Fifty-third Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 29 - October 2, 2015

On the Rate of Learning in Distributed Hypothesis Testing

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

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

Abstract—**This paper considers a problem of distributed hypothesis testing and cooperative learning. Individual nodes in a network receive noisy local (private) observations whose distribution is parameterized by a discrete parameter (hypotheses). The conditional distributions are known locally at the nodes, but the true parameter/hypothesis is not known. We consider a social ("non-Bayesian") learning rule from previous literature, in which nodes first perform a Bayesian update of their belief (distribution estimate) of the parameter based on their local observation, communicate these updates to their neighbors, and then perform a "non-Bayesian" linear consensus using the log-beliefs of their neighbors. For this learning rule, we know that under mild assumptions, the belief of any node in any incorrect parameter converges to zero exponentially fast, and the exponential rate of learning is a characterized by the network structure and the divergences between the observations' distributions. Tight bounds on the probability of deviating from this nominal rate in aperiodic networks is derived. The bounds are shown to hold for all conditional distributions which satisfy a mild bounded moment condition.**

I. Introduction

We study a model in which a network of individuals sample local observations (over time) governed by an unknown true parameter θ^* taking values in a discrete set Θ. We model the i-th node's observation by i.i.d conditional distribution (or local observation kernel, or likelihood) $f_i\left(\cdot;\theta^*\right)$ from a collection $\{f_i\left(\cdot;\theta\right) : \theta \in \Theta\}$. When these local channels are not sufficient to recover the underlying parameter locally, individuals must share and learn from each other in order to accurately estimate the parameter. Even though each individual cannot identify the parameter through local observations alone, the parameter may be collectively identifiable.

In this paper we study the learning rule proposed in [1], which is based on local Bayesian updating followed by consensus averaging on a reweighting of the *log beliefs* of nodes. Under the assumptions of network-wide observability and connectivity, in [1], we showed that the rate of convergence of this learning rule is given by the network divergence, $K(\theta^*, \theta)$, given by the expression $\sum_{j=1}^{n} v_j D\left(f_j\left(\cdot;\theta^*\right) \| f_j\left(\cdot;\theta\right)\right)$, where the weights in the sum are the nodes' influences as dictated by the consensus algorithm.

Furthermore, Theorem 2 in [2], under the assumption that log-likelihood ratios are bounded, provides bounds on the probability of concentration of rate around $K(\theta^*, \theta)$. These bounds show that under the assumption of boundedness of log-likelihoods in [1], the probability of the rate deviating from $K(\theta^*, \theta)$ decays exponentially fast. In this paper for aperiodic networks, we further relax the assumption of bounded log-likelihood ratios and obtain stronger results for probability of concentration of learning rate. This new assumption is less stringent and allows us to consider a much larger family of likelihood functions such as that of subgaussians and subexponentials. Moreover, for appropriately small deviations from the nominal learning rate around the network divergence, we obtain a large deviation principle, where tight (and matching) upper and lower bounds have on the probability of the deviations has been characterized.

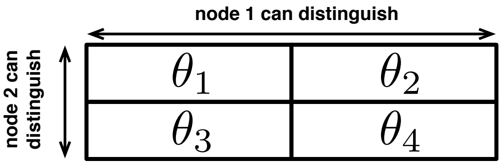

Fig. 1. Example of a parameter space in which no node can identify the true parameter. There are 4 parameters, $\{\theta_1, \theta_2, \theta_3, \theta_4\}$, and 2 nodes. The node 1 has $f_1\left(\cdot;\theta_1\right) = f_1\left(\cdot;\theta_3\right)$ and $f_1\left(\cdot;\theta_2\right) = f_1\left(\cdot;\theta_4\right)$,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

Jadbabaie et al. [3] provide an excellent overview of the large body of literature on distributed estimation, detection, hypothesis testing and learning over social networks. In [2], we have provided a comprehensive discussion of those most relevant to this paper [3]–[12]. In the interest of brevity, we refer the interested the reader to [2].

II. The Model

A. Nodes and Observations

The system consists of a set of n individual nodes. Let $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$ denote a finite set of M parameters which we call *hypotheses* and let each θ_i denotes a hypothesis. At every time instant t, each node $i \in [n]$ receives i.i.d (over time t) noisy observations $X_i^{(t)} \in \mathcal{X}_i$ of the system, where $\mathcal{X}_i$ denotes the *observation space* of node i. The observations are statistically governed by a fixed global "true hypothesis" $\theta^* \in \Theta$ which is unknown to the nodes. We consider the case where every node $i \in [n]$ knows its set

978-1-5090-1824-6/15/$31.00 ©2015 IEEE

of marginal conditional distributions $\{f_i(\cdot;\theta):\theta\in\Theta\}$, with $f_i(\cdot;\theta)$ denoting the distribution of $X_i^{(t)}$ conditioned on θ being the true hypothesis. The observations are independent across the nodes and over time. Furthermore, each node's observation sequence (in time) is conditionally independent and identically distributed (i.i.d).

In this setting, nodes can attempt to learn the "true hypothesis" θ^* using their knowledge of $\{f_i(\cdot;\theta):\theta\in\Theta\}$. It is not hard to see that if $f_i(\cdot;\theta)\neq f_i(\cdot;\theta^*)$, for some $\theta\neq\theta^*$, node i can exponentially rule out hypothesis θ in favor of θ^* with an exponent which is equal to $D\left(f_i(\cdot;\theta^*)\|\, f_i(\cdot;\theta)\right)$ [13, Section 11.7].

