Inaccuracy rates for distributed inference over random networks with applications to social learning

Dragana Bajović, *Member, IEEE*

Abstract

This paper studies probabilistic rates of convergence for consensus+innovations type of algorithms in random, generic networks. For each node, we find a lower and also a family of upper bounds on the large deviations rate function, thus enabling the computation of the exponential convergence rates for the events of interest on the iterates. Relevant applications include error exponents in distributed hypothesis testing, rates of convergence of beliefs in social learning, and inaccuracy rates in distributed estimation. The bounds on the rate function have a very particular form at each node: they are constructed as the convex envelope between the rate function of the hypothetical fusion center and the rate function corresponding to a certain topological mode of the node's presence. We further show tightness of the discovered bounds for several cases, such as pendant nodes and regular networks, thus establishing the first proof of the large deviations principle for consensus+innovations and social learning in random networks.

Index Terms

Large deviations, distributed inference, social learning, convex analysis, inaccuracy rates.

D. Bajović is with the Department of Power, Electronics and Communications Engineering, Faculty of Technical Sciences, University of Novi Sad. Email: dbajovic@uns.ac.rs.

Part of this work was done while the author was with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh 15213, PA, USA [1].

This work is partially supported by the European Union's Horizon 2020 Research and Innovation program under grant agreement No 957337. The paper reflects only the view of the author and the Commission is not responsible for any use that may be made of the information it contains.

I. INTRODUCTION

The theory of large deviations is the most prominent tool for studying *rare events* that occur with stochastic processes, offering a principled approach for estimating probabilities of such events. A typical setup concerns a sequence of probability measures induced by the studied process and parameterized by one of the process parameters (e.g., time, population size, learning rate etc.), with the goal of computing, or characterizing, the respective decay rate, for any given event (region) of interest. The practical value of such rates is in estimating the probability of a rare event of interest as an exponentially decaying function of the concerned process parameter, while neglecting the terms with slower than exponential dependence. The rates of rare events can additionally provide a ground for comparison of two statistical procedures, as originally proposed in the seminal work by Chernoff [2], and can therefore serve as a useful design criterion [3], [4], [5], [6]. This is of special interest in the cases when other performance metrics are intractable for optimization, such as probabilities of error with hypothesis testing.

In addition to the rate computation, large deviations analysis often reveals the most likely way through which the event of interest takes place, providing additional important insights that can guide system design. Most notable applications of large deviations theory are in statistics [7], communications and queuing theory [8], statistical mechanics [9], and information theory [10].

For example, in statistical estimation, an event of interest is the event that the estimator does not belong to a predefined close neighborhood of the parameter being estimated [11]. The decay rates of probabilities of such events are known in the estimation theory as *inaccuracy rates* and can, e.g., guide the decision on how many samples are needed for the estimator to reach the desired accuracy, with high probability [12]. To make the exposition concrete, let $X_t \in \mathbb{R}^d$, $t = 1, 2, ...$, be a sequence of estimators of a parameter $\theta \in \mathbb{R}^d$. Assuming that X_t converges to θ, an event of interest has the form $\{\|X_t - \theta\| \geq \epsilon\}$, where $\|\cdot\|$ denotes the l_2 norm (other vector norms can also be used). An equivalent way to represent this event is $\{X_t \in C_\epsilon\}$, where C_ϵ is the complement of the l_2 ball of diameter ϵ centered at θ, $C_\epsilon = B^c_\theta(\epsilon)$. Provided that X_t converges to θ, the probabilities of these events typically vanish exponentially fast with t. Large deviations analysis then aims at discovering the corresponding rate of decay, i.e., the inaccuracy rate $\mathbf{I}(C_\epsilon)$:

$$\mathbb{P}\left(X_t \in C_\epsilon\right) = e^{-t\mathbf{I}(C_\epsilon)+o(t)}, \tag{1}$$

where $o(t)$ denotes a function growing slower than linear with t. The inaccuracy rate $\mathbf{I}(C_\epsilon)$ has a very particular structure: it is given through the so called *rate function* $I : \mathbb{R}^d \mapsto \mathbb{R}$ by

$$\mathbf{I}(C_\epsilon) = \inf_{x \in C_\epsilon} I(x). \tag{2}$$

The rate function I is itself defined through the statistics of the inference sequence X_t. It should be noted that, in contrast with the set function $\mathbf{I}$, the rate function I does not depend on the inaccuracy region, i.e., when C_ϵ varies, only the domain of minimization on the right-hand side of (2) varies, while the rate function remains fixed. Also, this relation holds for an arbitrary set C_ϵ (e.g., not necessarily a ball complement). Hence, once the rate function is identified, the associated inaccuracy rate is readily computable through (2) for a new given region of interest, without the need to redo the large deviations analysis each time, i.e., for each new region. Large deviations rate for estimation were first studied by Bahadur in [12].

Another well-known application of large deviations analysis is hypothesis testing [2], where the sequence X_t is typically a decision statistics, e.g., obtained by summing up the log-likelihoods of the collected measurements up to the current time t, $X_t = 1/t \sum_{s=1}^{t} \log \frac{f_1(Y_s)}{f_0(Y_s)}$; f_0 and f_1 here are the marginal distributions of the measurements Y_s under the two hypotheses H_0 and H_1, respectively. If the acceptance threshold for H_1 at time t is γ_t, then rare events of interest are $\{X_t < \gamma_t\}$, when H_1 is true (i.e., when Y_s follow the distribution f_1) – resulting in missed detection, and $\{X_t \geq \gamma_t\}$, when H_0 is true (when Y_s follow the distribution f_0) – causing a false alarm. When C_ϵ in (1) is replaced by the preceding two events, the resulting large deviations rates $\mathbf{I}(C_\epsilon)$ are then the well-known *error exponents* that provide decay rates of the corresponding error probabilities.

In this paper we are concerned with large deviations rates of *distributed* statistical inference, where observations originate at different locations or different entities. Relevant works include algorithms such as consensus+innovations [13], [14], [15], [16], diffusion [17], [18], [19], and non-Bayesian or social learning [20], [21], [22], [23]. The common setup of the above works consists of networked nodes, each holding a local inference vector (parameter estimates, decision variables, beliefs) that is being updated over time. The updates are based on incorporating local, private signals that each agent observes over time, and then exchanging with immediate neighbors and averaging the received information through the well-known DeGroot averaging [24] (also known as consensus).

Asymptotic performance of distributed detection was studied in [13], for Gaussian observations, [14], for generic observations, and in [15], for networks with noisy communication links. In each of the named works, a randomly switching network topology is assumed and conditions for asymptotic equivalence of an arbitrary network node and a fusion center (with access to all observations) are studied. Reference [16] considers directed networks, both static and randomly varying, and studies the rate function for the vector of states, deriving the exact rate function for the case of static networks, and providing bounds on the exponential rates for randomly switching networks. The rate function for static networks is given as the weighted combination of the local rate functions, with weights being equal to the eigenvector centralities

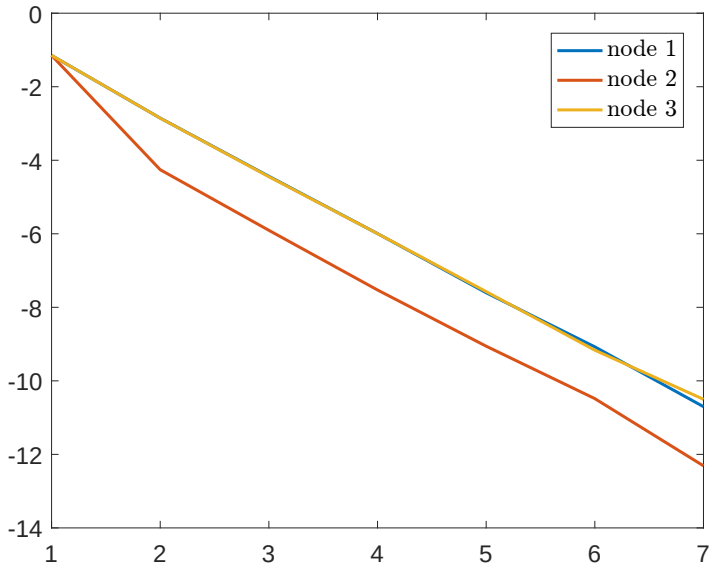
(i.e., the left Perron vector of the consensus matrix). Reference [17] studies distributed detection for static and symmetric networks and constant step size. For the limiting distribution of the local states, it proves the large deviations principle when the step size parameter decreases and shows that the rate function is equivalent to the centralized detector. These results are refined and extended in [18] by studying non-exponential terms and directed (static) networks. Reference [19] further considers distributed detection with 1-bit messages, while recent reference [6] addresses optimal aggregation strategies for social learning.

References [20], [21], [22], [23] study distributed M-ary hypothesis testing, where local updates are formed by applying Bayesian update on the vector of prior beliefs, based on the newly acquired local measurements. Assuming static, directed network, in [20] and [21], beliefs across immediate neighborhoods are merged through arithmetic average [20], while [22] adopts geometric average (or, equivalently, arithmetic average on the log-beliefs). A different merging rule is proposed and analyzed in [23], where instead of averaging, beliefs are updated by computing the minimum across the neighbors beliefs and the nodes' locally generated beliefs, showing improvement in the learning rate. Large deviations of the beliefs are addressed in [22], where it was proven that the log-ratios of beliefs with respect to the belief in the true distribution, satisfy the large deviations principle, with the rate function being equal to the eigenvector-centralities convex combination of the nodes' local rate functions, similarly as in [16] and [18]. Through the contraction principle, [22] also shows that the (log)-beliefs themselves satisfy the large deviations principle.

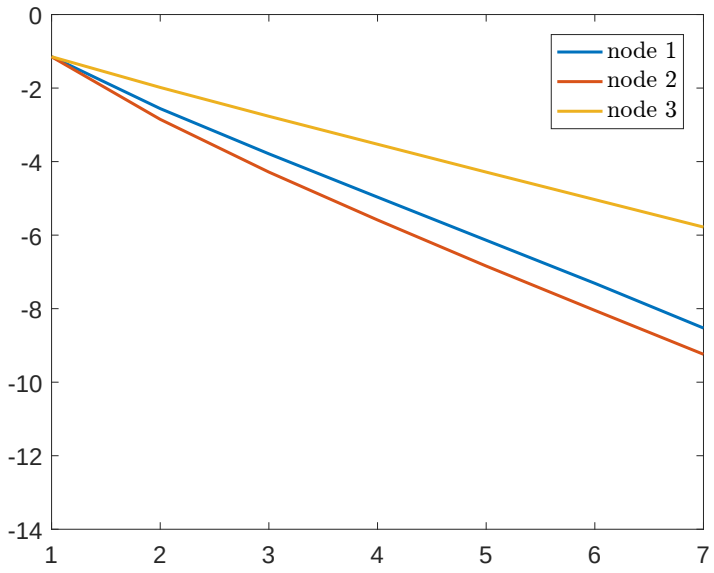
Contributions. In contrast with the works in [17]-[23], in this paper, we address computation of the rate function for distributed inference on *random networks*. This model shift from static to random networks has fundamental implications on the large deviations performance. To explain this at an intuitive level: when the underlying network is random, consensus mixing of local inference vectors might be disabled for an arbitrary long period of time due to the lack of communications. In general, the topology can then break down into several connected components of the original network[1]. When in this regime, neither of the nodes can "see" the observations beyond the connected component they belong to, and hence the resulting rate function will be strictly lower than that of the full network[2]. Figure 1 illustrates this effect with a toy example of a 3-node chain where each node produces scalar observations of standard Gaussian distribution.

[1]Note that this is very different from time-varying networks that are typically modelled by the assumption of the so called bounded intercommunication interval, which guarantees that the union graph formed of all communication links occurring in this interval is connected, after a strictly finite time, e.g., [23], [25]

[2]This is a consequence of the non-negativity of the rate function and the fact that it (roughly) scales linearly with the number of observation sources, as detailed in the paper.



(a) Static topology



(b) Randomly varying topology

Fig. 1: Decay of the log-probabilities in (1) for a fixed set C for static (top) and random (bottom) 3 node chain network.

In the top figure, we plot the logarithm of the probability in (1) for a ball complement inaccuracy set C, when the chain topology is fixed (static). In the bottom figure we plot the same probability, but when the two links of the chain graph alternate at random over time. We label the middle node as node 2, and we let the communication frequency between nodes 1 and 2 be higher (equal to 0.8) than the one between higher than the one between nodes 2 and 3 (equal to 0.2). It is clear from the figure that the static topology achieves much steeper decay, and, moreover, this decay is equal at each of the three nodes (and also equal to the decay of the hypothetical fusion center, cf. Section IV, as predicted by the theory). In contrast, in the random case, the difference between the nodes' decays is evident: node 2 achieves the steepest decay, followed by node 1, while node 3 has the worst performance.

In this work, we are interested in understanding the rate function of each node in the network and analytically expressing its dependence on the system parameters. For each node, we find a lower and a family of upper bounds on the rate function. This is achieved by carrying out node-specific large deviations analyses. We show that the two bounds match in several cases, such as for pendant nodes and also for nodes in a regular network. The family of upper bounds is indexed by different induced components of the given node, and each function in this family has the form of the convex envelope between the rate function of the full network and the rate function of the respective component, lifted up by the probability of the event that induces the component. The lower bound is given as the convex envelope between the rate function of the full network and the node's local rate function lifted up by the large deviations rate of consensus, whose existence was shown in [26]. With respect to references [13], [14], [15], there are several important novelties. First, we extend the results of [14] to the case of vector observations and vector inference state. Second, while [13]-[15] only provide a lower bound on the rate function, this work, as described in the above, finds also a family of upper bounds. This is achieved by carefully devising events that impact the rate function, and for which we develop novel large deviations techniques. The discovered upper bounds enable to establish, to the best of our knowledge, the first proof of the large deviations principle for nodes performing DeGroot-based distributed inference in randomly varying networks.

As an application of particular interest to this study, we consider social learning, specifically the form with the geometric average update [22]. We show that, with appropriate transformation of the belief iterates – namely, considering their log-ratios with respect to the belief in the true distribution, the algorithm studied in [22] exhibits full equivalence to the consensus+innovations algorithm that we analyze here. Building on this equivalence, we characterize the rate function of the beliefs in social learning and provide the first proof of the large deviations principle for social learning run over random networks.

A closely related work to ours is [27] that studies convergence properties of social learning over

random networks. This reference shows that, almost surely, each node is able to correctly identify the true hypothesis. We similarly focus on the case of random networks, but we are additionally concerned with characterizing the *rates* of probabilistic convergence of the iterates in the sense of large deviations. Finally, we show that almost sure convergence of the beliefs follows from the obtained large deviations rates.

