau)};\theta_M\right)}{f_j\left(X_j^{(t-\tau)};\theta_k\right)} = 0 \quad \text{P-a.s.} \tag{61}$$

Using (57) and (61), (60) becomes

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}$$
$$= \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=Nd}^{t}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}$$

with probability one. It is straightforward to see that the above equation can be rewritten as

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}$$
$$= \lim_{T\to\infty}\frac{1}{Td}\sum_{j=1}^{n}\sum_{m=N}^{T-1}\left\{\sum_{r=1}^{d}W^{md+r}(i,j)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$

with probability one. For every $\delta > 0$ and N such that for all $m \in N$ equations (58) and (59) hold true, using Lemma 1 we get that

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}$$

with probability one lies in the interval with end points

$$K(\theta_M,\theta_k) - \frac{\delta}{d}\sum_{j=1}^{n}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right]$$

and

$$K(\theta_M,\theta_k) + \frac{\delta}{d}\sum_{j=1}^{n}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right].$$

Since this holds for any $\delta > 0$, we have

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} = K(\theta_M,\theta_k) \quad \text{P-a.s.}$$

Hence, with probability one, for every $\epsilon > 0$ there exists a time T' such that $\forall t \geq T'$, $\forall k \in [M-1]$ we have

$$\left|\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} - K(\theta_M,\theta_k)\right| \leq \epsilon,$$

which implies

$$\frac{1}{1+\sum_{k\in[M-1]}e^{-K(\theta_M,\theta_k)t+\epsilon t}} \leq q_i^{(t)}(\theta_M) \leq 1.$$

Hence we have the assertion of the theorem.

Lemma 1. *For a given $\delta > 0$ and for some $N \in \mathbb{N}$ for which* (58) *and* (59) *hold true for all $m \geq N$, the following expression*

$$\lim_{T\to\infty}\frac{1}{Td}\sum_{j=1}^{n}\sum_{m=N}^{T-1}\left\{\sum_{r=1}^{d}W^{md+r}(i,j)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$

with probability one lies in an interval with end points

$$K(\theta_M,\theta_k) - \frac{\delta}{d}\sum_{j=1}^{n}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right],$$

and

$$K(\theta_M,\theta_k) + \frac{\delta}{d}\sum_{j=1}^{n}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right].$$

Proof: To the given expression we add and subtract $v_j d$ from $W^{md+r}(i,j)$ for all $j \in A_r$ to obtain

$$\lim_{T\to\infty}\frac{1}{Td}\sum_{j=1}^{n}\sum_{m=N}^{T-1}\left\{\sum_{r=1}^{d}W^{md+r}(i,j)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$
$$= \sum_{r=1}^{d}\sum_{j\notin A_r}\left\{\lim_{T\to\infty}\frac{1}{Td}\sum_{m=N}^{T-1}W^{md+r}(i,j)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$

$$+\sum_{r=1}^{d}\sum_{j\in A_r}\left\{\lim_{T\to\infty}\frac{1}{Td}\sum_{m=N}^{T-1}\left(W^{md+r}(i,j)-v_jd\right)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$
$$+\sum_{r=1}^{d}\sum_{j\in A_r}\left\{\lim_{T\to\infty}\frac{1}{Td}\sum_{m=N}^{T-1}v_jd\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}. \tag{62}$$

For each r and some $j\in A_r$, using (58) and the strong law of large numbers we have

$$\left|\lim_{T\to\infty}\frac{1}{Td}\left\{\sum_{m=N}^{T-1}\left(W^{md+r}(i,j)-v_jd\right)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}\right|$$
$$\leq\frac{\delta}{d}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right]\quad \text{P-a.s..}$$

Similarly for $j\notin A_r$, using (59) we have

$$\left|\lim_{T\to\infty}\frac{1}{Td}\sum_{m=N}^{T-1}W^{md+r}(i,j)\times \log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|$$
$$\leq\frac{\delta}{d}\mathbb{E}\left[\left|\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right|\right]\quad \text{P-a.s..}$$

Again, by the strong law of large numbers we have

$$\sum_{r=1}^{d}\sum_{j\in A_r}v_j\left\{\lim_{T\to\infty}\frac{1}{T}\sum_{m=N}^{T-1}\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right\}$$
$$=\sum_{r=1}^{d}\sum_{j\in A_r}v_j\mathbb{E}\left[\log\frac{f_j\left(X_j;\theta_M\right)}{f_j\left(X_j;\theta_k\right)}\right]$$
$$=K(\theta_M,\theta_k)\quad \text{P-a.s..}$$

Now combining this with (62) we have the assertion of the lemma.