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

Also, note that this assumption does not require the existence of a single node that can distinguish θ^* from all other hypotheses. We only require that for every pair $\theta_i\neq\theta_j$, there is at least one node $k\in[n]$ for which $f_k(\cdot;\theta_i)\neq f_k(\cdot;\theta_j)$. We make the following additional assumption on the observations.

Assumption 2. *For every pair $\theta_i\neq\theta_j$ and every node $k\in[n]$, $\log\mathbb{E}\left[e^{\alpha\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}}\right]$, i.e., the log moment generating function of $\frac{f_k(\cdot;\theta_i)}{f_k(\cdot;\theta_j)}$ is finite for α in some interval containing* 0.

B. Network

The nodes are connected in a network which facilitates communication and enables collaboration. The network is modeled via a directed graph. 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.

Assumption 3. *The underlying graph of the network is strongly connected and aperiodic, 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.

C. The Learning Rule

We begin by defining a few variables required in order to define the learning rule. At every time instant t each node i maintains an estimate vector $\mathbf{q_i^{(t)}}\in\mathcal{P}(\Theta)$ and a belief vector $\mathbf{b_i^{(t)}}\in\mathcal{P}(\Theta)$, each of which is a probability distribution 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 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 estimate $\mathbf{q_i^{(0)}}$. At each time $t=1,2,\ldots$ the following events happen:

1) Each node i draws a conditionally iid observation $X_i^{(t)}\sim f_i(\cdot;\theta^*)$.
2) Each node i performs a local Bayesian update on $\mathbf{q_i^{(t-1)}}$ to form a belief $\mathbf{b_i^{(t)}}$ using the following rule. For each $\theta\in\Theta$,
$$b_i^{(t)}(\theta)=\frac{f_i\left(X_i^{(t)};\theta\right)q_i^{(t-1)}(\theta)}{\sum_{\theta'\in\Theta}f_i\left(X_i^{(t)};\theta'\right)q_i^{(t-1)}(\theta')}. \tag{1}$$
3) Each node i sends the message $\mathbf{Y_i^{(t)}}=\mathbf{b_i^{(t)}}$ to all nodes j for which $i\in\mathcal{N}(j)$ and similarly receives messages from its neighbors.
4) Each node i forms an estimate of θ, by averaging the log beliefs it received from its neighbors. For each $\theta\in\Theta$,
$$q_i^{(t)}(\theta)=\frac{\exp\left(\sum_{j=1}^{n}W_{ij}\log b_j^{(t)}(\theta)\right)}{\sum_{\theta'\in\Theta}\exp\left(\sum_{j=1}^{n}W_{ij}\log b_j^{(t)}(\theta')\right)}. \tag{2}$$

Note that the estimate vectors $\mathbf{q_i^{(t)}}$ remain locally with the nodes while their belief vectors $\mathbf{b_i^{(t)}}$ are exchanged with other nodes.

Along with the weights, the network can be thought of as a weighted strongly connected network. Hence, from Assumption 3, we have that weight matrix W is irreducible and aperiodic. In this context we recall the following fact.

Fact 1 **(Section 2.5 of Hoel et. al. [14]).** *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*
$$v_i=\sum_{j=1}^{n}v_jW_{ji}. \tag{3}$$
Furthermore, all components of $\mathbf{v}$ *are strictly positive. If the Markov chain is aperiodic, then*
$$\lim_{t\to\infty}W^t(i,j)=v_j,\quad i,j\in[n]. \tag{4}$$

In the social network literature such as Jadbabaie et al. [3], the eigenvector $\mathbf{v}$ is known as the eigenvector centrality; it is a measure of social influence of a node in the network. The objective of learning rule is to ensure that the estimate vector $\mathbf{q_i^{(t)}}$ of each node $i\in[n]$ converges to $1_{\theta^*}(\cdot)$. Note that our learning rule is such that if the initial estimate of any $\theta\in\Theta$ for some node is zero then estimates of that θ remains zero in subsequent time intervals. Hence, we make the following assumption.

Assumption 4. *For all $i \in [n]$, the initial estimate $q_i^{(0)}(\theta) > 0$ for every $\theta \in \Theta$.*

In this paper for brevity we assume that each node starts with a uniform estimate.

III. Main Result

Definition 1. For any $\theta \neq \theta^*$, the network divergence between θ^* and θ, denoted by $K(\theta^*, \theta)$, is defined as

$$K(\theta^*, \theta) \triangleq \sum_{j=1}^{n} v_j D\left(f_j\left(\cdot;\theta^*\right) \| \, f_j\left(\cdot;\theta\right)\right). \tag{5}$$

Definition 2. Let $Y^{(t)}(\theta^*, \theta) \triangleq \sum_{j=1}^{n} v_j \log \frac{f_j\left(X_j^{(t)};\theta\right)}{f_j\left(X_j^{(t)};\theta^*\right)}$. Let $\Lambda_\theta(\alpha)$ denote the log moment generating function of $Y^{(t)}(\theta^*, \theta)$, which for every $\alpha \in \mathbb{R}$ is given by

$$\Lambda_\theta(\alpha) = \log \mathbb{E}[e^{\alpha Y(\theta^*, \theta)}] = \sum_{j=1}^{n} \log \mathbb{E}\left[\left\{\frac{f_j\left(X_j;\theta\right)}{f_j\left(X_j;\theta^*\right)}\right\}^{\alpha v_j}\right]. \tag{6}$$