From the technical perspective, this paper contributes with a novel set of techniques and approaches that could be of interest for further studies of social learning, and more generally, distributed inference in random networks.

Notation. For arbitrary $d \in \mathbb{N}$ we denote by 0_d the d-dimensional vector of all zeros; by 1_d the d-dimensional vector of all ones; by e_i the i-th canonical vector of $\mathbb{R}^d$ (that has value one on the i-th entry and the remaining entries are zero); by I_d the d-dimensional identity matrix; by J_d the $d \times d$ matrix whose all entries equal to $1/d$. For a matrix A, we let $[A]_{ij}$ and A_{ij} denote its i,j entry and for a vector $a \in \mathbb{R}^d$, we denote its i-th entry by a_i, $i,j = 1,...,d$. For the set of indices $C \subseteq \{1,2,...,N\}$, we let $[A]_C$ (or A_C) denote the submatrix of A that corresponds to indices in C. For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we denote its domain by $\mathcal{D}_f = \left\{x \in \mathbb{R}^d : -\infty < f(x) < +\infty\right\}$; for a set $D \subseteq \mathbb{R}$, $f^{-1}(D)$ is defined as $f^{-1}(D) = \{x \in \mathbb{R}^d : f(x) \in D\}$. $\log$ denotes the natural logarithm. For $N \in \mathbb{N}$, we denote by Δ_{N-1} the probability simplex in $\mathbb{R}^N$ and by α the generic element of this set: $\Delta_{N-1} = \left\{\alpha \in \mathbb{R}^N : \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1\right\}$. We let $\lambda_{\max}$ and λ_2, respectively, denote the maximal and the second largest (in modulus) eigenvalue of a square matrix; $\|\cdot\|$ denotes the spectral norm. For a matrix $S \in \mathbb{R}^{N \times N}$, we let $\mathcal{R}(S)$ denote the range of S, $\mathcal{R}(S) = \left\{Sx : x \in \mathbb{R}^N\right\}$. An open Euclidean ball in $\mathbb{R}^d$ of radius ρ and centered at x is denoted by $B_x(\rho)$; the closure, the interior, and the complement of an arbitrary set $D \subseteq \mathbb{R}^d$ are respectively denoted by $\overline{D}$, D^{o}, and D^{c}; $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel sigma algebra on $\mathbb{R}^d$; $\mathbb{P}$ and $\mathbb{E}$ denote the probability and the expectation operator; $\mathcal{N}(m,S)$ denotes Gaussian distribution with mean vector m and covariance matrix S. For a given graph H, $E(H)$ denotes the set of edges of H.

Paper organization. Section II describes the system model and the algorithm and Section III introduces the large deviations metric and defines the relevant large deviations quantities. Section IV states the main result of the paper, important corollaries and provides illustration examples. Section V provides applications of the results to social learning. Proofs of the main result are given in Section VI. Section VII concludes the paper.

II. System model

This section explains the system model and the consensus+innovations distributed inference algorithm accompanied by different application examples. Section II-A details the connection to social learning,

while Section II-B provides certain preliminaries.

Communication model. We consider a network of N identical agents connected by an arbitrary communication topology. The topology is represented by an undirected graph $\overline{G} = (V, \overline{E})$, where $V = \{1, 2, ..., N\}$ is the set of agents, and $\overline{E} \subseteq \binom{V}{2}$ is the set of possible communication links between agents. We assume that during operation of the network each link $\{i, j\} \in \overline{E}$ may fail, and that correlations between failures of different links are possible. Realization (i.e., a snapshot) of the communication topology at time slot t is denoted by $G_t = (V, E_t)$, for $t = 1, 2, \ldots$, where E_t is the set of links that are online at time t; note that $E_t \subseteq \overline{E}$. For an agent i, we let $O_{i,t}$ denote the set of neighbors of i at time t, $O_{i,t} = \{j \in V : \{i, j\} \in E_t\}$.

Consensus based distributed estimation. At each time t, each sensor i acquires a d-dimensional vector of measurements $Z_{i,t} \in \mathbb{R}^d$. We assume that the measurements $Z_{i,t}$ are independent and identically distributed across sensors and over time. The goal of each sensor is to estimate the state of nature θ, which is the expected value of sensor observations $Z_{i,t}$, $\theta = \mathbb{E}[Z_{i,t}]$. To achieve this, an agent i holds a local estimate, called also the state, $X_{i,t}$ and iteratively updates it over time slots t. At each slot t, agent i performs two steps: 1) the innovation step; and 2) the consensus step. In the innovation step, i acquires $Z_{i,t}$ and incorporates it into the current state $X_{i,t-1}$, by computing the following convex combination, forming an intermediate state:

$$\widehat{X}_{i,t} = \frac{t-1}{t} X_{i,t-1} + \frac{1}{t} Z_{i,t}. \tag{3}$$

It then subsequently transmits $\widehat{X}_{i,t}$ to (possibly, a subset of) its neighbors in $\overline{G}$, and, at the same time, receives the intermediate states $\widehat{X}_{j,t}$, $j \in O_{i,t}$, from its current neighbors. In the second, consensus, step, agent i computes the convex combination (DeGroot averaging) between its own and the neighbors estimates:

$$X_{i,t} = \sum_{j \in O_{i,t} \cup \{i\}} W_{ij,t} \widehat{X}_{j,t}, \tag{4}$$

where $W_{ij,t}$ is the weight that agent i at time t assigns to the estimate of agent j. For neat exposition, the weights of all nodes are collected in an N by N matrix W_t, such that the i, j entry of W_t equals $W_{ij,t}$, when $j \in O_{i,t} \cup \{i\}$, and equals zero otherwise. Thus, W_t respects the sparsity pattern of G_t: if $\{i, j\} \notin E_t$, then $[W_t]_{ij} = [W_t]_{ji} = 0$. Also, since the weights at each node form a convex combination, matrix W_t is stochastic. In addition, we assume that, at any time t, for any i, j, the weights are symmetric at each link, i.e., $W_{ij,t} = W_{ji,t}$, implying that W_t is symmetric.

Denoting by $\Phi(t, s) = W_t \cdots W_s$ for $1 \leq s \leq t$, algorithm (3)-(4) can be written as:

$$X_{i,t} = \frac{1}{t} \sum_{s=1}^{t} \sum_{j=1}^{N} [\Phi(t, s)]_{i,j} \, Z_{j,s}. \tag{5}$$

We analyse algorithm (3)-(4) under the following assumptions on the matrices W_t and observations $Z_{i,t}$.

Assumption 1 (Network and observations random model)**.**

1) *Observations $Z_{i,t}$, $i = 1, \dots, N$, $t = 1, 2, \dots$ are independent, identically distributed (i.i.d.) across nodes and over time;*
2) *The sequence of matrices W_t, $t = 1, 2, \dots$ is i.i.d. and for each t, every realization of W_t is stochastic, symmetric and has positive diagonals;*
3) *$\lambda_2 (\mathbb{E}[W_t]) < 1$, or, equivalently, the induced graph $\overline{G}$ of $\mathbb{E}[W_t]$ is connected.*
4) *Weight matrices W_t are independent from the nodes' observations $Z_{i,s}$ for all i, s, t.*

We now present different application examples of algorithm (3)-(4).

Example 2 (Estimating the distribution of opinions by social sampling)**.** *Consider the scenario where a group of N agents wishes to discover the distribution of opinions (e.g., about an event or phenomenon) across a certain, large population. To achieve this, agents continuously poll the population and register responses of individuals. We assume that the respondents' opinions are quantized to d preset opinion summaries: $\{r_1, ..., r_d\}$. We let $\mathcal{R}_{i,t}$ denote the opinion (summary) of the person that agent i interviewed at time t. Also, let p_l be the probability that the response of a person chosen uniformly at random is r_l. Consider now algorithm* (3)-(4) *and define the innovation vector $Z_{i,t}$ to be the vector of opinion indicators, $Z_{i,t} = \left(1_{\{\mathcal{R}_{i,t}=r_1\}}, ..., 1_{\{\mathcal{R}_{i,t}=r_d\}}\right)^\top$; again, let the W_t's be arbitrary stochastic matrices. Then, the states of all agents converge to the true opinion distribution, $(p_1, \dots, p_d)$, as we show in Section IV, i.e., algorithm* (3)-(4) *is able to correctly identify the distribution of opinions across a given population, while the rates of this convergence will prove to be highly dependent on the frequency of agents' interactions and interaction patterns.*

Example 3 (Distributed event detection)**.** *Suppose that a wireless sensor network is deployed in a certain area to detect in which of the two possible states the environment is. This problem can be modeled as a binary hypothesis testing problem, where under the state of nature (hypothesis) $\mathbf{H}_1$, the sensors measurements follow the distribution f_1, and similarly for f_0, where f_1 and f_0 are assumed known. We let $Y_{i,t}$ denote the measurement of sensor i at time t. We assume that $Y_{i,t}$'s are independent both over time and across different sensors. This hypothesis testing problem can be solved by algorithm* (3)-(4) *as follows. For each i and t, define the innovation $Z_{i,t}$ as the log-likelihood ratio of the node i's measurement at time t: $Z_{i,t} = \log \frac{f_1(Y_{i,t})}{f_0(Y_{i,t})}$. Then, any sensor in the system can, at any given time, make a decision simply by comparing its state $X_{i,t}$ against a prescribed threshold γ:*

$$X_{i,t} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \gamma. \tag{6}$$

For further details on distributed detection application, see also [14].

A generalization of the preceding example to M-ary hypothesis testing and an application to social learning is given in the next subsection.

A. Social learning

The idea of social learning is for a group of people to distinguish between M different hypotheses, potentially indistinguishable by any given individual, through local Bayesian updates and collaborative information exchange. Each node i over time draws observations $Y_{i,t}$ from (the true) distribution $f_{i,M}$ (hypothesis $\mathbf{H}_M$); the remaining $M-1$ candidate distributions that compete at node i in hypothesis testing are $f_{i,m}$ (hypothesis $\mathbf{H}_m$), $m = 1, ..., M-1$. It is assumed that, conditioned on the true hypothesis $\mathbf{H}_M$, observations at each node are independent over time, and they are also independent from the observations that are generated at any different node.

We consider here the algorithm for social learning proposed in [22]. Each node i maintains over time two sets of values (vectors), $q_{i,t} \in \mathbb{R}^M$ and $b_{i,t} \in \mathbb{R}^M$, called, respectively, *private* and *public belief* vectors, quantifying node i's beliefs in each of the M hypotheses. The m-th entry of $q_{i,t}$, denoted by $q_{i,t}^m \in \mathbb{R}$, corresponds to the private belief of node i in the m-th hypothesis; similarly, the m-th entry of $b_{i,t}$, denoted by $b_{i,t}^m \in \mathbb{R}$, corresponds to the public belief of node i in the m-th hypothesis. The values of both public and private belief vectors are between 0 and 1: the closer an entry of a belief vector is to 1 (0), the stronger (weaker) is the confidence of the respective node that the corresponding hypothesis is true; e.g., if for some m, $b_{i,t}^m$ equals 1, this means that node i is fully confident that hypothesis $\mathbf{H}_m$ is true.

The algorithm starts at each node with initial private beliefs $q_{i,t}^m > 0$, $m = 1, ..., M-1$. Upon receiving new local observation $Y_{i,t}$, each node i updates its m-th public belief as follows:

$$b_{i,t}^m = \frac{f_{i,m}(Y_{i,t}) q_{i,t-1}^m}{\sum_{l=1}^M f_{i,l}(Y_{i,t}) q_{i,t-1}^l}, \tag{7}$$

for each $m = 1, ..., M$. The node then sends its updated public belief vector $b_{i,t} = (b_{i,t}^1, ..., b_{i,t}^M)^\top$ to all of its neighbors $O_{i,t}$. Upon receiving the neighbors' (public) beliefs, the node updates its private beliefs as follows:

$$q_{i,t}^m = \frac{e^{\sum_{j \in O_{i,t}} W_{ij,t} \log b_{j,t}^m}}{\sum_{l=1}^M e^{\sum_{j \in O_{i,t}} W_{ij,t} b_{j,t}^l}}, \tag{8}$$

for each $m = 1, ..., M$.

It is easy to verify that both $q_{i,t}$ and $b_{i,t}$ represent valid probability vectors, i.e., $q_{i,t}$, $b_{i,t} \in \Delta_{M-1}$.

Connection with algorithm (3)**-**(4)**.** Consider the update for the private belief $q_{i,t}^m$ in (8). Computing the log-ratios of $q_{i,t}^m$ with $q_{i,t}^M$ (belief in the true hypothesis $\mathbf{H}_M$), the recursion in (8) transforms into:

$$\log \frac{q_{i,t}^m}{q_{i,t}^M} = \sum_{j \in O_{i,t}} W_{ij,t} \log \frac{b_{j,t}^m}{b_{j,t}^M}. \tag{9}$$

Similarly, it is easy to see that the log-ratios of the public beliefs $b_{j,t}^m$ with $b_{j,t}^M$ can be expressed as:

$$\log \frac{b_{i,t}^m}{b_{i,t}^M} = \log \frac{q_{i,t-1}^m}{q_{i,t-1}^M} + \log \frac{f_{i,m}(Y_{i,t})}{f_{i,M}(Y_{i,t})}. \tag{10}$$

Dividing both sides in (9) and (10) by t, we recognize the form in (3)-(4). Further, denoting, for each $m = 1, ..., M-1$,

$$L_{i,t}^m = \log \frac{f_{i,m}(Y_{i,t})}{f_{i,M}(Y_{i,t}} \tag{11}$$

$$\widehat{X}_{i,t}^m = \frac{1}{t} \log \frac{q_{i,t}^m}{q_{i,t}^1} \tag{12}$$

$$X_{i,t}^m = \frac{1}{t} \log \frac{b_{i,t}^m}{b_{i,t}^1} \tag{13}$$

and stacking the per-hypothesis quantities in vector form: $L_{i,t} = \left(L_{i,t}^1, ..., L_{i,t}^{M-1}\right) \in \mathbb{R}^{M-1}$, and $\widehat{X}_{i,t} = \left(\widehat{X}_{i,t}^1, ..., \widehat{X}_{i,t}^{M-1}\right) \in \mathbb{R}^{M-1}$, and $X_{i,t} = \left(X_{i,t}^1, ..., X_{i,t}^{M-1}\right) \in \mathbb{R}^{M-1}$, the exact form in (3)-(4) is obtained, where the innovation vectors $Z_{i,t}$ that algorithm (3)-(4) is fed with are the log-likelihood ratio vectors $L_{i,t}$; note also that, in this application instance, $d = M - 1$. Thus, the generic algorithmic form (3)-(4) subsumes also the social learning algorithm from (7)-(8) through the described variable transformation. Section V shows how results of this paper can be used to characterize convergence of beliefs and large deviations rates of social learning, specifically for the case when the weights $W_{ij,t}$ (neighborhoods $O_{i,t}$) in (8) are random.