■

B. Proof of Theorem 2

Recall the following equation:

$$\lim_{t\to\infty}\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}$$
$$=\lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}$$
$$+\lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=Nd}^{t}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}, \tag{63}$$

where N is such that for all $m\geq N, m\in\mathbb{N}$ equations (58) and (59) are satisfied. For any fixed t, using Assumption 4, the first term in the summation on the right hand side of (63) can be bounded as

$$\left|\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{Nd-1}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)}\right|\leq\frac{nNdL}{t}.$$

Also, the second term in the summation on the right hand side of (63) can be bounded as

$$\left|\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=Nd}^{t}W^{\tau}(i,j)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta_M\right)}{f_j\left(X_j^{(t-\tau+1)};\theta_k\right)} -\sum_{r=1}^{d}\sum_{j\in A_r}\frac{v_j}{Td}\sum_{m=0}^{T-1}\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|$$
$$\leq\delta\frac{1}{Td}\sum_{m=0}^{T-1}\left|\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|.$$

Using Assumption 4 we have

$$\frac{1}{Td}\sum_{m=0}^{T-1}\left|\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|\leq\frac{L}{d}.$$

Therefore, we have

$$\left|\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} -\sum_{r=1}^{d}\sum_{j\in A_r}\frac{v_j}{Td}\sum_{m=0}^{T-1}\log\frac{f_j\left(X_j^{(Td-md-r+1)};\theta_M\right)}{f_j\left(X_j^{(Td-md-r+1)};\theta_k\right)}\right|$$
$$\leq\frac{\delta nL}{d}.$$

Applying Hoeffding's inequality (Theorem 2 of [45]), for every $0<\epsilon\leq K(\theta_M,\theta_k)$, we can write (63) for $t\geq Nd$ as

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)}\leq K(\theta_M,\theta_k)-\epsilon+o\left(\frac{1}{t},\delta\right),$$

with probability at most $\exp\left(-\frac{\epsilon^2 T}{2L^2}\right)$ where $o\left(\frac{1}{t},\delta\right) = \frac{\delta nL}{d} + \frac{nNdL}{t}$. Similarly, for $0 < \epsilon \le L - K(\theta_M,\theta_k)$ we have

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} \ge K(\theta_M,\theta_k) + \epsilon + o\left(\frac{1}{t},\delta\right),$$

with probability at most $\exp\left(-\frac{\epsilon^2 T}{2L^2}\right)$ and for $\epsilon > L - K(\theta_M,\theta_k)$ we have

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta_M)}{q_i^{(t)}(\theta_k)} \ge K(\theta_M,\theta_k) + \epsilon + o\left(\frac{1}{t},\delta\right),$$

with probability 0. Now, taking limit and letting δ go to zero, for $0 < \epsilon \le K(\theta_M,\theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \le K(\theta_M,\theta_k) - \epsilon\right) \le -\frac{\epsilon^2}{2L^2 d},$$

for $0 < \epsilon \le L - K(\theta_M,\theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \ge K(\theta_M,\theta_k) + \epsilon\right) \le -\frac{\epsilon^2}{2L^2 d},$$

and for $\epsilon > L - K(\theta_M,\theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \ge K(\theta_M,\theta_k) + \epsilon\right) = -\infty.$$

Since $q_i^{(t)}(\theta_M) \le 1$, all the events ω which lie in the set $\{\omega : \rho_i^{(t)}(\theta_k) \le K(\theta_M,\theta_k) - \epsilon\}$ also lie in the set $\{\omega : \rho_i^{(t)}(\theta_k) \le K(\theta_M,\theta_k) - \epsilon + \rho_i^{(t)}(\theta_M)\}$. Hence, for every $0 < \epsilon \le K(\theta_M,\theta_k)$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \le K(\theta_M,\theta_k) - \epsilon\right) \le -\frac{\epsilon^2}{2L^2 d}. \quad (64)$$

For $k \in [M-1]$ and any $\alpha \ge 0$, the set

$$\left\{\rho_i^{(t)}(\theta_k) \ge K(\theta_M,\theta_k) + \epsilon\right\}$$

lies in the complement of the following set:

$$\left\{\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) < K(\theta_M,\theta_k) + \epsilon - \alpha\right\} \cap \left\{\rho_i^{(t)}(\theta_M) < \alpha\right\}.$$

This implies that

$$\begin{aligned}&\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \ge K(\theta_M,\theta_k) + \epsilon\right)\\ &\le \mathsf{P}\left(\rho_i^{(t)}(\theta_k) - \rho_i^{(t)}(\theta_M) \ge K(\theta_M,\theta_k) + \epsilon - \alpha\right)\\ &\quad + \mathsf{P}\left(\rho_i^{(t)}(\theta_M) \ge \alpha\right). \quad (65)\end{aligned}$$

Using Lemma 2 we have that for every $\delta > 0$ there exists a T such that for all $t \ge T$

$$\begin{aligned}&\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \ge K(\theta_M,\theta_k) + \epsilon\right)\\ &\le \exp\left(-\frac{(\epsilon-\alpha)^2}{2L^2 d}t + \delta t\right) \quad (66)\\ &\quad + \exp\left(-\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}t + \delta t\right). \quad (67)\end{aligned}$$