Let I_θ denote the Fenchel-Legendre transform of Λ_θ and for all $x \in \mathbb{R}$, it is given by

$$I_\theta(x) = \sup_{\alpha \in \mathbb{R}} \left(\alpha x - \Lambda_\theta(\alpha)\right). \tag{7}$$

Remark 1. Fact 1 together with Assumption 1 guarantees that $K(\theta^*, \theta)$ is strictly positive. Also, Fact 1 and Assumption 2 imply that $\left\{\frac{f_j(X_j;\theta)}{f_j(X_j;\theta^*)}\right\}^{v_j}$ has finite moments of all orders. Hence, we have $\Lambda_\theta(\alpha)$ is finite for all $\alpha \in \mathbb{R}$.

Theorem 1 (Rate of rejecting $\theta \neq \theta^*$)**.** *Under Assumptions 1–4, for all $\theta \neq \theta^*$, for $0 < \eta < \overline{\eta}$, we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*, \theta) + \eta\right)$$
$$= -I_\theta(-K(\theta^*, \theta) + \eta), \tag{8}$$

where $\overline{\eta}$ satisfies $I_\theta(-K(\theta^*, \theta) + \overline{\eta}) \leq \min_{\theta \neq \theta^*} I_\theta(0)$. For $0 < \eta < \underline{\eta}$, we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*, \theta) - \eta\right)$$
$$= -I_\theta(-K(\theta^*, \theta) - \eta), \tag{9}$$

where $\underline{\eta}$ satisfies $I_\theta(-K(\theta^*, \theta) - \underline{\eta}) \leq \min_{\theta \neq \theta^*} I_\theta(0)$.

Here we provide a sketch of the proof; the detailed proof has been provided in the appendix.

First, we obtain a Large Deviation Principle on $\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)}$ as follows

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \geq -K(\theta^*, \theta) + \eta\right)$$
$$= -I_\theta(-K(\theta^*, \theta) + \eta), \tag{10}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log \frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \leq -K(\theta^*, \theta) - \eta\right)$$
$$= -I_\theta(-K(\theta^*, \theta) - \eta). \tag{11}$$

We obtain the above LDP as follows. Using the learning rule and Fact 1 we have

$$\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)}$$
$$= \frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{t}\left(W^\tau(i,j) - v_j\right)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta\right)}{f_j\left(X_j^{(t-\tau+1)};\theta^*\right)}$$
$$+ \frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{t} v_j \log\frac{f_j\left(X_j^{(t-\tau+1)};\theta\right)}{f_j\left(X_j^{(t-\tau+1)};\theta^*\right)}$$
$$+ \frac{1}{t}\sum_{j=1}^{n}\frac{q_j^{(0)}(\theta)}{q_j^{(0)}(\theta^*)}. \tag{12}$$

We obtain LDP for the three terms on the right hand side of the equation (12). We show that the first term satisfies an LDP with rate faster than any exponential. Then, we apply Cramer's theorem on the second term to show that it satisfies an LDP with rate function I_θ. The last term vanishes uniformly on all sample paths. Now, we obtain that the right hand side of equation (12) satisfies an LDP with rate function equal to the rate of the slowest decaying term, i.e., it satisfies LDP with rate function I_θ. Thus, we have equations (10) and (11).

Now, from equations (10) and (11) we obtain the following two inequalities

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*, \theta) + \eta\right)$$
$$\leq -I_\theta(-K(\theta^*, \theta) + \eta), \tag{13}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*, \theta) - \eta\right)$$
$$\geq -I_\theta(-K(\theta^*, \theta) - \eta). \tag{14}$$

All that remains to be shown is the following half LDPs

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*, \theta) + \eta\right)$$
$$\geq -I_\theta(-K(\theta^*, \theta) + \eta), \tag{15}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*, \theta) - \eta\right)$$
$$\leq -I_\theta(-K(\theta^*, \theta) - \eta). \tag{16}$$

It is straightforward to see that, for $\alpha > 0$ events of the following form

$$\left\{\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \geq -K(\theta^*,\theta)+\eta+\alpha\right\}$$
$$\cap\left\{\frac{1}{t}\log q_i^{(t)}(\theta^*) \geq -\alpha\right\}$$
$$\subset\left\{\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right\}.$$

Then, we show that for $0 < \eta < \overline{\eta}$, we have

$$\lim_{t\to\infty}\lim_{\alpha\to 0}\frac{1}{t}\log \mathsf{P}\left(\left\{\frac{1}{t}\log q_i^{(t)}(\theta^*) \geq -\alpha\right\}\right.$$
$$\left.\cap\left\{\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)}) \geq -K(\theta^*,\theta)+\eta+\alpha\right\}\right)$$
$$\geq -I_\theta(-K(\theta^*,\theta)+\eta). \tag{17}$$

Hence, for $0 < \eta < \overline{\eta}$, we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)$$
$$\geq -I_\theta(-K(\theta^*,\theta)+\eta). \tag{18}$$

Combining equations (13) and (18), for $0 < \eta < \overline{\eta}$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)$$
$$= -I_\theta(-K(\theta^*,\theta)+\eta). \tag{19}$$

Now consider deviations below the mean. It is straightforward to see that for any $\alpha > 0$ we have

$$\left\{\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right\}$$
$$\subset\left\{\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} > -K(\theta^*,\theta)-\eta+\alpha\right\}^C$$
$$\cup\left\{\frac{1}{t}\log q_i^{(t)}(\theta^*) > -\alpha\right\}^C. \tag{20}$$

From this, for $0 < \eta < \underline{\eta}$, we obtain the following

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)$$
$$\leq -I_\theta(-K(\theta^*,\theta)-\eta). \tag{21}$$

Combining equations (14) and (21), $0 < \eta < \underline{\eta}$ we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)$$
$$= -I_\theta(-K(\theta^*,\theta)-\eta). \tag{22}$$

Now recall Assumption 4 from [1].