B. Probabilistic rate of consensus $\mathcal{J}$

We next define certain concepts and quantities pertinent to the underlying graph process that are needed for later analyses.

Components in union graphs. Since the sequence of matrices W_t is i.i.d., the sequence G_t of their underlying topologies is i.i.d. as well. We let $\mathcal{G}$ denote the set of all topologies on V that have non-zero probability of occurrence at a given time t, i.e., $\mathcal{G} = \{(V, E) : \mathbb{P}(G_t = (V, E)) > 0\}$. For convenience, for any undirected, simple graph H on the set of vertices V we denote $p_H = \mathbb{P}(G_t = H)$. Thus, for any $H \in \mathcal{G}$, $p_H > 0$. It will also be of interest to consider different subsets of the set of feasible graphs $\mathcal{G}$. For a collection of undirected simple graphs $\mathcal{H}$ on V we let $\Gamma_{\mathcal{H}} = (V, E_{\mathcal{H}})$ denote the corresponding union graph, that is, $\Gamma_{\mathcal{H}}$ is the graph with the set of vertices V and whose edge set $E_{\mathcal{H}}$ is the union of

edge sets of all the graphs in $\mathcal{H}$, $E_{\mathcal{H}} = \cup_{H\in\mathcal{H}} E(H)$. We let $p_{\mathcal{H}}$ denote the probability that G_t belongs to $\mathcal{H}$,

$$p_{\mathcal{H}} = \sum_{H\in\mathcal{H}} p_H.$$

We also introduce – what we refer to as – the component of a node in $\mathcal{H}$.

Definition 4 (Node component in union graph)**.** *Let $\mathcal{H}$ be a given collection of undirected simple graphs on V and let $C_1, ..., C_L$ be the components of the union graph $\Gamma(\mathcal{H})$. Then, the component of node i in $\mathcal{H}$, denoted by $C_{i,\mathcal{H}}$, is the component of $\Gamma(\mathcal{H})$ that contains i: i.e., if $i \in C_l$, then $C_{i,\mathcal{H}} = C_l$.*

Probabilistic rate of consensus $\mathcal{J}$. We recall here the rate of consensus, associated with a sequence of random stochastic symmetric matrices, introduced in [13] and subsequently analyzed in [26]. In [13] and [14] we showed that the quantity $\mathcal{J}$ below, termed the rate of consensus[3], captures well how the weight matrices W_t affect performance of the estimates $X_{i,t}$ when one is concerned with large deviations metrics:

$$\mathcal{J} := -\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|W_t \cdots W_1 - J\| > \frac{1}{t} \right). \tag{14}$$

Rate of consensus $\mathcal{J}$ is computed exactly in [26].

Theorem 5 ([26])**.** *Let Assumption 1, part 2 hold. Then the* $\limsup$ *in* (14) *is in fact a limit and the rate of consensus $\mathcal{J}$ is found by*

$$\mathcal{J} = |\log p_{\mathcal{H}^\star}|,$$

where $p_{\max}$ is the probability of the most likely collection of feasible graphs whose union graph is disconnected,

$$\mathcal{H}^\star = \arg \max_{\mathcal{H}\subseteq\mathcal{G}:\, \Gamma_{\mathcal{H}} \text{ disc.}} p_{\mathcal{H}}. \tag{15}$$

In the next example we consider an important special case when links in $\overline{G}$ fail independently at random.

[3]The rate of consensus $\mathcal{J}$ (in (14)) is defined slightly differently than the corresponding quantity from [13] and [14]. In [13] and [14], in the event $\|W_t \cdots W_1 - J_N\| > 1/t$, the probability of which we wish to compute, there is a constant $\varepsilon \in (0, 1]$ in the place of $1/t$. However, as we show in [26], the two rate quantities coincide when the weight matrices are i.i.d., which is the case that we consider here.

Example 6 (Random topologies with i.i.d. link failures)**.** *Consider the random model for W_t defined by Assumption 1.2 where each link in $\overline{G}$ fails independently from other links with probability $1-p$. Applying Theorem 5, it can be shown that*

$$\mathcal{J} = \min \text{cut}\,(\overline{G})|\log(1-p)|, \tag{16}$$

where $\min \text{cut}\,(\overline{G})$ *is the minimum edge cut of the graph $\overline{G}$; for example, if $\overline{G}$ is a chain, then* $\min \text{cut}\,(\overline{G}) = 1$. *The details of this derivation can be found in [26].*

For finite time analyses, of relevance is the following variant of (14): for any $\epsilon > 0$, there exists a positive constant K_ϵ such that for all t,

$$\mathbb{P}\left(\|W_t \cdots W_s - J_N\| > \frac{1}{t}\right) \leq K_\epsilon e^{-(t-s)\,(\mathcal{J}-\epsilon)}. \tag{17}$$

III. Problem formulation: The metric of large deviations

Section II illustrate uses of algorithm (3)-(4) for several applications: multi-agent polling with cooperation, in Example 2, fully distributed hypothesis testing, in Example 3, and social learning, in Section II-A. We now introduce the rates of large deviations that we adopt as performance metric for applications of algorithm (3)-(4).

Rate function I and the large deviations principle.

Definition 7 (Rate function I [28])**.** *Function $I : \mathbb{R}^d \mapsto [0, +\infty]$ is called a* rate function *if it is lower semicontinuous, or, equivalently, if its level sets are closed. If, in addition, the level sets of I are compact (i.e., closed and bounded), then I is called a good rate function.*

Definition 8 (The large deviations principle [28])**.** *Suppose that $I : \mathbb{R}^d \mapsto [0, +\infty]$ is lower semicontinuous. A sequence of measures μ_t on $\left(\mathbb{R}^d, \mathcal{B}\left(\mathbb{R}^d\right)\right)$, $t \geq 1$, is said to satisfy the large deviations principle (LDP) with rate function I if, for any measurable set $D \subseteq \mathbb{R}^d$, the following two conditions hold:*

1) $\limsup_{t\to+\infty} \frac{1}{t} \log \mu_t(D) \leq - \inf_{x\in\overline{D}} I(x)$;
2) $\liminf_{t\to+\infty} \frac{1}{t} \log \mu_t(D) \geq - \inf_{x\in D^\circ} I(x)$.

Differently than with the case of static topologies, when topologies and/or weight matrices W_t are random, finding the rate function of an arbitrary node performing distributed inference is a very difficult

problem [14], [29]. (In fact, even the existence of the LDP is not known a priori.) Our approach is to find functions $\overline{I}_i$ and $\underline{I}_i : \mathbb{R}^d \mapsto \mathbb{R}$, such that, for any measurable set D:

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in D\right) \leq - \inf_{x\in\overline{D}} \underline{I}_i(x), \tag{18}$$

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in D\right) \geq - \inf_{x\in D^{\circ}} \overline{I}_i(x). \tag{19}$$

At a high level, this is analytically achieved by carefully constructing events the probabilities of which upper and lower bound the probability of the event of interest in (18) and (19). We remark that functions $\underline{I}_i$ and $\overline{I}_i$ that we seek should satisfy (18) and (19) for any given set D, i.e., similarly as with the rate function I_i, to find bounds on the exponential rates for a given rare event $\{X_{i,t} \in D\}$, it suffices to perform minimizations of $\underline{I}_i$ and $\overline{I}_i$ over D. This property is very important, as once $\underline{I}_i$ and $\overline{I}_i$ are discovered, any inaccuracy rate can be easily estimated without the need to do any (further) large deviations analyses.

As we show in Appendix A, if for some node i the LDP holds and (18) and (19) are satisfied for any D, then

$$\underline{I}_i(x) \leq I_i(x) \leq \overline{I}_i(x),\ x \in \mathbb{R}^d, \tag{20}$$

i.e., the graph of the LDP rate function I_i lies between the graphs of $\overline{I}_i$ and $\underline{I}_i$.

Log-moment generating function of observations $Z_{i,t}$ and its conjugate. We proceed standardly by introducing the log-moment generating function of the observation vectors $Z_{i,t}$, which we denote by Λ. The log-moment generating function $\Lambda : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ corresponding to $Z_{i,t}$ is defined by:

$$\Lambda(\lambda) = \log \mathbb{E}\left[e^{\lambda^{\top} Z_{i,t}}\right],\ \text{for}\ \lambda \in \mathbb{R}^d. \tag{21}$$

We make the assumption that Λ is finite at all points.

Assumption 9. *$\mathcal{D}_\Lambda = \mathbb{R}^d$, i.e., $\Lambda(\lambda) < +\infty$ for all $\lambda \in \mathbb{R}^d$.*

Besides the log-moment generating function Λ, the second key object in large deviations analysis is the Fenchel-Legendre transform, or the conjugate, of Λ, defined by

$$I(x) = \sup_{\lambda\in\mathbb{R}^d} x^{\top}\lambda - \Lambda(\lambda),\ \text{for}\ x \in \mathbb{R}^d. \tag{22}$$

Log-moment generating function and its conjugate enjoy many nice properties, such as convexity and differentiability in the interior of the function's domain [28], [30]. We list the properties that are relevant for the current analysis in the next lemma. Recall that $\theta = \mathbb{E}[Z_{i,t}]$.

Lemma 10 (Properties of Λ and I)**.**

1) *Λ is convex and differentiable on $\mathbb{R}^d$;*

2) $\Lambda(0) = 0$ *and* $\nabla\Lambda(0) = \theta$*;*

3) I *is strictly convex;*

4) *if* $x = \nabla\Lambda(\lambda)$ *for some* $\lambda \in \mathbb{R}^d$*, then* $I(x) = \lambda^\top x - \Lambda(\lambda)$*;*

5) $I(x) \geq 0$ *with equality if and only if* $x = \theta$*.*

Proofs of 1-5 (with a weaker form of the claim in part 3 – with strict convexity replaced by convexity, and with non-negativity only in part 5) can be found in [28]. The proof of strict convexity of I under Assumption 9 can be found in [31]. We briefly comment on properties 2 and 5, to give some (mathematical) intuition as to why these properties hold, where we note that of particular, practical relevance is 5. Plugging in $\lambda = 0$ in the defining equation of Λ, (21), it is easy to see that $\Lambda(0) = 0$. Similarly, it can be shown that, for any λ, $\nabla\Lambda(\lambda) = \mathbb{E}[Z_{i,t}e^{\lambda^\top Z_{i,t}}]/\mathbb{E}[e^{\lambda^\top Z_{i,t}}]$. Evaluating at $\lambda = 0$, the property $\nabla\Lambda(0) = \theta$ follows. Property 5 has a very intuitive meaning: the rate function is non-negative and also equals zero at the mean value. To see why the latter holds, it suffices to invoke properties from part 2 in 4; note also that, since I is non-negative, θ is a minimizer of I. The if and only if part then follows from strict convexity of I, which implies uniqueness of its minimizer θ. We will show practical implications of this property when considering large deviations rate of the sequence $X_{i,t}$.

The following result, proven in [16], gives fundamental large deviations upper and lower bound for the inference sequence $X_{i,t}$. The result holds for arbitrary stochastic weight matrices W_t and, in particular, for directed topologies as well. This result will be invoked when proving tightness and optimality of our rate function bounds for certain classes of networks, in Section IV-B.

Lemma 11 (Fundamental distributed inference bounds)**.** *Consider algorithm* (3)-(4) *under Assumptions 1 and 9. Then* (18) *and* (19) *hold with* $\overline{I}_i = NI$ *and* $\underline{I}_i = I$*, for all* i*.*

Closed convex hull of a function. We recall the definitions of the epigraph and closed convex hull of a function.

Definition 12 (Epigraph and closed convex hull of a function, [32])**.** *Let* $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ *be a given function.*

1) *The epigraph of* f*, denoted by* $\mathrm{epi} f$*, is defined by*

$$\mathrm{epi} f = \left\{(x, r) : r \geq f(x), x \in \mathbb{R}^d\right\}. \tag{23}$$

2) *Consider the closed convex hull* $\overline{\mathrm{co}}\,\mathrm{epi}\, f$[4] *of the epigraph of* f*. The closed convex hull of* f*, denoted*

[4]The convex hull of a set A, where A is a subset of some Euclidean space, is defined as the set of all convex combinations of points in A [32].

by $\overline{\text{co}}f$*, is defined by:*

$$\overline{\text{co}}f(x) := \inf\{r : (x, r) \in \overline{\text{co}}\,\text{epi}\, f\}. \tag{24}$$

Hence, for a given function f, epigraph of f is the area above the graph of f. Closed convex hull of f is then constructed from $\text{epi}f$ by first finding the closed convex hull of the epigraph, $\overline{\text{co}}\,\text{epi}\, f$. Then, $\overline{\text{co}}f$ is defined as the function the epigraph of which matches $\overline{\text{co}}\,\text{epi}\, f$. Intuitively, $\overline{\text{co}}f$ is the best convex and lower semi-continuous (closed) approximation of f, as its epigraph contains (besides $\text{epi}\, f$) only those points that are needed for "convexification" and closure. Figure 2 further ahead gives an illustration of $\overline{\text{co}}f$, while construction of $\overline{\text{co}}f$ is explained in Section IV-A.

IV. Main result

The main result of this section, Theorem 13, finds functions $\underline{I}_i$ and $\overline{I}_i$ from (18) and (19). These functions enable computation of bounds on the exponential decay rate of an arbitrary rare event and, in the case of the existence of the LDP, by (20), provide approximations to the rate function I_i. A number of important corollaries of Theorem 13 is then presented in Subsection IV-B, including the large deviations principle for regular networks and for pendant nodes. Section V then studies application of the derived results to distributed hypothesis testing and social learning.

Theorem 13. *Consider distributed inference algorithm* (3)-(4) *under Assumptions 1 and 9. Then, for each node* i*, for any measurable set* D*:*

1) $$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in D\right) \leq - \inf_{x\in\overline{D}} I^\star(x), \tag{25}$$

where $I^\star(x) = \overline{\text{co}}\, \inf\{I(x) + \mathcal{J}, NI(x)\}$*;*

2) *for any collection* $\mathcal{H}$ *of graphs on* V*:*

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in D\right) \geq - \inf_{x\in D^\circ} I_{i,\mathcal{H}}(x), \tag{26}$$

where $I_{i,\mathcal{H}}(x) = \overline{\text{co}}\, \inf\{|C_{i,\mathcal{H}}|I(x) + |\log p_{\mathcal{H}}|, NI(x)\}$.

In words, Theorem 13 asserts that, for a fixed set D, for any node i, the probabilities $\mathbb{P}\left(X_{i,t} \in D\right)$ decay exponentially fast over iterations t and it also finds bounds on the rate of this decay. We now make a couple of additional remarks and such that aim at gaining further insights and intuition about this result and the relevant quantities.