Taking the limit as $\alpha \to 0^+$ for $0 < \epsilon \le L - K(\theta_M,\theta_k)$ we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \ge K(\theta_M,\theta_k) + \epsilon\right)\\ &\le -\frac{1}{2L^2 d}\min\left\{\epsilon^2, \min_{j\in[M-1]} K^2(\theta_M,\theta_j)\right\}. \quad (68)\end{aligned}$$

For $\epsilon \ge L - K(\theta_M,\theta_k)$ we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \ge K(\theta_M,\theta_k) + \epsilon\right)\\ &\le -\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}. \quad (69)\end{aligned}$$

Lemma 2. *For all $\alpha > 0$, we have the following for the sequence $q_i^{(t)}(\theta_M)$*

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\rho_i^{(t)}(\theta_M) \ge \alpha\right)\\ &\le -\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2 d}\right\}. \quad (70)\end{aligned}$$

Proof: For any $\alpha > 0$, consider

$$\begin{aligned}&\mathsf{P}\left(\rho_i^{(t)}(\theta_M) \ge \alpha\right)\\ &\le \sum_{k\in[M-1]} \mathsf{P}\left(\frac{1}{M-1}\left(1 - e^{-\alpha t}\right) \le q_i^{(t)}(\theta_k)\right)\\ &= \sum_{k\in[M-1]} \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \le K(\theta_M,\theta_k) - \eta_t(\theta_k)\right), \quad (71)\end{aligned}$$

where $\eta_t(\theta_k) = K(\theta_M,\theta_k) - \frac{1}{t}\log(M-1) + \frac{1}{t}\log\left(1 - e^{-\alpha t}\right)$. For every $\epsilon > 0$, there exists $T(\epsilon)$ such that for all $t \ge T(\epsilon)$ we have

$$\begin{aligned}&\mathsf{P}\left(\rho_i^{(t)}(\theta_k) \ge \alpha\right)\\ &\le \sum_{k\in[M-1]} \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \le K(\theta_M,\theta_k) - K(\theta_M,\theta_k) + \epsilon\right)\\ &= \sum_{k\in[M-1]} \mathsf{P}\left(\rho_i^{(t)}(\theta_k) \le \epsilon\right).\end{aligned}$$

Therefore, for every $\epsilon > 0$, $\delta > 0$, there exists $T = \max\{T(\epsilon), T(\delta)\}$ such that for all $t \ge T$ we have

$$\begin{aligned}&\mathsf{P}\left(\rho_i^{(t)}(\theta_M) \ge \alpha\right)\\ &\le (M-1)\max_{k\in[M-1]}\exp\left\{-\frac{(K(\theta_M,\theta_k) - \epsilon)^2}{2L^2 d}t + \delta t\right\}.\end{aligned}$$

By taking the limit and making ϵ arbitrarily small, we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\rho_i^{(t)}(\theta_M)\geq\alpha\right)\\&\leq-\min_{k\in[M-1]}\left\{\frac{K(\theta_M,\theta_k)^2}{2L^2d}\right\}.\end{aligned}$$

■

1) Proof of Corollary 3: From Theorem 2, we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\mu_i\geq\min_{k\in[M-1]}K(\theta_M,\theta_k)+\epsilon\right)\\&\leq-\frac{1}{2L^2d}\min\left\{\epsilon^2,\min_{k\in[M-1]}K(\theta_M,\theta_k)^2\right\}.\end{aligned}$$

Now, applying the Borel-Cantelli Lemma to the above equation we have

$$\mu_i\leq\min_{k\in[M-1]}K(\theta_M,\theta_k)\quad\mathsf{P}\text{-a.s.}$$

Letting $\epsilon\to0$ and by combining this with Corollary 1 we have

$$\mu_i=\min_{k\in[M-1]}K(\theta_M,\theta_k)\quad\mathsf{P}\text{-a.s.}$$

C. Proof of Theorem 3

Fact 2 (Cramer's Theorem, Theorem 3.8 [46])**.** *Consider a sequence of d-dimensional i.i.d random vectors $\{\mathbf{X}_n\}_{n=1}^{\infty}$. Let $\mathbf{S}_n=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i$. Then, the sequence of $\mathbf{S}_n$ satisfies a large deviation principle with rate function $\Lambda^*(\cdot)$, namely: For any set $F\subset\mathbb{R}^d$,*