Assumption 2′. There exists a finite $L \in \mathbb{R}$ such that for every $k \in [n]$ we have

$$-L \leq \max_{i\neq j}\sup_{X\in\mathcal{X}_k}\log\left(\frac{f_k(X;\theta_i)}{f_k(X;\theta_j)}\right) \leq L. \tag{23}$$

Remark 2. We obtained concentration of rate of rejecting $\theta \neq \theta^*$ for small deviations in Theorem 1, under Assumption 2 instead of Assumption 2′. The advantage of inclusion of Assumption 2 is that a larger class of distributions satisfy this. For instance, it allows the conditional distributions of observations, $f_k(\cdot;\theta)$, for every node k and $\theta \in \Theta$ to be continuous distributions such as the exponential distribution or the Gaussian distribution provided they satisfy Assumption 2. Also Assumption 2′ is special case of Assumption 2. This is because under Assumption 2′ we have $|\Lambda_\theta(\alpha)| \leq \alpha L$,which implies that $\Lambda_\theta(\alpha)$ is finite in the neighborhood of origin.

Corollary 1. *Suppose Assumption 2′ is satisfied for some finite $L \in \mathbb{R}$. Then for small η as specified in Theorem 1, we have*

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq K(\theta^*,\theta)+\eta\right) \leq -\frac{\eta^2}{2L^2}, \tag{24}$$

and

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq K(\theta^*,\theta)-\eta\right) \leq -\frac{\eta^2}{2L^2}. \tag{25}$$

Proof: Using equation (23) we get $-L \leq Y(\theta^*,\theta) \leq L$. Using Hoeffding's Lemma (Fact 5 in the Appendix) we have

$$\Lambda_\theta(\alpha) \leq \left(\frac{1}{2}\alpha^2 L^2 + \alpha K(\theta^*,\theta)\right). \tag{26}$$

Using the above equation and considering convex conjugate of Λ_θ at $K(\theta^*,\theta)+\eta$ for some $\eta > 0$, we have

$$\begin{aligned} I_\theta(K(\theta^*,\theta)+\eta) &= \sup_{\alpha\geq 0}\left(\alpha(K(\theta^*,\theta)+\eta) - \Lambda_\theta(\alpha)\right) \\ &\geq \sup_{\alpha\geq 0}\left(\alpha\eta - \frac{1}{2}\alpha^2 L^2\right) \\ &= \sup_{\alpha\geq 0}\left(\frac{\eta^2}{2L^2} - \left(\frac{\alpha L}{\sqrt{2}} - \frac{\eta}{\sqrt{2}L}\right)^2\right) \\ &= \frac{\eta^2}{2L^2}. \end{aligned} \tag{27}$$

Similarly, we have

$$I_\theta(K(\theta^*,\theta)-\eta) \geq \frac{\eta^2}{2L^2}. \tag{28}$$

Using equations (27) and (28) with Theorem 1 we have the assertion of the lemma. ■

Remark 3. Corollary 1 shows that under the boundedness assumption on ratio of log-likelihoods, $\log\left(\frac{f_k(X;\theta_i)}{f_k(X;\theta_j)}\right)$ for every $k \in [n]$ and $i \neq j, i, j \in [M]$, the concentration result obtain from Theorem 1 gives the same exponent obtained using Hoeffding's inequality (Theorem 2 in [2]).

The following lemma provides a sufficient condition on the conditional distributions of observations, $f_k(\cdot;\theta)$, for every node k and $\theta \in \Theta$, which ensures that Assumption 2 will hold.

Lemma 1. *The log moment generating function of $\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}$, for $k \in [n]$ and $i, j \in [M]$ such that $i \neq j$, is finite in an interval containing origin if and only if it has exponentially decaying tails, i.e., there exist some positive constants C and β such that*

$$\mathsf{P}\left(\frac{f_k\left(X_k;\theta_i\right)}{f_k\left(X_k;\theta_j\right)} \geq x\right) \leq Ce^{-\beta x}. \tag{29}$$

Proof: Suppose

$$\mathbb{E}\left[e^{\beta\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}}\right] < \infty, \tag{30}$$

for some β in an interval containing 0. Then, by Markov inequality we have

$$\begin{aligned}\mathsf{P}\left(\frac{f_k\left(X_k;\theta_i\right)}{f_k\left(X_k;\theta_j\right)} > x\right) &= \mathsf{P}\left(e^{\beta\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}} > e^{\beta x}\right) \\ &\leq e^{-\beta x}\mathbb{E}\left[e^{\beta\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}}\right]. \end{aligned}\tag{31}$$