Remark 14. *Consider an arbitrary disconnected collection* $\mathcal{H}$*. By the construction of* $C_{i,\mathcal{H}}$*, for any node* i*, there holds* $\{i\} \subseteq C_{i,\mathcal{H}}$ *and, by non-negativity of* I*, it follows that* $I \leq |C_{i,\mathcal{H}}|I$ *(point-wise). Further, from Theorem 5 we know that* $\mathcal{J} = |\log p_{\mathcal{H}^\star}| \leq |\log p_{\mathcal{H}}|$*. Therefore, we have that for any disconnected*

collection $\mathcal{H}$, $I + \mathcal{J} \leq |C_{i,\mathcal{H}}|I + |\log p_{\mathcal{H}}|$. The latter obviously implies $I^\star \leq I_{i,\mathcal{H}}$, serving as a first feasibility check for (20) *(and also* (18) *and* (19)*).*

Comparing the upper bound from Theorem 13 with (18), we see that (18) is satisfied for

$$\underline{I}_i \equiv I^\star, \text{ for all } i \in V. \tag{27}$$

That is, we have a uniform (lower) bound $I^\star$ on each of the nodes' rate functions I_i, $i \in V$.

With respect to the lower bound from Theorem 13, there is in fact a whole family of functions $\overline{I}_i$, one per each collection of graphs $\mathcal{H}$, that validate (19). To find the best bound for a given D, we might optimize the right hand side of (26) over all collections $\mathcal{H}$. This, however, might be computationally infeasible. Instead, we can focus only on those collections $\mathcal{P} \subseteq \mathcal{G}$ that have a certain property, e.g., $\mathcal{P} = \{\mathcal{H} : |C_{i,\mathcal{H}}| = n\}$, for some n, $1 \leq n \leq N$. Then, $\overline{I}_i$ from (19) can be found by finding $\mathcal{H} \in \mathcal{P}$ that yields uniformly lowest (i.e., closest to I_i) $I_{i,\mathcal{H}}$:

$$\overline{I}_i = \inf_{\mathcal{H}\in\mathcal{P}} I_{i,\mathcal{H}}, \text{ for } i \in V. \tag{28}$$

The following corollary follows directly from (20) and the definition of LDP.

Corollary 15. 1) *If, for a given i, the sequence $X_{i,t}$, $t = 1, 2, ...$ satisfies the LDP with rate function I_i, then, for any collection of graphs $\mathcal{H}$,*

$$I^\star \leq I_i \leq I_{i,\mathcal{H}}. \tag{29}$$

2) *If, for a given i, for some $\mathcal{P}$ (possibly, a single element set $\mathcal{P} = \{\mathcal{H}\}$), $I^\star \equiv \inf_{\mathcal{H}\in\mathcal{P}} I_{i,\mathcal{H}}$, then the sequence $X_{i,t}$, $t = 1, 2, ...$ satisfies the LDP with rate function $I_i = I^\star \equiv \inf_{\mathcal{H}\in\mathcal{P}} I_{i,\mathcal{H}}$.*

In the next remark, through simple convex analyses, we make a connection between Corollary 15 (Theorem 13) and Lemma 11, completing the established bounds in (29) with the general bounds from Lemma 11, hence establishing a coherent view of the derived results.

Remark 16 (Recovery of fundamental bounds in Lemma 11)**.** *From the point-wise non-negativity of I and non-negativity of $\mathcal{J}$, it is easy to see that $I \leq NI$ and $I \leq I + \mathcal{J}$. Thus,* $\operatorname{epi}\inf\{NI, I + \mathcal{J}\} \subseteq \operatorname{epi} I$. *Since I is closed and convex,* $\overline{\operatorname{co}}\operatorname{epi} I = \operatorname{epi} I$*, thus implying* $\overline{\operatorname{co}}\operatorname{epi}\inf\{NI, I + \mathcal{J}\} \subseteq \operatorname{epi} I$. *The latter directly implies $I \leq I^\star$. Similarly, we have $NI \geq \inf\{NI, |C_{i,\mathcal{H}}|I + |\log p_{\mathcal{H}}|\}$, where the latter holds for any disconnected collection $\mathcal{H}$. Thus* $\operatorname{epi} NI \subseteq \operatorname{epi}\inf\{NI, |C_{i,\mathcal{H}}|I + |\log p_{\mathcal{H}}|\}$*, which in turn implies* $\overline{\operatorname{co}}\operatorname{epi} NI \subseteq \overline{\operatorname{co}}\operatorname{epi}\inf\{NI, |C_{i,\mathcal{H}}|I + |\log p_{\mathcal{H}}|\}$. *Since NI is convex and closed (the properties inherited*

from I), $\overline{\text{co}}\,\text{epi}NI = \text{epi}NI$, and therefore $\text{epi}NI = \overline{\text{co}}\,\text{epi}NI \subseteq \overline{\text{co}}\,\text{epi}\inf\{NI, |C_{i,\mathcal{H}}|I + |\log p_{\mathcal{H}}|\}$. The latter implies $NI \geq I_{i,\mathcal{H}}$. Combining with (29) *establishes:*

$$I \leq I^\star \leq I_i \leq I_{i,\mathcal{H}} \leq NI. \tag{30}$$

The above chain of inequalities is a capture of the so far established bounds in the literature on the large deviations rate function for consensus+innovations distributed inference iterates on random networks.

As a byproduct, we note in passing that (30) *verifies Lemma 11 for the special case of stochastic symmetric weight matrices.*

Remark 17 (Zero rate at θ)**.** *Since I is non-negative, both NI and $|C_{i,\mathcal{H}}|I(\theta) + |\log p_H|$ are also non-negative, implying $I_{i,\mathcal{H}} \geq 0$. Further, from Lemma 10, we have $I(\theta) = 0$, and noting now that $NI(\theta) = 0 < |C_{i,\mathcal{H}}|I(\theta) + |\log p_H|$, it follows that $I_{i,\mathcal{H}}(\theta) = 0$. It can be similarly shown that $I^\star(\theta) = 0$. From the preceding properties it follows that for any set C containing the mean value θ*

$$\inf_{x \in C} I^\star(x) = \inf_{x \in C} I_{i,\mathcal{H}}(x) = 0. \tag{31}$$

It follows that $\mathbf{I}_i(C) = 0$, i.e., the inaccuracy rate for any C containing θ equals zero. This means that probabilities of events that $X_{i,t}$ belong to C do not exhibit an exponential decay – specifically, for any norm ball centered at θ, and of an arbitrary radius $\rho > 0$, $B_\theta(\rho) > 0$, there holds

$$\lim_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in B_\theta(\rho)\right) = 0. \tag{32}$$

Observing the form of the algorithm, eq. (3)-(4)*, where innovations $Z_{i,t}$ – the mean vector of which is θ, are incorporated and mixed via weighted averaging (both over time and across nodes), it is intuitive to expect that $X_{i,t}$ will converge to θ (consider the ideal averaging case – $W_t = J_d$, for which $X_{i,t} = \frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{N}\frac{1}{N}Z_{j,s}$, which converges to θ by the law of large numbers). Hence, the zero decay in* (32) *is intuitive, i.e., the probabilities that $X_{i,t}$ belongs to a neighborhood of θ should not vanish with t.*

We use the result of Theorem 13, together with the uniqueness of the minimizer of I, property 5 from Lemma 10, to establish a sort of a converse to (32) - i.e., whenever we seek the inaccuracy rate $\mathbf{I}_i(C)$ for a set C not containing θ, this rate will be strictly positive. Practical relevance of this (technical) property is given in Theorem 19 below, where almost sure convergence of $X_{i,t}$ to θ is formally established.

Remark 18 (Strictly non-zero rate at $x \neq \theta$)**.** *Consider an arbitrary point $x \neq \theta$. From Lemma 10, part 5 we know that $I(x) > 0$ for any $x \neq \theta$.*

Consider now an arbitrary set C such that $\theta \notin C$. By strict convexity of I and uniqueness of the minimizer of I, it follows that I is coercive [33]. Pick an arbitrary point $x_0 \in C$ and let $\alpha = I(x_0)$.

Define $S_\alpha = \{x \in \mathbb{R}^d : I(x) \leq \alpha\}$, i.e., S_α is the α-level set of I. By coercivity of I, it follows that S_α is compact. We now note

$$\inf_{x \in C} I(x) = \inf_{x \in C \cap S_\alpha} I(x) =: a. \tag{33}$$

Compactness of S_α implies compactness of $C \cap S_\alpha$ and since I is continuous and strictly greater than 0, it follows by the Weierstrass theorem that the infimum of I over C is strictly greater than zero, $a = \inf_{x \in C \cap S_\alpha} I(x) > 0$. Finally, By the fact that $I^\star \geq I$ (the left-hand side inequality in (30)*), we in turn obtain:*

$$\inf_{x \in C} I^\star(x) \geq \inf_{x \in C} I(x) = a > 0. \tag{34}$$

Therefore, for any set C such that $\theta \notin C$, we have

$$\limsup_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in C\right) \leq -a < 0, \tag{35}$$

where the constant a bounding the exponential decay rate depends on the chosen set C.

With preceding considerations at hand, almost sure convergence of nodes' iterates $X_{i,t}$ follows by standard arguments.

Theorem 19 (Almost sure convergence of $X_{i,t}$)**.** *Consider distributed inference algorithm* (3)-(4) *under Assumptions 1 and 9. Then, for each node i, the state vectors $X_{i,t}$ converge almost surely to $\theta = E[Z_{i,t}]$.*

Proof. Fix node $i \in V$. Pick an arbitrary $\epsilon > 0$ and consider $C = \mathbb{B}^c_\theta(\epsilon)$. We start by noting that inequality in (35) implies existence of a finite $t_0 = t_0(C)$ such that, for all $t \geq t_0$, $\mathbb{P}(X_{i,t} \in C) \leq e^{-t\frac{a}{2}}$. Then, for all $t \geq t_0$, we have

$$\mathbb{P}\left(\|X_{i,t} - \theta\| \geq \epsilon\right) \leq e^{-t\frac{a}{2}}. \tag{36}$$

Thus,

$$\mathbb{P}\left(\|X_{i,t} - \theta\| > \epsilon, \text{ i.o.}\right) \leq \sum_{t=1}^{\infty} e^{-t\frac{a}{2}} < \infty, \tag{37}$$

where the last inequality follows from strict positivity of a. Applying the Borel-Cantelli lemma [34], the claim of the theorem follows. □

A. A closer look at functions $I^\star$ and $I_{i,\mathcal{H}}$

This subsection finds closed form expressions for the functions $I^\star$ and $I_{i,\mathcal{H}}$ for the case when $Z_{i,t}$ is a Gaussian vector, and provides a graphical interpretation of the obtained result.

Lemma 20. *Let $Z_{i,t}$ be Gaussian with mean vector m and covariance matrix S. Then*

$$I^\star(x) = \begin{cases} NI(x), & x \in \mathcal{R}_1^\star \\ N\sqrt{2c_1}\, H(x) - Nc_1, & x \in \mathcal{R}_2^\star \\ I(x) + \mathcal{J}, & x \in \mathcal{R}_3^\star \end{cases}, \tag{38}$$

where $\mathcal{R}_1^\star = \{x : NI(x) \le c_1\}$, $\mathcal{R}_2^\star = \{x : c_1 < I(x) \le Nc_1\}$, and $\mathcal{R}_3^\star = \{x : I(x) > Nc_1\}$, $I(x) = \frac{1}{2}(x-m)^\top S^{-1}(x-m)$, $H(x) = \sqrt{(x-m)^\top S^{-1}(x-m)}$, and $c_1 = \frac{\mathcal{J}}{N(N-1)}$. Also, for any fixed collection of graphs $\mathcal{H}$

$$I_{i,\mathcal{H}}(x) = \begin{cases} NI(x), & x \in \mathcal{R}_1^{i,\mathcal{H}} \\ N\sqrt{2c_2}\, H(x) - Nc_2, & x \in \mathcal{R}_2^{i,\mathcal{H}} \\ |C_{i,\mathcal{H}}|\, I(x) + |\log p_{\mathcal{H}}|, & x \in \mathcal{R}_3^{i,\mathcal{H}} \end{cases}, \tag{39}$$

where $\mathcal{R}_1^{i,\mathcal{H}} = \left\{x : \frac{N}{|C_{i,\mathcal{H}}|} I(x) \le c_2\right\}$, $\mathcal{R}_2^{i,\mathcal{H}} = \left\{x : c_2 < I(x) \le \frac{N}{|C_{i,\mathcal{H}}|} c_2\right\}$, $\mathcal{R}_3^{i,\mathcal{H}} = \left\{x : I(x) > \frac{N}{|C_{i,\mathcal{H}}|} c_2\right\}$, and $c_2 = \frac{|C_{i,\mathcal{H}}||\log p_{\mathcal{H}}|}{N(N-|C_{i,\mathcal{H}}|)}$.

Proof of Lemma 20 is given in Appendix B.

Three regions of $I^\star$. We provide a graphical illustration for $I^\star$ in Figure 2. We consider an instance of algorithm (3)-(4) running on a $N = 3$-node chain, with i.i.d. link failures of probability $(1-p) = e^{-5}$, and where the observations $Z_{i,t}$ are standard Gaussian (zero mean and variance equal to one). For standard Gaussian, $I(x) = \frac{1}{2}x^2$, and we obtain from Example 6 that the rate of consensus equals $\mathcal{J} = |\log(1-p)| = 5$. The more curved blue dotted line plots the function $NI(x) = \frac{1}{2}Nx^2$, the less curved blue dotted line plots the function $I(x) + \mathcal{J} = \frac{1}{2}x^2 + 5$, and the solid red line plots $I^\star$. Observing the figure and the corresponding formula (38), we see that $I^\star$ is defined by three regions. In the region around the zero mean, $\mathcal{R}_1^\star$, $I^\star$ matches the optimal rate function NI. On the other hand, in the outer region, $\mathcal{R}_3^\star$, where values of x are sufficiently large, $I^\star$ follows the slower growing function, $I + \mathcal{J}$. Finally, in the middle region, $\mathcal{R}_2^\star$, $I^\star$ is linear (more generally, when $d > 1$, $I^\star$ will exhibit linear intervals over any direction that crosses the mean value). This linear part is the tangent line that touches both the epigraph of $NI(\cdot)$ and the epigraph of $I + \mathcal{J}$ and is responsible for the convexification of the point-wise infimum $\inf\{I + \mathcal{J}, NI\}$. Function $I_{i,\mathcal{H}}$ has similar properties.