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathsf{P}(\mathbf{S}_n\in F)\geq-\inf_{\mathbf{x}\in F^o},\tag{72}$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathsf{P}(\mathbf{S}_n\in F)\leq-\inf_{\mathbf{x}\in\bar{F}},\tag{73}$$

where $\Lambda^*(\cdot)$ is given by

$$\Lambda^*(\mathbf{x})\overset{\triangle}{=}\sup_{\boldsymbol{\lambda}\in\mathbb{R}^d}\left\{\langle\boldsymbol{\lambda},\mathbf{x}\rangle-\Lambda(\boldsymbol{\lambda})\right\}.\tag{74}$$

and $\Lambda(\cdot)$ is the log moment generating function of $\mathbf{S}_n$ which is given by

$$\Lambda(\boldsymbol{\lambda})\overset{\triangle}{=}\log\mathbb{E}[e^{\langle\boldsymbol{\lambda},\mathbf{Y}\rangle}].\tag{75}$$

Fact 3 (Contraction Principle, Theorem 3.20 [46])**.** *Let $\{\mathsf{P}_t\}$ be a sequence of probability measures on a Polish space $\mathcal{X}$ that satisfies LDP with rate function I. Let*

$$\left\{\begin{array}{ll}\mathcal{Y} & \textit{be a Polish space}\\ T:\mathcal{X}\to\mathcal{Y} & \textit{a continuous map}\\ \mathsf{Q}_t=\mathsf{P}_t\circ T^{-1} & \textit{an image probability measure.}\end{array}\right.\tag{76}$$

Then $\{\mathsf{Q}_t\}$ satisfies the LDP on $\mathcal{Y}$ with rate function J given by

$$J(y)=\inf_{x\in\mathcal{X}:T(x)=y}I(x).\tag{77}$$

To prove that $\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}$ satisfies the LDP, first we establish the LDP satisfied by the following vector:

$$\mathbf{Q}_i^{(t)}=\left[\frac{q_i^{(t)}(\theta_1)}{q_i^{(t)}(\theta_M)},\frac{q_i^{(t)}(\theta_2)}{q_i^{(t)}(\theta_M)},\ldots,\frac{q_i^{(t)}(\theta_{M-1})}{q_i^{(t)}(\theta_M)}\right]^T.\tag{78}$$

Note that $\mathbf{Q}_i^{(t)}=\frac{\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}}{q_i^{(t)}(\theta_M)}$. From Lemma 3, we obtain that $\frac{1}{t}\log\mathbf{Q}_i^{(t)}$ satisfies the LDP with rate function $I(\cdot)$, as given by (37). Now we apply the Contraction Principle (Fact 3), for

$$\begin{aligned}&\mathcal{X}=\mathbb{R}^{M-1},\quad\mathcal{Y}=\mathbb{R}^{M-1},\\&T(\mathbf{x})=g(\mathbf{x}),\quad\forall\mathbf{x}\in\mathbb{R}^{M-1},\\&\mathsf{P}_t=\mathsf{P}\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\in\cdot\right),\\&\mathsf{Q}_t=\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)\in\cdot\right),\end{aligned}$$

and we get that $g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)$ satisfies an LDP with a rate function $J(\cdot)$, i.e., for every $F\subset\mathbb{R}^{M-1}$ we have

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)\in F\right)\geq-\inf_{\mathbf{y}\in F^o}J(\mathbf{y}),\tag{79}$$

and

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)\in F\right)\leq-\inf_{\mathbf{y}\in\bar{F}}J(\mathbf{y}).\tag{80}$$

Combining Lemma 4 with (79) and (80), we obtain that $\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}$ satisfies the LDP with rate function $J(\cdot)$ as well. Hence, we have the assertion of the theorem.

Lemma 3. *The random vector $\frac{1}{t}\log\mathbf{Q}_i^{(t)}$ satisfies the LDP with rate function given by $I(\cdot)$ in* (36)*. That is, for any set $F\subset\mathbb{R}^{M-1}$ with interior F^o and closure $\bar{F}$, we have*

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\in F\right)\geq-\inf_{\mathbf{x}\in F^o}I(\mathbf{x}),\tag{81}$$

and

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\in F\right)\leq-\inf_{\mathbf{x}\in\bar{F}}I(\mathbf{x}).\tag{82}$$

Proof: Using the learning rule we have

$$\begin{aligned}\frac{1}{t}\log\mathbf{Q}_i^{(t)}&=\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}W^{\tau}(i,j)\mathbf{L}_j^{(t-\tau+1)}\\&=\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}\left(W^{\tau}(i,j)-v_j\right)\mathbf{L}_j^{(t-\tau+1)}\\&\quad+\frac{1}{t}\sum_{\tau=1}^{t}\mathbf{Y}^{(\tau)},\end{aligned}\tag{83}$$

where $\mathbf{L}$ is given by (33) and $\mathbf{Y}$ by (32). Using Cramer's Theorem (Fact 2) in $\mathbb{R}^{M-1}$, for any set $F \subset \mathbb{R}^{M-1}$, we have

$$\liminf_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} \mathbf{Y}^{(\tau)} \in F\right) \geq -\inf_{\mathbf{x}\in F^{o}} I(\mathbf{x}), \quad (84)$$

and

$$\limsup_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\sum_{\tau=1}^{t} \mathbf{Y}^{(\tau)} \in F\right) \leq -\inf_{\mathbf{x}\in \bar{F}} I(\mathbf{x}). \quad (85)$$