Take $C = \mathbb{E}\left[e^{\beta\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}}\right]$, hence we have equation (29). For the other direction consider $\alpha > 0$, then we have the following

$$\begin{aligned}\mathbb{E}\left[e^{\alpha\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}}\right] &\leq \int_0^\infty \mathsf{P}\left(e^{\alpha\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}} > x\right)dx \\ &\leq 1 + \int_1^\infty Cx^{-\frac{\beta}{\alpha}}dx. \end{aligned}\tag{32}$$

This implies for $0 < \alpha < \beta$, $\mathbb{E}\left[e^{\alpha\frac{f_k(X_k;\theta_i)}{f_k(X_k;\theta_j)}}\right] < \infty$, i.e., the log moment generating function is finite in the interval $[0, \beta)$. Now, it is straightforward to see that log moment generating function is finite in the interval $(-\beta, 0]$ if we have

$$\mathsf{P}\left(\frac{f_k\left(X_k;\theta_i\right)}{f_k\left(X_k;\theta_j\right)} \leq \frac{1}{x}\right) \leq Ce^{-\beta x}. \tag{33}$$

This is true since equation (29) holds for $\frac{f_k(X_k;\theta_j)}{f_k(X_k;\theta_i)}$. ■

Note that the sufficient and necessary condition provided by above lemma is a weaker condition than Assumption 2′ which can be satisfied by a larger class of continuous distributions such as Gaussian and exponential distributions.

IV. Conclusion

We consider the learning rule proposed in [2]. We relax the assumption on log-likelihood ratios to be bounded with an assumption that is be satisfied by a larger class of continuous distributions such as Gaussian and exponential distributions. Under the weaker assumption, we show concentration of rate of rejection untrue hypothesis for any aperiodic network under mild assumptions. For appropriately small deviations in the learning rate around the network divergence, we obtain a large deviation principle, where the exact exponent of decay has been characterized.

Appendix

Fact 2 (Facts on convergence in [15]). *At any time t, an irreducible and aperiodic stochastic matrix W satisfies*

$$\left|W^t(i,j) - v_j\right| \leq n\lambda_{\max}(W)^t, \tag{34}$$

for any $j \in [n]$, where v_j are components of v which is the left eigenvector of W associated with eigenvalue 1 and where $\lambda_{\max}(W) < 1$ is the second largest absolute value of eigenvalues of W.

Fact 3 (Cramer's Theorem in $\mathbb{R}$ (Theorem 2.2.3 of [16])). *Let $X \in \mathbb{R}$ be a random variable, let $\{X_i, i \in \mathbb{N}\}$, be i.i.d random variables distributed like X, and let $S_n = \sum_{i=1}^n X_i$. Then for any measurable set $B \in \mathbb{R}$, we have*

$$\liminf_{n\to\infty} \frac{1}{n}\log\mathsf{P}\left(\frac{S_n}{n} \in B\right) \geq -\inf_{x\in B^o} I(x), \tag{35}$$

$$\limsup_{n\to\infty} \frac{1}{n}\log\mathsf{P}\left(\frac{S_n}{n} \in B\right) \leq -\inf_{x\in \bar{B}} I(x), \tag{36}$$

where B^o denotes the interior of B and $\bar{B}$ denotes the closure of B.

Fact 4 (Lemma 2.2.5 and Lemma 2.2.20 of [16]). *Let $\epsilon > 0$ and let $\mu := \mathbb{E}[X]$. If $\Lambda(\alpha)$ is finite in an interval containing origin, then $I(x)$ is non-decreasing for all $x \in [\mu+\epsilon, \infty)$ and we have*

$$I(x) = \sup_{\alpha\geq 0}\left(\alpha x - \Lambda(\alpha)\right) > 0. \tag{37}$$

Similarly, $I(x)$ is non-increasing for $x \in (-\infty, \mu-\epsilon]$ and we have

$$I(x) = \sup_{\alpha\leq 0}\left(\alpha x - \Lambda(\alpha)\right) > 0. \tag{38}$$

Also $\Lambda(\alpha)$ is differentiable at 0 *and $\Lambda'(0) = \mu$. Moreover,* $\inf_{x\in\mathbb{R}} I(x) = I(\mu) = 0$.

Fact 5 (Hoeffding's Lemma). *Let X be any real valued random variable with expected value $\mathbb{E}[X] = \mu$ and such that $a \leq X \leq b$ almost surely. Then, for all $\alpha \in \mathbb{R}$,*

$$\mathbb{E}[e^{\alpha X}] \leq \exp\left(\frac{1}{8}\alpha^2(b-a)^2 + \alpha\mu\right). \tag{39}$$

A. Proof of Theorem 1

Using the learning rule and Fact 1 we have

$$\begin{aligned}&\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \\ &= \frac{1}{t}\sum_{j=1}^n\sum_{\tau=1}^t \left(W^\tau(i,j) - v_j\right)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta\right)}{f_j\left(X_j^{(t-\tau+1)};\theta^*\right)} \\ &\quad + \frac{1}{t}\sum_{j=1}^n\sum_{\tau=1}^t v_j\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta\right)}{f_j\left(X_j^{(t-\tau+1)};\theta^*\right)} \\ &\quad + \frac{1}{t}\sum_{j=1}^n \frac{q_j^{(0)}(\theta)}{q_j^{(0)}(\theta^*)}. \end{aligned}\tag{40}$$