B. Illustrations and LDP for special cases

In this subsection, we use Theorem 13 to establish the LDP for certain classes of random models. As explained in the remarks after Theorem 13, to prove the LDP at some node i, it is sufficient to show that $I^\star$ and $I_{i,\mathcal{H}}$ coincide for some collection $\mathcal{H}$.

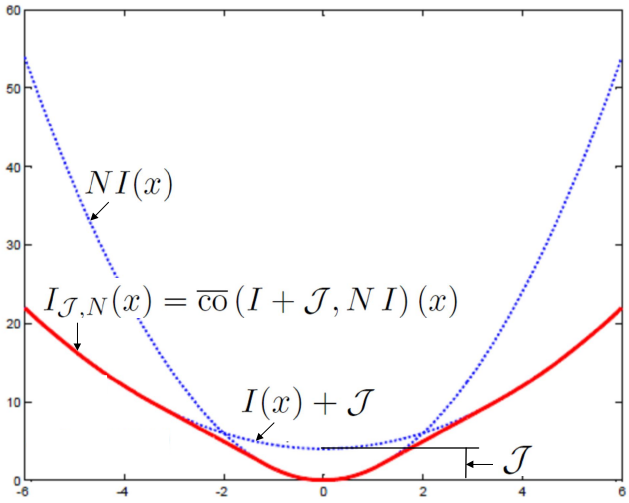


Fig. 2: Illustration of $I^\star$ for a chain network of size $N = 3$, with $\mathcal{J} = 5$, and $Z_{i,t} \sim \mathcal{N}(0, 1)$. The more curved blue dotted line plots $NI(x) = \frac{1}{2}Nx^2$, the less curved blue dotted line plots $I(x)+\mathcal{J} = \frac{1}{2}x^2+\mathcal{J}$. The solid red line plots $I^\star = \text{co}\,(NI, I + \mathcal{J})$.

The first corollary of Theorem 13 asserts that if every realization of the network topology is connected, then, for any node i, the sequence of states $X_{i,t}$ satisfies the LDP with rate function NI. In our recent work [16], we prove that NI is the best (highest) possible rate function for any distributed inference algorithms of the form (3)-(4) with N nodes. It is also the rate function of a hypothetical fusion node that has access to all the observations. Thus, when every instance of the network topology is connected, then each node in the network is, in the asymptotic sense, effectively acts as a fusion center. Corollary 21 was, for the special case of Gaussian observations, previously proved in [35].

Corollary 21. *Let, for each t, G_t be connected. Then, for any $i \in V$, $X_{i,t}$ satisfies the large deviations principle with rate function NI.*

Proof. By Theorem 2 from [16], we know that, for any node i and for any set D

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\,(X_{i,t} \in D) \geq - \inf_{x\in D^\circ} NI(x). \tag{40}$$

Comparing with the conditions for LDP in Definition 8, we see that we only need to prove that $I^\star \equiv NI$. For the latter identity it suffices to show that $\mathcal{J} = +\infty$, because then $\inf\{NI, I + \mathcal{J}\} \equiv NI$, and since NI is closed and convex, we obtain $I^\star = \overline{\text{co}}(NI) = NI$. Suppose for the sake of contradiction that there exists a disconnected collection of graphs $\mathcal{H}$ such that $p_{\mathcal{H}} > 0$. Then, there must be a graph $H \in \mathcal{H}$ such that both H is disconnected and $p_H > 0$. But this contradicts the assumption that every possible

(i.e., non-zero probability) topology is connected. Thus, it must be that for every disconnected collection $p_{\mathcal{H}} = 0$ implying $\mathcal{J} = +\infty$, and proving the claim. □

In particular, Corollary 21 implies that if the nodes' interactions are deterministic, i.e., $W_t \equiv A$, for some stochastic symmetric A, and A is such that $|\lambda_2(A)| < 1$, then, for each i, $X_{i,t}$ satisfy the LDP with the optimal rate function NI. This recovers the large deviations principle for deterministic networks, established in [16], for the special case of symmetric networks (cf. Theorem 1 in [16]).

LDP for critical nodes. Consider now a situation when there exists a node i such that $\mathcal{J} = |\log p_{i,\text{isol}}|$, where $p_{i,\text{isol}}$ denotes the probability that i operates in isolation due to network randomness, $p_{i,\text{isol}} = \mathbb{P}(O_{i,t} = \emptyset)$. Comparing with Theorem 5, this means that the most likely way to disconnect $\overline{G}$ is to isolate i, i.e.,

$$p_{\max} = \sum_{H \in \mathcal{H}_{i,\text{isol}}} p_H, \tag{41}$$

where $\mathcal{H}_{i,\text{isol}} = \{H : p_H > 0,\ C_{i,H} = \{i\}\}$. Since $C_{i,\mathcal{H}_{i,\text{isol}}} = \{i\}$, we have $\left|C_{i,\mathcal{H}_{i,\text{isol}}}\right| = 1$. Consider now the lower bound in (26) for $\mathcal{H} = \mathcal{H}_{i,\text{isol}}$. Noting that $\left|p_{\mathcal{H}_{i,\text{isol}}}\right| = \mathcal{J}$, we see that the two functions $I^\star$ and $I_{i,\mathcal{H}_{i,\text{isol}}}$ coincide, thus implying the LDP for node i. This is formally stated in the next corollary.

Corollary 22 (LDP for critical nodes)**.** *Suppose that for some i, $\mathcal{J} = |\log p_{i,\text{isol}}|$. Then, the sequence of states $X_{i,t}$ satisfies the LDP with the rate function $\overline{\text{co}}\{NI(x), I(x) + |\log p_{i,\text{isol}}|\}$.*

In the next two corollaries we assume the random model from Assumption 1.2 where each link in the graph $\overline{G}$ fails independently with the same probability $1 - p$, $p \in [0, 1]$.

Corollary 23 (LDP for pendant nodes)**.** *Suppose that the random model for W_t is such that all links in $\overline{E}$ fail independently from each other with probability $1-p$. Then, for any node i whose degree in $\overline{G}$ is equal to one, its sequence of states $X_{i,t}$ satisfies the LDP with the rate function $\overline{\text{co}}\{NI(x), I(x) + |\log(1-p)|\}$.*

Proof. Suppose that i is a degree one node. By Corollary 22, it suffices to show that $\mathcal{J} = |\log(1-p)|$. From Example 6, we know that $\mathcal{J}$ equals $|\log(1-p)|$ times the minimum edge cut of $\overline{G}$. In this case, minimum edge cut equals one (and is achieved, for instance, when the edge adjacent to i is removed from the network), which proves the result. □

Corollary 24 (LDP for regular networks)**.** *Suppose that $\overline{G}$ is a circulant network in which each node is connected to $d/2$ nodes on the left and $d/2$ nodes on the right, where $d \leq N-1$ is even. We assume that each link, independently of all other links, fails with probability $1-p$. Then, for any node i its sequence of states $X_{i,t}$ satisfies the LDP with the rate function $\overline{\text{co}}\{NI, I + d\log|1-p|\}$.*

Proof. Note that $p_{i,\text{isol}} = (1-p)^d$ for any i. Hence, by Corollary 22, it suffices to show that $\mathcal{J} = d|\log(1-p)|$. Observing that the minimum cut in this case equals d, the result follows. □

V. Application to distributed hypothesis testing and social learning

In this subsection we show how results from Section IV can be used to characterize large deviations rates of distributed hypothesis testing and social learning that are run over random networks. We recall the algorithm and relevant quantities defined in Section II-A. We assume that the measurement distributions corresponding to the same hypothesis are equal across all nodes, i.e., when hypothesis $\mathbf{H}_m$ is true, the measurements at all nodes are drawn from the same distribution f_m: $Y_{i,t} \sim f_{i,m} \equiv f_m$, for all i.

Following the identified role of the vector of log-likelihood ratios $L_{i,t}$ as the innovation vector $Z_{i,t}$ in (3)-(4), we introduce the log-moment generating function Λ_M of $L_{i,t}$ at node i, when the measurements are drawn from f_M (hypothesis $\mathbf{H}_M$ is true):

$$\Lambda_M(\lambda) = \mathbb{E}\left[e^{\lambda^\top L_{i,t}} \middle| \mathbf{H} = \mathbf{H}_M\right] \tag{42}$$

$$= \mathbb{E}\left[e^{\sum_{m=1}^{M-1} \lambda_m \log \frac{f_m(Y_{i,t})}{f_M(Y_{i,t})}} \middle| \mathbf{H} = \mathbf{H}_M\right], \tag{43}$$

for $\lambda = (\lambda_1, ..., \lambda_{M-1})^\top \in \mathbb{R}^{M-1}$; we note that index M in Λ_M indicates the dependence on the assumed true distribution f_M. Similarly as in Section III, the conjugate of Λ_M is denoted by I_M. We assume that Λ_M satisfies Assumption 9.

A. Large deviations rates of the belief log-ratios

The following result follows as a direct application of Theorem 13 to the log-ratios $X_{i,t}$ of public beliefs, defined in Section II-A, eq. (13), $X_{i,t}^m = \frac{1}{t}\log\frac{b_{i,t}^m}{b_{i,t}^M}$, $m = 1, ..., M-1$.

Theorem 25. *Consider the social learning algorithm* (7)-(8) *under Assumptions 1 and 9, for* $\Lambda = \Lambda_M$. *Then, when* $\mathbf{H} = \mathbf{H}_M$, *for each node* i, *for any measurable set* D,

1) $$\limsup_{t\to+\infty} \frac{1}{t}\log \mathbb{P}\left(X_{i,t} \in D\right) \leq -\inf_{x\in\overline{D}} I_M^\star(x), \tag{44}$$

where $I_M^\star(x) = \overline{\text{co}}\ \inf\{I_M(x) + \mathcal{J}, NI_M(x)\}$;

2) *for any collection* $\mathcal{H}$ *of graphs on* V:

$$\liminf_{t\to+\infty} \frac{1}{t}\log \mathbb{P}\left(X_{i,t} \in D\right) \geq -\inf_{x\in D^\circ} I_{i,\mathcal{H};M}(x), \tag{45}$$

where $I_{i,\mathcal{H};M}(x) = \overline{\text{co}}\ \inf\{|C_{i,\mathcal{H}}|I_M(x) + |\log p_{\mathcal{H}}|, NI_M(x)\}$.

Consequently, all considerations, corollaries and results from Section IV also carry over without any changes for the log-ratios $X_{i,t}$ of beliefs in social learning. In particular, the LDP results for regular networks and pendant nodes also carry over to the social learning setup.

Theorem 26 (Almost sure convergence of $X_{i,t}$ in social learning)**.** *Consider the social learning algorithm* (7)-(8) *under Assumptions 1 and 9, for $\Lambda = \Lambda_M$. Then, for each node i, for each $m = 1, ..., M-1$, $\frac{1}{t} \log \frac{b^m_{i,t}}{b^M_{i,t}}$ converges almost surely to $-D_{KL}(f_M||f_m) = -\mathbb{E}\left[\log \frac{f_m(Y_{i,t})}{f_M(Y_{i,t}} \middle| \mathbf{H} = \mathbf{H}_M\right]$.*

The result follows as a direct application of Theorem 19 for the case when the innovations $Z_{i,t}$ in (3)-(4) are instantiated by the log-likelihood ratios $L_{i,t}$ defined in (11), $L^m_{i,t} = \log \frac{f_m(Y_{i,t})}{f_M(Y_{i,t}}$, for $m = 1, ..., M-1$, and by recognizing that the expected value of $\log \frac{f_m(Y_{i,t})}{f_M(Y_{i,t}}$ under distribution f_M is the negative of the KL divergence between f_m and f_M.

To illustrate the setup and the relevant quantities, we consider the example of M scalar Gaussian distributions of different mean values and equal variances.

Example 27 (Gaussian case: different mean values and equal variances)**.** *Let $Y_{i,t}$ be Gaussian scalars, with mean value μ_m under hypothesis m, and (equal) variance σ^2. It is easy to show that, for this case, $L_{i,t}$ is computed as:*

$$L_{i,t} = \frac{1}{\sigma^2} \left(Y_{i,t} - \mu_M\right) d - D_{\mathrm{KL}}, \tag{46}$$

where $d = (d_1, ..., d_{M-1})^\top$, and each $d_m = \mu_m - \mu_M$ is the difference between the mean value for the m-th hypothesis and the mean value for the true hypothesis, and $D_{\mathrm{KL}} = (D_{\mathrm{KL},1}, ..., D_{\mathrm{KL},M-1})^\top$, where $D_{\mathrm{KL},m} = \frac{(\mu_m - \mu_M)^2}{2\sigma^2}$ is the KL divergence between the distribution f_m and the true distribution f_M, $m = 1, ..., M-1$. It is easy to see that, for each $i = 1, ..., N$ and each t, $L_{i,t}$ is Gaussian with mean vector $-D_{\mathrm{KL}}$ and covariance matrix $\frac{1}{\sigma^2} dd^\top$. Using the standard formula for the log-moment generating function of multivariate Gaussian distribution, we get:

$$\Lambda_M(\lambda) = -\lambda^\top D_{\mathrm{KL}} + \frac{(\lambda^\top d)}{2\sigma^2}. \tag{47}$$

Simple calculus shows that the conjugate function I_M is given by:

$$I_M(x) = \begin{cases} \frac{\zeta^2}{2\sigma^2}, & \text{if } x = \frac{\zeta}{2\sigma^2} d - D_{\mathrm{KL}}, \text{ for some } \zeta \in \mathbb{R} \\ +\infty, & \text{if } x + D_{\mathrm{KL}} \notin \mathrm{span}(d) \end{cases}. \tag{48}$$

Thus, I_M is essentially a one-dimensional quadratic function that changes only along the direction $-D_{KL} + \alpha d$, $\alpha \in \mathbb{R}$, while being equal to $+\infty$ in the rest of the $\mathbb{R}^d$ space. This is intuitive as the log-likelihood ratios for different m are coupled through a common (scalar) variable $Y_{i,t}$, and hence the events that vector $L_{i,t}$ lies outside of the line $-D_{KL} + \alpha d$ must have zero probability (and thus rate function equal to $+\infty$). The convex conjugates of I_M from Theorem 13, $I^\star_M$ and $I_{i,\mathcal{H};M}$, can be found similarly as in Section IV-A.

B. Large deviations rates for beliefs in social learning

For each $m = 1, .., M-1$, define $g_m : \mathbb{R}^{M-1} \mapsto \mathbb{R}$ as $g_m(x) = x_m - \max\{0, x_1, ..., x_{M-1}\}$, for $x \in \mathbb{R}^d$.