Consider

$$\begin{aligned} &\left|\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}\left(W^{\tau}(i,j)-v_j\right)\mathbf{L}_j^{(t-\tau+1)}\right| \\ &\leq \frac{n}{t}\sum_{\tau=1}^{t}\left|\lambda_{\max}^{\tau}(W)\right|\left(\sum_{j=1}^{n}\left|\mathbf{L}_j^{(t-\tau+1)}\right|\right). \end{aligned} \quad (86)$$

From Assumption 5, we have that $\Lambda(\boldsymbol{\lambda})$ is finite for $\boldsymbol{\lambda} \in \mathbb{R}^n$. Now, using Lemma 5, we have

$$\begin{aligned} \lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\left|\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}\left(W^{\tau}(i,j)-v_j\right)\mathbf{L}_j^{(t-\tau+1)}\right| \geq \boldsymbol{\delta}\right) \\ = -\infty. \end{aligned} \quad (87)$$

Using Lemma 6 on $\frac{1}{t}\log\mathbf{Q}_i^{(t)}$, we have the assertion of the theorem. ■

Lemma 4. *For every set $F \subset \mathbb{R}^{M-1}$ and for all $i \in [n]$, we have*

$$\begin{aligned} &\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \\ &\geq \liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \in F\right), \end{aligned} \quad (88)$$

and

$$\begin{aligned} &\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \\ &\leq \limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \in F\right). \end{aligned} \quad (89)$$

Proof: For all $t \geq 0$, we have

$$\begin{aligned} &\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} = g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \\ &-\frac{1}{t}\log\left(e^{-C^{(t)}t}+\sum_{j=1}^{M-1}e^{g_j\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)t}\right)\mathbf{1}, \end{aligned} \quad (90)$$

where

$$\begin{aligned} C^{(t)} = \max\Bigg\{0, \frac{1}{t}\log\frac{q_i^{(t)}(\theta_1)}{q_i^{(t)}(\theta_M)}, \frac{1}{t}\log\frac{q_i^{(t)}(\theta_2)}{q_i^{(t)}(\theta_M)}, \\ \ldots, \frac{1}{t}\log\frac{q_i^{(t)}(\theta_{M-1})}{q_i^{(t)}(\theta_M)}\Bigg\}. \end{aligned}$$

Also for all $t \geq 0$, we have

$$1 \leq e^{-C^{(t)}t}+\sum_{j=1}^{M-1}e^{g_j\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)t} \leq M.$$

Hence for all $\epsilon > 0$, there exists $T(\epsilon)$ such that for all $t \geq T(\epsilon)$ we have

$$g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right)-\epsilon\mathbf{1} \leq \frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \leq g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right). \quad (91)$$

For any $F \subset \mathbb{R}^{M-1}$, let $F_{\epsilon^{+}} = \{\mathbf{x}+\delta\mathbf{1}, \forall\, 0 < \delta \leq \epsilon \text{ and } \mathbf{x} \in F\}$, $F_{\epsilon^{-}} = \{\mathbf{x}-\delta\mathbf{1}, \forall\, 0 < \delta \leq \epsilon \text{ and } \mathbf{x} \in F\}$. Therefore, for every $\epsilon > 0$ we have

$$\begin{aligned} &\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \in F\right) \\ &\leq \liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F_{\epsilon^{-}}\right). \end{aligned} \quad (92)$$

Making ϵ arbitrarily small, $F_{\epsilon^{-}} \to F$, and by monotonicity and continuity of probability measure we have

$$\begin{aligned} &\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \in F\right) \\ &\leq \liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right). \end{aligned} \quad (93)$$

For $t \geq T(\epsilon)$ we also have

$$\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \leq g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \leq \frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})}+\epsilon\mathbf{1}. \quad (94)$$

This implies for every $\epsilon > 0$ we have

$$\begin{aligned} &\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \\ &\leq \limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \in F_{\epsilon^{+}}\right). \end{aligned} \quad (95)$$

Again, by making ϵ arbitrarily small we have

$$\begin{aligned} &\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log\tilde{\mathbf{q}}_{\mathbf{i}}^{(\mathbf{t})} \in F\right) \\ &\leq \limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(g\left(\frac{1}{t}\log\mathbf{Q}_i^{(t)}\right) \in F\right). \end{aligned} \quad (96)$$

Hence, we have the assertion of the lemma. ■

D. Proof of the Lemmas

Lemma 5. *Let q be a real number such that $q \in (0,1)$. Let $\mathbf{X}_i$ be a sequence of non-negative i.i.d random vectors in $\mathbb{R}^n$, distributed as $\mathbf{X}$ and let $\Lambda(\boldsymbol{\lambda})$ denote its log moment generating function which is finite for $\boldsymbol{\lambda} \in \mathbb{R}^n$, then for every $\delta > 0$, we have*