Details of obtaining the above equation are provided in [2]. The second term in equation (40), denoted by S_t, is the following

$$S_t = \frac{1}{t}\sum_{\tau=T}^{t} Y^{(t-\tau+1)}(\theta^*,\theta). \tag{41}$$

Now, applying Cramer's Theorem in $\mathbb{R}$ (Fact 3 in the Appendix) on the set $B_1 = [-K(\theta^*,\theta)+\eta,\infty)$ and we get

$$\begin{aligned}\limsup_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(S_t \geq -K(\theta^*,\theta)+\eta\right)&\\ \leq -\inf_{x\geq -K(\theta^*,\theta)+\eta} I_\theta\left(x\right),&\end{aligned} \tag{42}$$

and

$$\begin{aligned}\liminf_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(S_t > -K(\theta^*,\theta)+\eta\right)&\\ \geq -\inf_{x> -K(\theta^*,\theta)+\eta} I_\theta\left(x\right).&\end{aligned} \tag{43}$$

Now using Remark 2 and Fact 4 in Appendix, we have

$$\inf_{x\geq -K(\theta^*,\theta)+\epsilon} I_\theta\left(x\right) = I_\theta\left(-K(\theta^*,\theta)+\eta\right) \tag{44}$$

and

$$\begin{aligned}\inf_{x> -K(\theta^*,\theta)+\epsilon} I_\theta\left(x\right) &= \lim_{x\downarrow -K(\theta^*,\theta)+\epsilon} I_\theta\left(-K(\theta^*,\theta)+\eta\right)\\ &= I_\theta\left(-K(\theta^*,\theta)+\eta\right).\end{aligned} \tag{45}$$

Hence, we get

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(S_t \geq -K(\theta^*,\theta)+\eta\right)\\ &= -I_\theta\left(-K(\theta^*,\theta)+\eta\right).\end{aligned} \tag{46}$$

Similarly applying Cramer's Theorem on the set $B_2 = (-\infty, -K(\theta^*,\theta)-\eta]$, we get

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(S_t \leq -K(\theta^*,\theta)-\eta\right)\\ &= -I_\theta\left(-K(\theta^*,\theta)-\eta\right).\end{aligned} \tag{47}$$

Now, consider the first term in equation (40), denoted by E_t, we have

$$\begin{aligned}|E_t| &= \left|\frac{1}{t}\sum_{j=1}^{n}\sum_{\tau=1}^{t}\left(W^\tau(i,j)-v_j\right)\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta^*\right)}{f_j\left(X_j^{(t-\tau+1)};\theta\right)}\right|\\ &\leq \frac{n}{t}\sum_{\tau=1}^{t}\lambda_{\max}(W)^\tau\left(\sum_{j=1}^{n}\left|\log\frac{f_j\left(X_j^{(t-\tau+1)};\theta^*\right)}{f_j\left(X_j^{(t-\tau+1)};\theta\right)}\right|\right).\end{aligned} \tag{48}$$

From Lemma 2, we have the right hand side of above inequality vanishes to zero in probability at a rate faster than any exponential. Hence, we have

$$\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(|E_t|\geq\delta\right) = -\infty \tag{49}$$

Now, applying Lemma 3, for all $\eta > 0$ we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \geq -K(\theta^*,\theta)+\eta\right)\\ &= -I_\theta(-K(\theta^*,\theta)+\eta),\end{aligned} \tag{50}$$

and

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \leq -K(\theta^*,\theta)-\eta\right)\\ &= -I_\theta(-K(\theta^*,\theta)-\eta).\end{aligned} \tag{51}$$

The above LDP implies the following half LDPs for $\frac{1}{t}\log q_i^{(t)}(\theta)$. For $0<\eta<K(\theta^*,\theta)$, we get

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)\\ &\leq -I_\theta(-K(\theta^*,\theta)+\eta),\end{aligned} \tag{52}$$

and for $\eta > 0$ we get

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)\\ &\geq -I_\theta(-K(\theta^*,\theta)-\eta).\end{aligned} \tag{53}$$

Consider $\theta\neq\theta^*$ and $0<\eta<\overline{\eta}$. For any $\alpha>0$, we have

$$\begin{aligned}&\left\{\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \geq -K(\theta^*,\theta)+\eta+\alpha\right\}\\ &\cap\left\{\frac{1}{t}\log q_i^{(t)}(\theta^*) \geq -\alpha\right\}\\ &\subset\left\{\frac{1}{t}\log q_i^{(t)}(\theta^*) \geq -K(\theta^*,\theta)+\eta\right\}.\end{aligned} \tag{54}$$

Then, we get

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta^*) \geq -K(\theta^*,\theta)+\eta\right)\\ &\geq \mathsf{P}\left(\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \geq -K(\theta^*,\theta)+\eta+\alpha\right)\\ &\quad - \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta^*) < -\alpha\right),\end{aligned} \tag{55}$$

where the last inequality follows from the inequality $\mathsf{P}(A) = \mathsf{P}(A\cap B)+\mathsf{P}(A\cap B^C) \leq \mathsf{P}(A\cap B)+\mathsf{P}(B^C)$, i.e. $\mathsf{P}(A)-\mathsf{P}(B^C)\leq\mathsf{P}(A\cap B)$. We know that for every $\epsilon>0$ exists a T such for all $t\geq T$

$$\frac{1}{t}\log \mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta^*) < -\alpha\right) \leq -\min_{\theta\neq\theta^*} I_\theta(0)+\epsilon. \tag{56}$$

Similarly, for every $\epsilon>0$, there exists a T_1 such that for all $t\geq T_1$ we have