Theorem 28. *Consider the social learning algorithm* (7)-(8) *under Assumptions 1 and 9, for* $\Lambda = \Lambda_M$. *Then, for each node* $i \in V$ *and hypothesis* $m = 1, ..., M-1$, *for any given interval* $F \subseteq \mathbb{R}$*:*

1)

$$\limsup_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(\frac{1}{t} \log b_{i,t}^m \in F\right) \leq - \inf_{x: g_m(x) \in F} I_M^\star(x); \tag{49}$$

2) *for any disconnected collection* $\mathcal{H}$*,*

$$\liminf_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(\frac{1}{t} \log b_{i,t}^m \in F\right) \geq - \inf_{x: g_m(x) \in F} I_{i,\mathcal{H};M}(x). \tag{50}$$

The proof is very similar to the proof of Lemma 4 from [22]. The key distinction is that here full LDP for the log-ratios of the beliefs, $X_{i,t}$, is not available due to the complexity of the network model, and we have to work instead with the upper and the lower rate function bounds. However, the key arguments remain unaltered. For completeness, we provide the main steps of the proof in Appendix C.

The result in Theorem 28 is very general, as it holds for arbitrary distributions f_m, $m = 1, ..., M$, such that the log-moment generating function Λ_M satisfies Assumption 9; this is for example the case for Gaussian distributions from Example 27.

Remark 29. *It can be shown by carrying out the same analyses as in the proof of Theorem 28, that, if for some node* i *the sequence* $X_{i,t}$ *satisfies the LDP with rate function* I_i*, then, for each* $m = 1, ..., M$ *the sequence of log beliefs* $\frac{1}{t} \log b_{i,t}^m$ *also satisfies the LDP with rate function*

$$R_{i,m}(z) = \inf_{x:\, g_m(x) = z} I_i(x), \tag{51}$$

for $x \in \mathbb{R}$.

We can see that, to find the large deviations rates of the beliefs, first the rate function I_i (or bounds on this function) for the log-belief ratios $X_{i,t}$ are found, and then the contraction principle is applied with functions g_m acting as the bridge between the two domains. This relation is established in [22] for static networks, but the same behaviour carries over to the general case, with the difference that the rate function of log-beliefs can differ across different nodes as a result of network randomness. To shed some light on function g_m, we revisit Example 27 for which we derive a closed form expression for g_m.

Example 30 (Computation of g_m for the Gaussian case)**.** *Consider the setup from Example 27. Recall that* $g_m(x) = x_m - \max\{0, x_1, ..., x_{M-1}\}$ *and also that* $I_M(x) = +\infty$ *outside of the line* $-D_{KL} + \alpha d$, $\alpha \in \mathbb{R}$. *Define also*

$$f(\zeta) = \max\{0, \zeta\frac{d_1}{\sigma^2} - D_{KL,1}, ..., \zeta\frac{d_{M-1}}{\sigma^2} - D_{KL,M-1}\}. \tag{52}$$

and note that

$$g_m(x) = \zeta\frac{d_m}{\sigma^2} - D_{KL,m} - f(\zeta), \tag{53}$$

for any $\zeta \in \mathbb{R}$ *and* $x \in \mathbb{R}^{M-1}$ *such that* $x = (\zeta\frac{d_1}{\sigma^2} - D_{KL,1}, ..., \zeta\frac{d_{M-1}}{\sigma^2} - D_{KL,M-1})$, *for* $m = 1, ..., M-1$.

Without loss of generality, assume that $\mu_1 < \mu_2 < ... < \mu_{M-1}$, *implying also* $d_1 < d_2 < ... < d_{M-1}$. *Let* $m^\star$ *be the largest* m *such that* $\mu_m < \mu_M$, $m^\star = \max\{m \in \{0, 1, ..., M-1\} : \mu_m < \mu_M\}$, *wherein we additionally define* $\mu_0 \equiv -\infty$ *to account for the case that* $\mu_M < \mu_1$. *Then* $d_1 < ... < d_m^\star < 0 < d_{m^\star+1} < ... < d_{M-1}$. *By the preceding ordering, and exploiting also that* $D_{KL,m} = \frac{d_m^2}{2\sigma^2}$, *it can be easily verified that for any pair* $l < m$, *the intersection between the lines* $\zeta\frac{d_l}{\sigma^2} - D_{KL,l}$ *and* $\zeta\frac{d_m}{\sigma^2} - D_{KL,m}$ *occurs at* $\frac{d_l+d_m}{2}$, *with the* l*-indexed line dominating to the left of this point, for* $\zeta < \frac{d_l+d_m}{2}$, *while the* m*-indexed line dominates to the right. It also clearly follows that the first intersection point occurs for the first neighboring index, thus, as* ζ *increases, the lines must dominate in the same order as their* d_m *values. Summarizing,* f *is given in the following form:*

$$f(\zeta) = \begin{cases} \zeta\frac{d_1}{\sigma^2} - D_{KL,1}, & \zeta < \frac{d_1+d_2}{2} \\ \zeta\frac{d_2}{\sigma^2} - D_{KL,2}, & \frac{d_1+d_2}{2} \leq \zeta < \frac{d_2+d_3}{2} \\ ... & \\ 0 & \frac{d_{m^\star}}{2} \leq \zeta < \frac{d_{m^\star+1}}{2} \\ ... & \\ \zeta\frac{d_m}{\sigma^2} - D_{KL,m}, & \frac{d_{m-1}+d_m}{2} \leq \zeta < \frac{d_m+d_{m+1}}{2} \\ ... & \\ \zeta\frac{d_{M-1}}{\sigma^2} - D_{KL,M-1}, & \zeta \geq \frac{d_{M-2}+d_{M-1}}{2} \end{cases}. \tag{54}$$

From (53) *and* (54)*, we can obtain for* $x = \frac{\zeta}{\sigma^2} d - D_{KL}$*:*

$$g_m(x) = \begin{cases} \frac{(d_m - d_1)}{\sigma^2}\left(\zeta - \frac{d_m + d_1}{2}\right) & \zeta < \frac{d_1+d_2}{2} \\ \zeta \frac{d_2}{\sigma^2} - D_{KL,2} & \frac{d_1+d_2}{2} \leq \zeta < \frac{d_2+d_3}{2} \\ \ldots & \\ \zeta \frac{d_m}{\sigma^2} - D_{KL,m} & \frac{d_{m^\star}}{2} \leq \zeta < \frac{d_{m^\star+1}}{2} \\ \ldots & \\ 0 & \frac{d_{m-1}+d_m}{2} \leq \zeta < \frac{d_m+d_{m+1}}{2} \\ \ldots & \\ \frac{(d_m - d_{M-1})}{\sigma^2}\left(\zeta - \frac{d_m + d_{M-1}}{2}\right) & \zeta \geq \frac{d_{M-2}+d_{M-1}}{2} \end{cases}. \tag{55}$$

The derived closed form expression for g_m *is a step towards deriving the closed form expression for the rate function* $R_{i,m}$*, and, in particular, it suggests an analytical validation for the piece-wise behaviour of the rate function of beliefs discovered numerically in [22], Figure 9. This is out of scope of the current paper and is left for future work. To provide an illustration towards characterizing* $R_{i,m}$*, we consider the value of the rate function at* $-D_{KL,m}$*. From* (55)*, we see that* $g_m(x) = -D_{KL,m}\}$ *for* $x = \frac{\zeta}{\sigma^2} - D_{KL}$ *and* $\zeta = 0$ *(note that, by construction,* $d_{m^\star} < 0$ *and* $d_{m^\star+1} > 0$*, and hence* $\zeta = 0 \in [\frac{d_{m^\star}}{2}, \frac{d_{m^\star+1}}{2})$*). Thus, we have*

$$R_{i,m}(-D_{KL,m}) = \inf_{x: g_m(x) = -D_{KL,m}} I_i(x) \leq I_i(-D_{KL}). \tag{56}$$

the preceding inequality holds trivially by the fact that $-D_{KL} \in \{x : g_m(x) = -D_{KL,m}\}$*. On the other hand, we have proved that* $I^\star_M(-D_{KL}) = I_{i,\mathcal{H};M}(-D_{KL}) = 0$ *(see Remark 17). By* (29)*, we thus have* $I_i(-D_{KL}) = 0$*. It follows that* $R_{i,m}(-D_{KL,m}) = 0$*, i.e., the derived expression for* g_m *reveals that the value of the rate function* $R_{i,m}$ *at* $-D_{KL,m}$ *is zero. This is in accordance with almost sure convergence of* $\frac{1}{t}\log b^m_{i,t}$ *to* $-D_{KL,m}$ *which follows by combining Theorems 26 and 32.*

When the two functions from (44) and (45), namely, $I^\star_M$ and $I_{i,\mathcal{H};M}$ match, this implies that the corresponding $\limsup$ and the $\liminf$ are equal. Hence, whenever for a given node i its sequence $X_{i,t}$ exhibits LDP, this implies LDP for the sequence of beliefs $\frac{1}{t}\log b^m_{i,t}$, for each $m = 1, ..., M-1$. Here we give an example for regular networks.

Corollary 31 (LDP for social learning in regular networks)**.** *Suppose that* $\overline{G}$ *is a circulant network as in Corollary 24, i.e., each node is connected to* $d/2$ *nodes on the left and* $d/2$ *nodes on the right, where* $d \leq N-1$ *is even. We assume that each link, independently of all other links, fails with probability* $1-p$*. Then, for any node* i*, for each* m*,* $\frac{1}{t}\log b^m_{i,t}$ *satisfies the LDP with the rate function*

$$R_m(z) = \inf_{x \in \mathbb{R}^{M-1}: g_m(x) = z} \overline{\text{co}}\left\{NI_M, I_M + d\log|1-p|\right\}(x). \tag{57}$$

A similar result holds also for pendant nodes with i.i.d. link failures.

Convergence to the correct hypothesis. The next result establishes, through the use of large deviations analysis, that the social learning algorithm (7)-(8) correctly identifies the true hypothesis. We remark that this recovers the result of [27] for the special case of identical distributions across nodes.

Theorem 32. *Consider the social learning algorithm* (7)-(8) *under Assumption 1 and 9, for* $\Lambda = \Lambda_M$. *Then, when* $\mathbf{H} = \mathbf{H}_M$, *for each node* i, *the sequence of beliefs* $b_{i,t}^M$ *converges to one almost surely.*

Proof. From the construction of the beliefs $b_{i,t}^m$, for each i, t, $b_{i,t}^M = 1 - b_{i,t}^1 - ... - b_{i,t}^{M-1}$. Combining this with the relations $X_{i,t}^m = \frac{1}{t} \log \frac{b_{i,t}^m}{b_{i,t}^M}$, yields

$$b_{i,t}^M = \frac{1}{1 + \sum_{m=1}^{M-1} e^{tX_{i,t}^m}}. \tag{58}$$

By Theorem 26, for each $m = 1, ..., M-1$, $X_{i,t}^m$ converges almost surely to $-D_{KL}(f_M || f_m) < 0$. Hence, each of the terms $e^{tX_{i,t}^m}$ in the sum above vanishes with probability one. Since M is finite, there exists a set of probability one such that $\sum_{m=1}^{M} e^{tX_{i,t}^m}$ vanishes, proving that $b_{i,t}^M$ converges to one almost surely. □

The next two sections prove Theorem 13; Section VI-A proves the upper bound (25) and Section VI-B proves the lower bound (26).

VI. Proof of Theorem 13

This section proves Theorem 13 by proving separately the upper and the lower bound. Before giving the respective proofs, we first give some important lemmas that are used in both the upper and the lower bound proof.

Lemma 33 will be used to find the log-moment generating function of the estimate $X_{i,t}$ from the log-moment generating functions of each of the terms in the sum (5). This result follows from convexity and zero value at the origin property of Λ.

Lemma 33. *For any set of convex multipliers* $\alpha \in \Delta_{N-1}$, *for each* $j = 1, ..., N$, *the log-moment generating function* Λ *satisfies,*

$$N\Lambda\left(1/N\lambda\right) \leq \sum_{i=1}^{N} \Lambda(\alpha_i \lambda) \leq \Lambda(\lambda), \tag{59}$$

for any $\lambda \in \mathbb{R}^d$.

The proof of Lemma 33 can be found in [16].

The claims in Lemma 34 are standard results from convex analysis, the proofs of which can be found, e.g., in [32]. Let the superscript * denote the conjugacy operation, i.e., for a given function $f : \mathbb{R}^d \mapsto \mathbb{R}$,

$$f^\star(x) = \sup_{s\in\mathbb{R}^d} s^\top x - f(s), \quad x \in \mathbb{R}^d. \tag{60}$$

The following relations hold between a function f and its conjugate $f^\star$.

Lemma 34. 1) *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a given a function. Then:*

a) $[f(\cdot) + r]^\star = f^\star(\cdot) - r$;

b) *for $\alpha > 0$ and $\beta \neq 0$,* $[\alpha f(\beta(\cdot))]^\star = \alpha f^\star\left(1/(\alpha\beta)(\cdot)\right)$.

2) *Let f_1 and f_2 be two given functions. Then, the conjugate of the pointwise supremum of f_1 and f_2 is the convex hull of the pointwise infimum of $f_1^\star$ and $f_2^\star$:*

$$[\sup\{f_1, f_2\}]^\star = \overline{\mathrm{co}}\inf\left\{f_1^\star, f_2^\star\right\}. \tag{61}$$

A. Proof of the upper bound (25)

In our previous work [16], we have proved that, at any node i, the sequence $X_{i,t}$ is exponentially tight. This intuitively means that the probabilities of the tail events of $X_{i,t}$ vanish sufficiently fast (i.e., the exponential rates of the tail probabilities grow unbounded when the tails move to infinity). Lemma 35 uses this result to derive an elegant sufficient condition for a certain function to satisfy the large deviations upper bound from Definition 8. In our case, this function will be the conjugate of a certain modification of Λ that accounts for the effects of intermittent communications. We remark that at the core of the proof of Lemma 35 is a modification of the finite cover argument from the proof of Cramér's theorem in $\mathbb{R}^d$ (see, e.g., [28]); the detailed proof of Lemma 35 is provided in Appendix D.