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\sum_{i=1}^{t}(q)^{i}\mathbf{X}_i \geq \delta\mathbf{1}\right) = -\infty. \quad (97)$$

Proof: Applying Chebychev's inequality and using the definition of log moment generating function, for $\boldsymbol{\lambda} \in \mathbb{R}^n$, we have

$$\mathsf{P}\left(\frac{1}{t}\sum_{i=1}^{t}(q)^i\mathbf{X}_i \geq \delta\mathbf{1}\right) \leq e^{-t\left(\langle\boldsymbol{\lambda},\delta\mathbf{1}\rangle - \frac{1}{t}\sum_{i=1}^{t}\Lambda((q)^i\boldsymbol{\lambda})\right)}. \tag{98}$$

From convexity of Λ, we have $\sum_{i=1}^{t}\Lambda((q)^i\boldsymbol{\lambda}) \leq \Lambda(\boldsymbol{\lambda})\sum_{i=1}^{t}(q)^i$. Since $\Lambda(\boldsymbol{\lambda})$ is finite and $\sum_{i=1}^{\infty}(q)^i < \infty$, for all $\delta > 0$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\sum_{i=1}^{t}(q)^i\mathbf{X}_i \geq \delta\mathbf{1}\right) \leq -\langle\boldsymbol{\lambda},\delta\mathbf{1}\rangle. \tag{99}$$

Since, the above equation is true for all $\boldsymbol{\lambda} \in \mathbb{R}^n$, we have the assertion of the lemma. ■

Lemma 6. *Consider a sequence* $\{\mathbf{Z}^{(t)}\}_{t=0}^{\infty}$ *where* $\mathbf{Z}^{(t)} \in \mathbb{R}^d$ *such that*

$$\mathbf{Z}^{(t)} = \mathbf{X}^{(t)} + \mathbf{Y}^{(t)}, \tag{100}$$

where sequences $\{\mathbf{X}^{(t)}\}_{t=0}^{\infty}$ *and* $\{\mathbf{Y}^{(t)}\}_{t=0}^{\infty}$ *have the following properties:*

1) *The sequence* $\{\mathbf{X}^{(t)}\}_{t=0}^{\infty}$ *satisfies*

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) \geq -\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x}), \tag{101}$$

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) \leq -\inf_{\mathbf{x}\in\bar{F}} I_X(\mathbf{x}), \tag{102}$$

where $I_X : \mathbb{R}^d \to \mathbb{R}$ *is a well-defined LDP rate function.*

2) *For every* $\epsilon > 0$*, sequence* $\{\mathbf{Y}^{(t)}\}_{t=0}^{\infty}$ *satisfies*

$$\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}(|\mathbf{Y}^{(t)}| \geq \epsilon\mathbf{1}) = -\infty. \tag{103}$$

Then $\{\mathbf{Z}^{(t)}\}_{t=0}^{\infty}$ *satisfies*

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}(\mathbf{Z}^{(t)} \in F) \geq -\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x}), \tag{104}$$

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathsf{P}(\mathbf{Z}^{(t)} \in F) \leq -\inf_{\mathbf{x}\in\bar{F}} I_X(\mathbf{x}). \tag{105}$$

Proof: For every $t \geq 0$, we have

$$\begin{aligned}&\mathsf{P}\left(\mathbf{Z}^{(t)} \in F_{\epsilon^+} \cup F_{\epsilon^-}\right)\\ &\geq \mathsf{P}\left(\{\mathbf{X}^{(t)} \in F\} \cap \{|\mathbf{Y}^{(t)}| \leq \epsilon\mathbf{1}\}\right)\\ &\geq \mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) - \mathsf{P}\left(|\mathbf{Y}^{(t)}| > \epsilon\mathbf{1}\right).\end{aligned}$$

For all $\delta > 0$, there exists a $T(\delta)$ such that for all $t \geq T(\delta)$ we have

$$\mathsf{P}\left(\mathbf{X}^{(t)} \in F\right) \geq e^{-\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x})t - \delta t}.$$

For all $B > 0$, there exists a $T(B)$ such that for all $t \geq T(B)$ we have

$$\mathsf{P}\left(|\mathbf{Y}^{(t)}| > \epsilon\mathbf{1}\right) \geq e^{-Bt}.$$

Now choose $B > \inf_{\mathbf{x}\in F^o} I_X(\mathbf{x}) + \delta$ and $t \geq \max\{T(\delta), T(B)\}$, then we have

$$\begin{aligned}&\mathsf{P}\left(\mathbf{Z}^{(t)} \in F_{\epsilon^+} \cup F_{\epsilon^-}\right)\\ &\geq e^{-\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x})t - \delta t}\left(1 - e^{-Bt + \inf_{\mathbf{x}\in F^o} I_X(\mathbf{x})t + \delta t}\right).\end{aligned}$$