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \geq -K(\theta^*,\theta)+\eta+\alpha\right)\\ &\geq e^{-I_\theta(-K(\theta^*,\theta)+\eta+\alpha)t-\epsilon t}.\end{aligned} \tag{57}$$

Therefore, for every $\epsilon > 0$ there exists T such that for all $t \geq T$ we have

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)\\ &\geq e^{-I_\theta(-K(\theta^*,\theta)+\eta+\alpha)t-\epsilon t} - e^{-\min_{\theta\neq\theta^*} I_\theta(0)t+\epsilon t}.\end{aligned} \tag{58}$$

From right continuity of I_θ we have

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)\\ &\geq \lim_{\alpha\to 0^+} e^{-I_\theta(-K(\theta^*,\theta)+\eta+\alpha)t-\epsilon t} - e^{-I_{\hat{\theta}}(0)t+\epsilon t}\\ &= e^{-I_\theta(-K(\theta^*,\theta)+\eta)t-\epsilon t} - e^{-\min_{\theta\neq\theta^*} I_\theta(0)t+\epsilon t}.\end{aligned} \tag{59}$$

In other words, for every $\epsilon > 0$, there exists T such that for all $t \geq T$ we have

$$\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right) \tag{60}$$

$$\begin{aligned}&\geq -I_\theta(-K(\theta^*,\theta)+\eta)-\epsilon\\ &+\frac{1}{t}\log\left(1-e^{I_\theta(-K(\theta^*,\theta)+\eta)t-\min_{\theta\neq\theta^*} I_\theta(0)t+2\epsilon t}\right).\end{aligned} \tag{61}$$

For all $0 < \eta < \overline{\eta}$ we have $I_\theta(-K(\theta^*,\theta)+\eta) \leq \min_{\theta\neq\theta^*} I_\theta(0)$, hence there exists a T such that for all $t \geq T$ we have

$$\begin{aligned}&\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)\\ &\geq -I_\theta(-K(\theta^*,\theta)+\eta)-2\epsilon.\end{aligned} \tag{62}$$

Therefore, for $0 < \eta < \overline{\eta}$, we have the other half of the LDP for deviations above the mean

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)\\ &\geq -I_\theta(-K(\theta^*,\theta)+\eta).\end{aligned} \tag{63}$$

Combining above equation with equation (52), for $0 < \eta < \overline{\eta}$ we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \geq -K(\theta^*,\theta)+\eta\right)\\ &= -I_\theta(-K(\theta^*,\theta)+\eta).\end{aligned} \tag{64}$$

Consider the following for any $\alpha > 0$

$$\begin{aligned}&\left\{\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right\}\\ &\subset \left\{\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} > -K(\theta^*,\theta)-\eta+\alpha\right\}^C\\ &\cup \left\{\frac{1}{t}\log q_i^{(t)}(\theta^*) > -\alpha\right\}^C,\end{aligned} \tag{65}$$

which implies for all

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)\\ &\leq \mathsf{P}\left(\frac{1}{t}\log\frac{q_i^{(t)}(\theta)}{q_i^{(t)}(\theta^*)} \leq -K(\theta^*,\theta)-\eta+\alpha\right)\\ &\quad+\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta^*) \leq -\alpha\right)\\ &\leq e^{-I_\theta(-K(\theta^*,\theta)-\eta+\alpha)t} + e^{-\min_{\theta\neq\theta^*} I_\theta(0)t}.\end{aligned} \tag{66}$$

From right continuity of I_θ we have

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)\\ &\leq \lim_{\alpha\to 0^+} e^{-I_\theta(-K(\theta^*,\theta)-\eta+\alpha)t} + e^{-\min_{\theta\neq\theta^*} I_\theta(0)t}\\ &\leq e^{-I_\theta(-K(\theta^*,\theta)-\eta)t} + e^{-\min_{\theta\neq\theta^*} I_\theta(0)t}.\end{aligned} \tag{67}$$

For $0 < \eta < \underline{\eta}$, we have

$$\begin{aligned}&\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)\\ &\leq 2e^{-I_\theta(-K(\theta^*,\theta)-\eta)t}.\end{aligned} \tag{68}$$

Therefore, for $0 < \eta < \underline{\eta}$, we have the other half LDP for deviations below mean

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)\\ &\leq -I_\theta(-K(\theta^*,\theta)-\eta).\end{aligned} \tag{69}$$

Combining above equation with equation (53), for $0 < \eta < \underline{\eta}$, we have

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{t}\log\mathsf{P}\left(\frac{1}{t}\log q_i^{(t)}(\theta) \leq -K(\theta^*,\theta)-\eta\right)\\ &= -I_\theta(-K(\theta^*,\theta)-\eta).\end{aligned} \tag{70}$$

Lemma 2. *Let $0 < q < 1$. Let X_i be a sequence of non-negative i.i.d random variables distributed as X and $\Lambda(\alpha)$ is its log-moment generating function which is finite for all $\alpha \in \mathbb{R}$, then for every $\delta > 0$ we have*

$$\lim_{n\to\infty}\frac{1}{n}\log\mathsf{P}\left(\frac{\sum_{i=1}^n q^i X_i}{n} \geq \delta\right) = -\infty. \tag{71}$$