Lemma 35. *Let X_t be an arbitrary sequence of random variables where each X_t takes values in $\mathbb{R}^d$. Suppose that for some function f, for any measurable set D there holds*

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_t \in D\right) \leq f(\lambda) - \inf_{x\in D} \lambda^\top x, \tag{62}$$

for any $\lambda \in \mathbb{R}^d$. Then, if f is finite for all $\lambda \in \mathbb{R}^d$, for any compact set F

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_t \in F\right) \leq - \inf_{x\in F} f^\star(x), \tag{63}$$

where $f^\star$ is the conjugate of f. If in addition X_t is exponentially tight, then (63) *holds for any closed set F.*

Fix an arbitrary node $i \in V$. Replicating the steps of the proof of Theorem 5 from [14], we obtain that, for any measurable set D, and any fixed $\lambda \in \mathbb{R}^d$,

$$\begin{aligned}&\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in D\right)\\ &\leq \max\left\{N\Lambda\left(\frac{1}{N}\lambda\right), \Lambda(\lambda) - \mathcal{J}\right\} - \inf_{x\in D} \lambda^\top x. \end{aligned} \tag{64}$$

By Lemma 17 from [16], the sequence of estimates $X_{i,t}$ is exponentially tight. (We remark that this result is proven under more general assumptions on the weight matrices than assumed here.) Hence, to prove the upper bound (25), it only remains to show that $I^\star$ from Theorem 13 is the conjugate of $f(\lambda) := \max\left\{N\Lambda\left(1/N\lambda\right), \Lambda(\lambda) - \mathcal{J}\right\}$, $\lambda \in \mathbb{R}^d$. From part 2 of Lemma 34, we have that the conjugate of f is the closed convex hull of the infimum of the conjugates of $f_1(\lambda) := \lambda \mapsto N\Lambda\left(1/N\lambda\right)$ and $f_2(\lambda) := \lambda \mapsto \Lambda(\lambda) - \mathcal{J}$. Using the conjugacy rules from parts 1b and 1a of Lemma 34, we obtain that the respective conjugates of f_1 and f_2 are $NI(x)$, $x \in \mathbb{R}^d$, and $I(x) + \mathcal{J}$, $x \in \mathbb{R}^d$. The upper bound 25 follows by part 2 of Lemma 34.

B. Proof of the lower bound (26)

Fix an arbitrary node $i \in V$. Fix a collection of feasible graphs $\mathcal{H}$. To simplify the notation, we denote the component of i in $\mathcal{H}$, $C_{i,\mathcal{H}}$, by C. We also let M denote the number of nodes in C, $M = |C|$. For each fixed t, we define the family of events $\left\{\mathcal{E}^t_\theta : \theta \in [0,1]\right\}$, such that for any $\theta \in [0,1]$,

$$\begin{aligned}\mathcal{E}^t_\theta = \Bigg\{&G_s \in \mathcal{H},\ \lceil \theta t\rceil \leq s \leq t,\ \ \|[\Phi(t, t-o_t)]_C - J_M\| \leq \frac{1}{t},\\ &\|\Phi(\lceil \theta t\rceil, \lceil \theta t\rceil - o_t) - J_N\| \leq \frac{1}{t}\Bigg\},\end{aligned} \tag{65}$$

where $o_t = \lceil \log t\rceil$; we recall that, for a square matrix A, A_C denotes the block of A corresponding to the intersection of columns and rows of A the indices of which belong to C. For convenience, we introduce $\mathcal{T}_\theta = \{\lceil \theta t\rceil, ..., t\}$.

Lemma 36. *Let θ be an arbitrary number in $[0,1]$. For any $\omega \in \mathcal{E}^t_\theta$,*

1) *for any $s \in \mathcal{T}_\theta$,*
$$[\Phi(t,s)]_{ij} = 0, \ \textit{for}\ \ j \notin C;$$
2) *for $t - o_t \geq s \geq \lceil \theta t\rceil$,*
$$\left|[\Phi(t,s)]_{ij} - \frac{1}{M}\right| \leq \frac{1}{t}, \ \textit{for all}\ \ j \in C;$$
3) *for $\lceil \theta t\rceil - o_t \geq s \geq 1$,*
$$\left|[\Phi(t,s)]_{ij} - \frac{1}{N}\right| \leq \frac{1}{t}, \ \textit{for all}\ \ j \in V.$$

Proof. Fix $\omega \in \mathcal{E}_\theta^t$ and, for $s = 1, ..., t$, denote $A_s = W_s(\omega)$. Consider first part 1, and suppose, without loss of generality, that $C = \{1, ..., M\}$. By construction of $\mathcal{E}_\theta^t$, none of the graphs that appear during $\mathcal{T}_\theta$ have links that connect C with the remaining part of the network $C^{\mathrm{c}} = V \setminus C$. Hence, each of the matrices A_s, $s \in \mathcal{T}_\theta$ has the following block diagonal form

$$A_s = \begin{bmatrix} [A_s]_C & 0_{M\times(N-M)} \\ 0_{M\times(N-M)} & [A_s]_{V\setminus C} \end{bmatrix}, \tag{66}$$

and the same structure is therefore preserved in their products $\Phi(t, s) = A_t \cdots A_s$, $s \in \mathcal{T}_\theta$, i.e.,

$$\Phi(t, s) = \begin{bmatrix} [A_t]_C \cdot \ldots \cdot [A_s]_C & 0_{M\times(N-M)} \\ 0_{M\times(N-M)} & [A_t]_{C^{\mathrm{c}}} \cdot \ldots \cdot [A_s]_{C^{\mathrm{c}}} \end{bmatrix}.$$

We next consider part 2. Since for an arbitrary matrix A, for any i, j there holds $|A_{ij}| \leq \|A\|$, it is sufficient to show that $\|[\Phi(t, s)]_C - J_M\| \leq 1/t$, for any fixed $s \in \mathcal{T}_\theta$ such that $s \leq t - o_t$. By part 1, we know that for any $s_1, s_2 \in \mathcal{T}_\theta$, the C block of $\Phi(s_1, s_2)$ is computed as the product of blocks $[A_{s_1}]_C$ through $[A_{s_2}]_C$. Since each of these blocks is a symmetric, stochastic, M by M matrix, we have that $[\Phi(s_1, s_2)]_C$ is a doubly stochastic (M by M) matrix. Consider now a fixed $s \in \mathcal{T}_\theta$ such that $s \leq t - o_t$. Factoring out $[\Phi(t, s)]_C$ as the product $[\Phi(t, t-o_t)]_C\Phi(t-o_t-1, s)]_C$, and using the double-stochasticity of the latter two matrices, we obtain $[\Phi(t, s)]_C - J_M = ([\Phi(t, t-o_t)]_C - J_M)(\Phi(t-o_t-1, s)]_C - J_M)$. By construction of $\mathcal{E}_\theta^t$, the spectral norm of the first factor is not greater than $1/t$, while the double-stochasticity of $\Phi(t - o_t - 1, s)]_C$ yields that the spectral norm of the second factor is not greater than 1. Using submultiplicativity of the spectral norm, the claim in part 2 follows:

$$\begin{aligned} &\|[\Phi(t, s)]_C - J_M\| \\ &\leq \|[\Phi(t, t - o_t)]_C - J_M\| \, \|[\Phi(t - o_t - 1, s)]_C - J_M\| \\ &\leq 1/t. \end{aligned} \tag{67}$$

Part 3 can be proven by factoring out $\Phi(t, s)$ as the product $\Phi(t, \lceil\theta t\rceil)\Phi(\lceil\theta t\rceil - 1, \lceil\theta t\rceil - o_t)\Phi(\lceil\theta t\rceil - o_t - 1, s)$ and applying similar arguments as in the proof of part 2. □

Fix $\theta \in [0, 1]$ and consider the probability distribution $\nu_t^\theta : \mathcal{B}\left(\mathbb{R}^d\right) \to [0, 1]$ defined by

$$\nu_t^\theta(D) = \frac{\mathbb{P}\left(\{X_{i,t} \in D\} \cap \mathcal{E}_\theta^t\right)}{\mathbb{P}\left(\mathcal{E}_\theta^t\right)}, \tag{68}$$

that is, ν_t^θ is the probability distribution of $X_{i,t}$ conditioned on the event $\mathcal{E}_\theta^t$ (we note that $\mathbb{P}\left(\mathcal{E}_\theta^t\right) > 0$ for t sufficiently large, as we show later in the proof, see Lemma 38 further ahead).

Let Υ_t be the (normalized) logarithmic moment generating function associated with ν_t^θ,

$$\Upsilon_t(\lambda) = \frac{1}{t} \log \mathbb{E}\left[e^{t\lambda^\top X_{i,t}} \,\middle|\, \mathcal{E}_\theta^t\right], \quad \text{for } \lambda \in \mathbb{R}^d. \tag{69}$$

Using the properties of entries of $\Phi(t,s)$ for different intervals on s listed in Lemma 36, we establish in Lemma 37 that the sequence of functions Υ_t has a point-wise limit for every $\lambda \in \mathbb{R}^d$. This will allow to apply the Gärtner-Ellis theorem [28] to compute the large deviations rate function for the sequence of measures ν_t^θ. We first state and prove Lemma 37.

Lemma 37. *For any $\lambda \in \mathbb{R}^d$ and any $\theta \in [0,1]$:*

$$\lim_{t\to+\infty} \Upsilon_t(\lambda) = (1-\theta)M\Lambda\left(\frac{1}{M}\lambda\right) + \theta N\Lambda\left(\frac{1}{N}\lambda\right), \tag{70}$$

where, we recall, $M = |C|$.

Proof. Fix $\theta \in [0,1]$, $\lambda \in \mathbb{R}^d$. We have:

$$\begin{aligned}
\mathbb{E}\left[e^{t\lambda^\top X_{i,t}} \left| \mathcal{E}_\theta^t\right.\right] &= \frac{1}{\mathbb{P}\left(\mathcal{E}_\theta^t\right)}\mathbb{E}\left[1_{\mathcal{E}_\theta^t} e^{t\lambda^\top X_{i,t}}\right] \\
&= \frac{1}{\mathbb{P}\left(\mathcal{E}_\theta^t\right)}\mathbb{E}\left[\mathbb{E}\left[1_{\mathcal{E}_\theta^t} e^{t\lambda^\top X_{i,t}} | W_1, ..., W_t\right]\right] \\
&= \frac{1}{\mathbb{P}\left(\mathcal{E}_\theta^t\right)}\mathbb{E}\left[1_{\mathcal{E}_\theta^t}\mathbb{E}\left[e^{t\lambda^\top X_{i,t}} | W_1, ..., W_t\right]\right],
\end{aligned} \tag{71}$$

where in the last equality we used that the indicator $1_{\mathcal{E}_\theta^t}$ is a function of $W_1, ..., W_t$. Further, as the summands in (5) are independent given $W_1, ..., W_t$, we obtain

$$\mathbb{E}\left[e^{t\lambda^\top X_{i,t}} | W_1, ..., W_t\right] = e^{\sum_{s=1}^t \sum_{j=1}^N \Lambda([\Phi(t,s)]_{ij}\lambda)}. \tag{72}$$

Consider now a fixed $\omega \in \mathcal{E}_\theta^t$. We split the sum in the exponent of (72) according to the intervals used in the construction of $\mathcal{E}_\theta^t$. With this in mind, we define also

$$\overline{\chi}_t := \max_{\alpha\in[1/M-1/t,1/M+1/t]} \Lambda(\alpha\lambda), \tag{73}$$

$$\underline{\chi}_t := \min_{\alpha\in[1/M-1/t,1/M+1/t]} \Lambda(\alpha\lambda), \tag{74}$$

and

$$\overline{\zeta}_t := \max_{\alpha\in[1/N-1/t,1/N+1/t]} \Lambda(\alpha\lambda), \tag{75}$$

$$\underline{\zeta}_t := \min_{\alpha\in[1/N-1/t,1/N+1/t]} \Lambda(\alpha\lambda), \tag{76}$$

for $\lambda \in \mathbb{R}^d$. We remark that, by the continuity of Λ and compactness of the intervals, in each of the preceding optimization problems there exists a maximizer. Further, as $t \to +\infty$, the corresponding intervals shrink to a single point, and by using again continuity of Λ, we obtain that $\overline{\chi}_t, \underline{\chi}_t \to \Lambda(1/M\lambda)$, and $\overline{\zeta}_t, \underline{\zeta}_t \to \Lambda(1/N\lambda)$, as $t \to +\infty$. Then, by part 1 of Lemma 36 and the fact that $\Lambda(0) = 0$, we have

$$\sum_{j\notin C} \Lambda([\Phi(t,s)]_{ij}\lambda) = 0, \quad \text{for each } s \in \mathcal{T}_\theta.$$

Further, by part 2 of Lemma 36

$$M\underline{\chi}_t \leq \sum_{j\in C} \Lambda\left([\Phi(t,s)]_{ij}\lambda\right) \leq M\overline{\chi}_t, \text{ for } t-o_t \geq s \geq \lceil \theta t \rceil,$$

and, similarly, by part 3 of Lemma 36

$$N\underline{\zeta}_t \leq \sum_{j=1}^{N} \Lambda\left([\Phi(t,s)]_{ij}\lambda\right) \leq N\overline{\zeta}_t, \text{ for } \lceil \theta t \rceil - o_t \geq s \geq 1.$$

As for the summands in the intervals $\{t, ..., t-o_t\}$ and $\{\lceil \theta t \rceil, ..., \lceil \theta t \rceil - o_t\}$, we apply Lemma 33 to get

$$M\Lambda\left(\frac{1}{M}\lambda\right) \leq \sum_{j\in C} \Lambda\left([\Phi(t,s)]_{ij}\lambda\right) \leq \Lambda(\lambda),$$

$$\text{for } t \geq s \geq t-o_t,$$

and

$$N\Lambda\left(1/N\lambda\right) \leq \sum_{j=1}^{N} \Lambda\left([\Phi(t,s)]_{ij}\lambda\right) \leq \Lambda(\lambda),$$

$$\text{for } \lceil \theta t \rceil \geq s \geq \lceil \theta t \rceil - o_t.$$

Summing out the upper and lower bounds over all s in the preceding five inequalities yields:

$$t\,\underline{\Upsilon}_t\left(\lambda\right) \leq \sum_{s=1}^{t}\sum_{i=1}^{N} \Lambda\left([\Phi(t,s)]_{i,j}\right) \leq t\,\overline{\Upsilon}_t\left(\lambda\right), \tag{77}$$

where

$$\begin{aligned}\underline{\Upsilon}_t\left(\lambda\right) = &\frac{\lceil \theta t \rceil - o_t}{t} N\underline{\zeta}_t + \frac{o_t}{t}\left(N\Lambda\left(\frac{1}{N}\lambda\right) + M\Lambda\left(\frac{1}{M}\lambda\right)\right) \\ &+ \frac{t - \lceil \theta t \rceil - o_t}{t} M\underline{\chi}_t,\end{aligned}$$

and

$$\begin{aligned}\overline{\Upsilon}_t(\lambda) = &\frac{\lceil \theta t \rceil - o_t}{t} N\overline{\zeta}_t + \frac{o_t}{t}\left(N\Lambda\left(\frac{1}{N}\lambda\right) + M\Lambda\left(\frac{1}{M}\lambda\right)\right) \\ &+ \frac{t - \lceil \theta t \rceil - o_t}{t} M\overline{\chi}_t.\end{aligned}$$