Sending ϵ to zero and taking the limit we have

$$\liminf_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\mathbf{Z}^{(t)} \in F\right) \geq -\inf_{\mathbf{x}\in F^o} I_X(\mathbf{x}).$$

Similarly, using the fact that $\mathsf{P}(\{\mathbf{Z}^{(t)} \in F\} \cap \{|\mathbf{Y}^{(t)}| \leq \epsilon\mathbf{1}\}) \leq \mathsf{P}\left(\mathbf{X}^{(t)} \in F_{\epsilon^+}\right)$ we have the other LDP bound. ■

References

[1] A. Lalitha, A. Sarwate, and T. Javidi, "Social learning and distributed hypothesis testing," in *Proceedings of the 2014 IEEE International Symposium on Information Theory*, June 2014, pp. 551–555. [Online]. Available: http://dx.doi.org/10.1109/ISIT.2014.6874893

[2] A. Lalitha and T. Javidi, "On the rate of learning in distributed hypothesis testing," in *Proceedings of the 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, September 2015, pp. 1–8. [Online]. Available: http://dx.doi.org/10.1109/ALLERTON.2015.7446979

[3] ——, "Large deviation analysis for learning rate in distributed hypothesis testing," in *Proceedings of the 49th Asilomar Conference on Signals, Systems and Computers*, November 2015, pp. 1065–1069. [Online]. Available: http://dx.doi.org/10.1109/ACSSC.2015.7421302

[4] R. Ahlswede and I. Csiszar, "Hypothesis testing with communication constraints," *IEEE Transactions on Information Theory*, vol. 32, no. 4, pp. 533–542, July 1986. [Online]. Available: http://dx.doi.org/10.1109/TIT.1986.1057194

[5] T. Han, "Hypothesis testing with multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 33, no. 6, pp. 759–772, November 1987. [Online]. Available: http://dx.doi.org/10.1109/TIT.1987.1057383

[6] M. Longo, T. D. Lookabaugth, and R. M. Gray, "Quantization for decentralized hypothesis testing under communication constraints," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 241–255, March 1990. [Online]. Available: http://dx.doi.org/10.1109/18.52470

[7] V. V. Veeravalli, T. Basar, and H. V. Poor, "Decemberentralized sequential detection with a fusion center performing the sequential test," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 433–442, March 1993. [Online]. Available: http://dx.doi.org/10.1109/18.212274

[8] H. Shimokawa, T. S. Han, and S. Amari, "Error bound of hypothesis testing with data compression," in *Proceedings of the 1994 IEEE International Symposium on Information Theory*, June 1994, pp. 114–119. [Online]. Available: http://dx.doi.org/10.1109/ISIT.1994.394874

[9] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2300–2324, October 1998. [Online]. Available: http://dx.doi.org/10.1109/18.720540

[10] Y. Xiang and Y. H. Kim, "Interactive hypothesis testing against independence," in *Proceedings of the 2013 IEEE International Symposium on Information Theory Proceedings*, July 2013, pp. 2840–2844. [Online]. Available: http://dx.doi.org/10.1109/ISIT.2013.6620744

[11] Y. Mei, "Asymptotic optimality theory for decemberentralized sequential hypothesis testing in sensor networks," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2072–2089, May 2008. [Online]. Available: http://dx.doi.org/10.1109/TIT.2008.920217

[12] M. S. Rahman and A. B. Wagner, "On the optimality of binning for distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6282–6303, October 2012. [Online]. Available: http://dx.doi.org/10.1109/TIT.2012.2206793

[13] B. Chen, R. Jiang, T. Kasetkasem, and P. K. Varshney, "Channel aware decemberision fusion in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 52, no. 12, pp. 3454–3458, December 2004. [Online]. Available: http://dx.doi.org/10.1109/TSP.2004.837404

[14] B. Chen and P. K. Willett, "On the optimality of the likelihood-ratio test for local sensor decemberision rules in the presence of nonideal channels," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 693–699, February 2005. [Online]. Available: http://dx.doi.org/10.1109/TIT.2004.840879

[15] A. Anandkumar and L. Tong, "Distributed statistical inference using type based random access over multi-access fading channels," in *Proceedings of the 40th Annual Conference on Information Sciences and Systems*, March 2006, pp. 38–43. [Online]. Available: http://dx.doi.org/10.1109/CISS.2006.286427

[16] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "A collaborative training algorithm for distributed learning," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1856–1871, April 2009. [Online]. Available: http://dx.doi.org/10.1109/TIT.2009.2012992

[17] S. Kar, J. M. F. Moura, and H. V. Poor, "$\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1848–1862, April 2013. [Online]. Available: http://dx.doi.org/10.1109/TSP.2013.2241057