Proof: Applying Chebycheff's inequality and using the definition of log moment generating function, for $\alpha \in \mathbb{R}$, we have

$$\begin{aligned}\mathsf{P}\left(\frac{1}{t}\sum_{i=1}^t q^i X_i \geq \delta\right) &\leq \mathbb{E}\left[e^{\alpha t\left(\frac{1}{t}\sum_{i=1}^t q^i X_i-\delta\right)}\right]\\ &= e^{-t\left(\alpha\delta-\frac{1}{t}\sum_{i=1}^t \Lambda(q^i\alpha)\right)}.\end{aligned} \tag{72}$$

From convexity of Λ, we have

$$\sum_{i=1}^t \Lambda(q^i\alpha) \leq \Lambda(\alpha)\sum_{i=1}^t q^i. \tag{73}$$

Since $\Lambda(\alpha)$ is finite and $\sum_{i=1}^{\infty} q^i < \infty$, for every $\delta > 0$ we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t} \sum_{i=1}^{t} q^i X_i \geq \delta \right) \leq -\alpha\delta. \tag{74}$$

Since, the above equation is true for all $\alpha \in \mathbb{R}$, we have the assertion of the lemma. ∎

We provide the next lemma without proof for the sake of brevity.

Lemma 3. *Consider sequences $\{E_t\}_{t=0}^{\infty}$ and $\{S_t\}_{t=0}^{\infty}$ such that S_t has mean K and satisfies an LDP with rate function $I(\cdot)$. For all $\eta > 0$ we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(S_t \geq K + \eta\right) = -I(K+\eta), \tag{75}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(S_t \leq K - \eta\right) = -I(K-\eta). \tag{76}$$

The sequence E_t has the following property for every $\delta > 0$

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(|E_t| \geq \delta\right) = -\infty, \tag{77}$$

then the sequence $S_t + E_t$ satisfies LDP with rate function $I(\cdot)$. In other, words we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(S_t + E_t \geq K + \eta\right) = -I(K+\eta), \tag{78}$$

and

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(S_t + E_t \leq K - \eta\right) = -I(K-\eta). \tag{79}$$

Lemma 4. *For all $\alpha > 0$, we have*

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t} \log q_i^{(t)}(\theta^*) \leq -\alpha \right) \leq -\min_{\theta\neq\theta^*} I_\theta(0). \tag{80}$$

Proof:

For any $\alpha > 0$, consider

$$\begin{aligned}
&\mathsf{P}\left(\frac{1}{t} \log q_i^{(t)}(\theta^*) \leq -\alpha \right) \\
&= \mathsf{P}\left(q_i^{(t)}(\theta^*) \leq e^{-\alpha t} \right) \\
&\leq \mathsf{P}\left(1 - e^{-\alpha t} \leq \sum_{\theta\neq\theta^*} q_i^{(t)}(\theta) \right) \\
&\leq \sum_{\theta\neq\theta^*} \mathsf{P}\left(\frac{1}{M}\left(1 - e^{-\alpha t}\right) \leq q_i^{(t)}(\theta) \right) \\
&\leq \sum_{\theta\neq\theta^*} e^{-I_\theta\left(-K(\theta^*,\theta)+K(\theta^*,\theta)-\frac{1}{t}\log M+\frac{1}{t}\log\left(1-e^{-\alpha t}\right)\right)t}.
\end{aligned} \tag{81}$$

By taking limit we have

$$\lim_{t\to\infty} \frac{1}{t} \log \mathsf{P}\left(\frac{1}{t} \log q_i^{(t)}(\theta^*) \leq -\alpha \right) \leq -\min_{\theta\neq\theta^*} I_\theta(0). \tag{82}$$

∎

References

[1] A. Lalitha, A. Sarwate, and T. Javidi, "Social learning and distributed hypothesis testing," in *Information Theory (ISIT), 2014 IEEE International Symposium on*, June 2014, pp. 551–555.

[2] A. Lalitha, T. Javidi, and A. Sarwate, "Social Learning and Distributed Hypothesis Testing," *ArXiv e-prints*, Oct. 2014.

[3] 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.

[4] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, Dec 2013, pp. 6196–6201.

[5] A. Jadbabaie, P. Molavi, and A. Tahbaz-salehi, "Information heterogeneity and the speed of learning in social networks," Working Paper 2013.

[6] K. Rahnama Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *Decision and Control (CDC), 2010 49th IEEE Conference on*, Dec 2010, pp. 5050–5055.

[7] 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.

[8] V. Saligrama, M. Alanyali, and O. Savas, "Distributed detection in sensor networks with packet losses and finite capacity links," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4118–4132, Nov 2006.

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

[10] R. Rahman, M. Alanyali, and V. Saligrama, "Distributed tracking in multihop sensor networks with communication delays," *Signal Processing, IEEE Transactions on*, vol. 55, no. 9, pp. 4656–4668, Sept 2007.

[11] A. Nedić, A. Olshevsky, and C. A. Uribe, "Nonasymptotic Convergence Rates for Cooperative Learning Over Time-Varying Directed Graphs," *ArXiv e-prints*, Oct. 2014.

[12] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed Detection : Finite-time Analysis and Impact of Network Topology," *ArXiv e-prints*, Sep. 2014.

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

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

[15] J. S. Rosenthal, "Convergence rates for markov chains," *Siam Review*, vol. 37, no. 3, pp. 387–405, 1995.

[16] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, ser. Applications of mathematics. New York, Berlin, Heidelberg: Springer, 1998. [Online]. Available: http://opac.inria.fr/record=b1093895