[18] D. Mosk-Aoyama and D. Shah, "Fast distributed algorithms for computing separable functions," *IEEE Transactions on Information Theory*, vol. 54, no. 7, pp. 2997–3007, July 2008. [Online]. Available: http://dx.doi.org/10.1109/TIT.2008.924648

[19] P. Bianchi, G. Fort, and W. Hachem, "Performance of a distributed stochastic approximation algorithm," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7405–7418, November 2013. [Online]. Available: http://dx.doi.org/10.1109/TIT.2013.2275131

[20] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, June 2006. [Online]. Available: http://dx.doi.org/10.1109/TIT.2006.874516

[21] F. Benezit, A. G. Dimakis, P. Thiran, and M. Vetterli, "Order-optimal consensus through randomized path averaging," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5150–5167, October 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2060050

[22] T. C. Aysal and K. E. Barner, "Convergence of consensus models with stochastic disturbances," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 4101–4113, August 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2050940

[23] Y. Yang and R. S. Blum, "Broadcast-based consensus with non-zero-mean stochastic perturbations," *IEEE Transactions on Information Theory*, vol. 59, no. 6, pp. 3971–3989, June 2013. [Online]. Available: http://dx.doi.org/10.1109/TIT.2013.2243816

[24] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, March 2010. [Online]. Available: http://dx.doi.org/10.1109/TSP.2009.2036046

[25] 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. [Online]. Available: https://doi.org/10.1016/j.geb.2012.06.001

[26] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *Proceedings of the 52nd Annual IEEE Conference on Decision and Control (CDC), 2013*, December 2013, pp. 6196–6201. [Online]. Available: http://dx.doi.org/10.1109/CDC.2013.6760868

[27] A. Jadbabaie, P. Molavi, and A. Tahbaz-salehi, "Information heterogeneity and the speed of learning in social networks," *Columbia Business School Research Paper*, no. 13-28, May 2013. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2266979

[28] K. Rahnama Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *Proceedings of the 49th IEEE Conference on Decision and Control*, December 2010, pp. 5050–5055. [Online]. Available: http://dx.doi.org/10.1109/CDC.2010.5717946

[29] 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. [Online]. Available: https://doi.org/10.1007/11533382_11

[30] 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

[31] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3256–3268, November 2016. [Online]. Available: http://dx.doi.org/10.1109/TAC.2015.2506903

[32] A. Nedić, A. Olshevsky, and C. A. Uribe, "Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs," in *Proceedings of the 2015 American Control Conference (ACC)*, July 2015, pp. 5884–5889. [Online]. Available: http://dx.doi.org/10.1109/ACC.2015.7172262

[33] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3575–3605, June 2012. [Online]. Available: http://dx.doi.org/10.1109/TIT.2012.2191450

[34] V. Saligrama, M. Alanyali, and O. Savas, "Distributed detection in sensor networks with packet losses and finite capacity links," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4118–4132, November 2006. [Online]. Available: http://dx.doi.org/10.1109/TSP.2006.880227

[35] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, "Distributed bayesian hypothesis testing in sensor networks," in *Proceedings of the 2004 American Control Conference*, vol. 6, June 2004, pp. 5369–5374 vol.6.

[36] M. Harel, E. Mossel, P. Strack, and O. Tamuz, "On the speed of social learning," *CoRR*, vol. abs/1412.7172, 2014. [Online]. Available: http://arxiv.org/abs/1412.7172

[37] M. Mueller-Frank, "A general framework for rational learning in social networks," *Theoretical Economics*, vol. 8, no. 1, pp. 1–40, 2013. [Online]. Available: http://dx.doi.org/10.3982/TE1015

[38] A. K. Sahu and S. Kar, "Distributed sequential detection for gaussian shift-in-mean hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 89–103, January 2016. [Online]. Available: http://dx.doi.org/10.1109/TSP.2015.2478737

[39] D. Bajovic, D. Jakovetic, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+innovations distributed detection with non-gaussian observations," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5987–6002, November 2012. [Online]. Available: http://dx.doi.org/10.1109/TSP.2012.2210885

[40] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York, NY, USA: Wiley-Interscience, 1991.

[41] A. G. Dimakis, S. Kar, J. M. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, November 2010. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2010.2052531

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

[43] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast Convergence Rates for Distributed Non-Bayesian Learning," *IEEE Transactions on Automatic Control*, vol. PP, no. 99, pp. 1–1, 2017. [Online]. Available: http://dx.doi.org/10.1109/TAC.2017.2690401

[44] S. Salehkalaibar, M. A. Wigger, and L. Wang, "Hypothesis testing in multi-hop networks," vol. abs/1708.05198, 2017. [Online]. Available: http://arxiv.org/abs/1708.05198

[45] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. pp. 13–30, 1963. [Online]. Available: http://www.jstor.org/stable/2282952

[46] F. den Hollander, *Large Deviations*, ser. Fields Institute Monographs. American Mathematical Society, 2000, vol. 14.