The inequalities in (77) hold for any fixed $\omega \in \mathcal{E}_\theta^t$. Thus,

$$1_{\mathcal{E}_\theta^t} e^{t\underline{\Upsilon}_t(\lambda)} \leq 1_{\mathcal{E}_\theta^t}\mathbb{E}\left[e^{t\lambda^\top X_{i,t}}|W_1, ..., W_t\right] \leq 1_{\mathcal{E}_\theta^t} e^{t\overline{\Upsilon}_t(\lambda)}. \tag{78}$$

Finally, by monotonicity of the expectation:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_\theta^t\right) e^{t\underline{\Upsilon}_t(\lambda)} &\leq \mathbb{E}\left[1_{\mathcal{E}_\theta^t}\mathbb{E}\left[e^{t\lambda^\top X_{i,t}}|W_1, ..., W_t\right]\right] \\ &\leq \mathbb{P}\left(\mathcal{E}_\theta^t\right) e^{t\overline{\Upsilon}_t(\lambda)},\end{aligned}$$

which combined with (71) implies

$$e^{t\underline{\Upsilon}_t(\lambda)} \leq \mathbb{E}\left[e^{t\lambda^\top X_{i,t}}|\mathcal{E}_\theta^t\right] \leq e^{t\overline{\Upsilon}_t(\lambda)}. \tag{79}$$

Now, taking the logarithm and dividing by t,

$$\underline{\Upsilon}_t(\lambda) \leq \Upsilon_t(\lambda) \leq \overline{\Upsilon}_t(\lambda),$$

and noting that

$$\begin{aligned}\lim_{t\to+\infty} \overline{\Upsilon}_t(\lambda) &= \lim_{t\to+\infty} \underline{\Upsilon}_t(\lambda) \\ &= (1-\theta)M\Lambda\left(\frac{1}{M}\lambda\right) + \theta N\Lambda\left(\frac{1}{N}\lambda\right),\end{aligned}$$

the claim of Lemma 37 follows. □

By the Gärtner-Ellis theorem it follows then that the sequence of measures ν_t^θ satisfies the large deviations principle[5], with the rate function equal to the conjugate of

$$f_\theta(\lambda) := (1-\theta)M\Lambda\left(\frac{1}{M}\lambda\right) + \theta N\Lambda\left(\frac{1}{N}\lambda\right), \tag{80}$$

for $\lambda \in \mathbb{R}^d$. Therefore, for every open set $E \subseteq \mathbb{R}^d$, there holds

$$\liminf_{t\to+\infty} \frac{1}{t}\log \mathbb{P}\left(X_{i,t} \in E|\mathcal{E}_\theta^t\right) \geq -\inf_{x\in E}\left\{\sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - f_\theta(\lambda)\right\}. \tag{81}$$

We next turn to computing the probability of the event $\mathcal{E}_\theta^t$.

Lemma 38. *For any $\theta \in [0,1]$, for all t sufficiently large:*

$$\frac{1}{4}p_{\mathcal{H}}^{t-\lceil\theta t\rceil} \leq \mathbb{P}\left(\mathcal{E}_\theta^t\right) \leq p_{\mathcal{H}}^{t-\lceil\theta t\rceil}. \tag{82}$$

Proof. By the disjoint blocks theorem [34] applied to the matrices in $\mathcal{T}_\theta$ and its complement $\{1,...,t\}\setminus\mathcal{T}_\theta$, we obtain

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_\theta^t\right) = \mathbb{P}&\left(\|\Phi(\lceil\theta t\rceil, \lceil\theta t\rceil - o_t) - J_N\| \leq \frac{1}{t}\right) \times \\ \mathbb{P}&\left(G_s \in \mathcal{H},\ \text{for } s \in \mathcal{T}_\theta,\ \|[\Phi(t, t-o_t)]_C - J_M\| \leq \frac{1}{t}\right).\end{aligned} \tag{83}$$

[5]We use here the variant of the Gärtner-Ellis theorem which claims the (full) LDP for the case when the domain of the limiting function is the whole space $\mathbb{R}^d$, as given in [28]; see also Exercise 2.3.20 in [28] for the statement and the sketch of the proof of this result.

We show using (17) that the first term in the right-hand side of the preceding equality goes to 1 as $t \to +\infty$. Fix an arbitrary $\epsilon \in (0,1)$. Then, for all t sufficiently large,

$$\begin{aligned}&\mathbb{P}\left(\|\Phi(\lceil\theta t\rceil, \lceil\theta t\rceil - o_t) - J_N\| \leq \frac{1}{t}\right)\\ &\geq 1 - K_\epsilon e^{-t(\mathcal{J}-\epsilon)} \geq 1/2.\end{aligned} \tag{84}$$

Clearly, being a probability, this term is also smaller than 1 (for all t). Consider now the second factor in the right-hand side of (83). Conditioning on the event $\{G_s \in \mathcal{H}, \text{ for } s \in \mathcal{T}_\theta\}$, and using the fact that the probability of this event equals $p_{\mathcal{H}}^{t-\lceil\theta t\rceil}$ (note that the latter holds by the independence of weight matrices, Assumption 1.2), we obtain

$$\begin{aligned}&\mathbb{P}\left(G_s \in \mathcal{H}, \text{ for } s \in \mathcal{T}_\theta, \ \|[\Phi(t, t-o_t)]_C - J_M\| \leq \frac{1}{t}\right) =\\ &\mathbb{P}\left(\|[\Phi(t, t-c_t)]_C - J_M\| \leq \frac{1}{t} | G_s \in \mathcal{H}, \text{ for } s \in \mathcal{T}_\theta\right) p_{\mathcal{H}}^{t-\lceil\theta t\rceil}.\end{aligned}$$

Similarly as in (84), it can be shown that the conditional probability term in (85), for all t sufficiently large, greater than $1/2$. On the other hand, it is obviously smaller than 1 for all t. Summarizing the preceding findings, the claim of the lemma follows. □

To bring the two key arguments together – Lemma 38 and the lower bound (81), we start from the following simple bound

$$\begin{aligned}\mathbb{P}\left(X_{i,t} \in E\right) &\geq \mathbb{P}\left(\{X_{i,t} \in E\} \cap \mathcal{E}_\theta^t\right)\\ &= \nu_t^\theta(E)\mathbb{P}\left(\mathcal{E}_\theta^t\right).\end{aligned} \tag{85}$$

From superadditivity of the $\liminf$, followed by an application of (81) and (82), we obtain

$$\begin{aligned}&\liminf_{t\to+\infty} \frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in E\right)\\ &\geq \liminf_{t\to+\infty} \frac{1}{t}\log\nu_t^\theta(E) + \lim_{t\to+\infty}\frac{1}{t}\log\mathbb{P}\left(\mathcal{E}_\theta^t\right)\\ &\geq -\inf_{x\in E}\left\{\sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - f_\theta(\lambda)\right\} - (1-\theta)|\log p_{\mathcal{H}}|.\end{aligned}$$

The preceding inequality holds for each θ in $[0,1]$. Optimizing over all such values yields:

$$\begin{aligned}&\liminf_{t\to+\infty} \frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in E\right) \geq\\ &-\inf_{\theta\in[0,1]}\left\{\inf_{x\in E}\sup_{\lambda\in\mathbb{R}^d}\left\{\lambda^\top x - f_\theta(\lambda)\right\} + (1-\theta)|\log p_{\mathcal{H}}|\right\}\\ &= -\inf_{x\in E}\inf_{\theta\in[0,1]}\left\{\sup_{\lambda\in\mathbb{R}^d}\left\{\lambda^\top x - f_\theta(\lambda)\right\} + (1-\theta)|\log p_{\mathcal{H}}|\right\}.\end{aligned}$$

Now, fix $x \in E$ and consider the function

$$g(\theta, \lambda) := \lambda^\top x - (1-\theta)\left(M\Lambda\left(\frac{1}{M}\lambda\right) - |\log p_{\mathcal{H}}|\right) - \theta N\Lambda\left(\frac{1}{N}\lambda\right). \tag{86}$$

As an affine function of θ, g is convex in θ. Further, by convexity of Λ, g is concave in λ, for any $\theta \in [0,1]$. Finally, sets $[0,1]$ and $\mathbb{R}^d$ are convex and set $[0,1]$ is compact. Thus, conditions for applying the Minimax theorem [36] are fulfilled and we obtain:

$$\begin{aligned}
&\inf_{\theta\in[0,1]} \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - (1-\theta)\left(M\Lambda\left(1/M\lambda\right) - |\log p_{\mathcal{H}}|\right) \\
&\qquad - \theta N\Lambda\left(1/N\lambda\right) = \\
&\sup_{\lambda\in\mathbb{R}^d} \inf_{\theta\in[0,1]} \lambda^\top x - (1-\theta)\left(M\Lambda\left(1/M\lambda\right) - |\log p_{\mathcal{H}}|\right) \\
&\qquad - \theta N\Lambda\left(1/N\lambda\right) \\
&= \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - \max\left\{M\Lambda\left(1/M\lambda\right) - |\log p_{\mathcal{H}}|, \Lambda\left(1/N\lambda\right)\right\}.
\end{aligned}$$

Similarly as in the proof of the upper bound, using the conjugacy rules from Lemma 34,

$$\begin{aligned}
&\sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - \min\left\{M\Lambda\left(\frac{1}{M}\lambda\right) - |\log p_{\mathcal{H}}|, \Lambda\left(\frac{1}{N}\lambda\right)\right\} \\
&= \overline{\mathrm{co}}\inf\left(NI, MI + |\log p_{\mathcal{H}}|\right)(x),
\end{aligned}$$

which finally yields,

$$\begin{aligned}
&\liminf_{t\to+\infty} \frac{1}{t}\log \mathbb{P}\left(X_{i,t} \in E\right) \\
&\geq -\inf_{x\in E} \overline{\mathrm{co}}\inf\left\{NI, MI + |\log p_{\mathcal{H}}|\right\}(x).
\end{aligned}$$

This completes the proof of the lower bound and the proof of Theorem 13.

VII. Conclusion

We studied large deviations inaccuracy rates for consensus+innovations based distributed inference for generic random networks. We assume vector measurements with possibly non-i.i.d. entries. Our goal was to find bounds or exact rate function for each node in the network, accounting for the specificities of the node's interactions. For each node, we found a node-specific family of lower bounds, induced by the family of network subgraphs in which the node participates. Specifically, each bound in the family is given as the convex envelope of the centralized rate function and the effective rate function corresponding to a given subgraph, and lifted by the probability that this subgraph remains isolated from the remainder

of the network. The upper bound is defined as the convex envelope of the centralized rate function and the rate function corresponding to an isolated node, lifted by the rate of consensus. We show that, for certain cases such as pendant nodes and d-cyclic graphs, the two bounds match, hence proving the large deviations principle for these classes of random networks. We illustrate the results with an application to social learning, providing also the first proof of the large deviations principle for social learning beliefs with random network models.

APPENDIX A
PROOF OF (20)

Fix $i \in V$ and suppose that the inequalities in (18) and (19) hold for any set D. Suppose also that the sequence of node i's states, $X_{i,t}$, satisfies the LDP with rate function I_i.

We prove (20) by contradiction. Consider first the right hand side of (20) and suppose, for the sake of contradiction, that there exists a point x_0 such that $I_i(x_0) > \overline{I}_i(x_0)$. Let $\epsilon = I_i(x_0) - \overline{I}_i(x_0)$ and introduce $S = \left\{x \in \mathbb{R}^d : I_i(x) > \overline{I}_i(x_0) + \epsilon/2\right\}$. By the lower semi-continuity of I_i, S is open. Also, $x_0 \in S$. Thus, for $\delta > 0$ sufficiently small, the closed ball $\overline{B}_{x_0}(\delta)$ entirely belongs to S. Combining the LDP upper bound (1) for $D = \overline{B}_{x_0}(\delta)$, with the bound (18) for $D = B_{x_0}(\delta)$, we obtain:

$$-\inf_{x \in B_{x_0}(\delta)} \overline{I}_i(x) \leq \liminf_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in B_{x_0}(\delta)\right) \tag{87}$$

$$\leq \limsup_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in \overline{B}_{x_0}(\delta)\right) \leq -\inf_{x \in \overline{B}_{x_0}(\delta)} I_i(x). \tag{88}$$

Since $\inf_{x \in B_{x_0}(\delta)} \overline{I}_i(x) \leq \overline{I}_i(x_0)$, we have that the left hand side in (87) is greater than $-\overline{I}_i(x_0)$. On the other hand, for any $x \in \overline{B}_{x_0}(\delta)$, $I_i(x) > \overline{I}_i(x_0) + \epsilon/2$, implying $\inf_{x \in \overline{B}_{x_0}(\delta)} I_i(x) \geq \overline{I}_i(x_0) + \epsilon/2$. This finally yields contradiction since the right hand side in (87) cannot be smaller than $-\overline{I}_i(x_0)$.

APPENDIX B
PROOF OF LEMMA 20

We start by noting that $\operatorname{epi} \inf\{NI, I + \mathcal{J}\} = S_1 \cup S_2$, where S_1 and S_2 are the epigraphs of NI and $I + \mathcal{J}$, $S_1 = \operatorname{epi}(NI)$ and $S_2 = \operatorname{epi}(I + \mathcal{J})$. To prove Lemma 20, we need to show that $\operatorname{epi} F = \overline{\operatorname{co}}\{S_1 \cup S_2\}$, where F is the function defined in the right hand side of eq. (38). To do this it suffices to show that: 1) $\operatorname{epi} F$ is a convex set, and 2) $\operatorname{epi} F \subseteq \operatorname{co}(S_1 \cup S_2)$. We first prove 1). It suffices to show that F is convex, which we do using generalized second order characterizations of convex functions, e.g. [37]. Note that F is continuous and that $\mathcal{D}_F = \mathbb{R}^d$. For each x and d, let $F'_+(x,d)$ and $F''_+(x,d)$

denote, respectively, the upper directional derivatives of the first and the second order at the point x and in the direction d,

$$F'_+(x;d) = \limsup_{\epsilon \downarrow 0} \frac{F(x+\epsilon d) - F(x)}{\epsilon} \tag{89}$$

$$F''_+(x;d) = \limsup_{\epsilon \downarrow 0} \frac{F(x+\epsilon d) - F(x) - F'_+(x;d)}{2\,\epsilon^2}. \tag{90}$$

We will show that F is in fact differentiable. Then, by Theorem 2.1. part (i) from [37], proving convexity of F would reduce to proving that $F''_+(x;d) \geq 0$ for any x and d. Note that I and H are differentiable, with their respective gradients given by $\nabla I(x) = S^{-1}(x-m)$ and $\nabla H(x) = S^{-1}(x-m)/\sqrt{(x-m)^\top S^{-1} x - m}$. Thus, F is differentiable in each of the three open sets (note that I is continuous and differentiable): $\{x : I(x) < c_1\}$, $\{x : c_1 < I(x) < Nc_1\}$, and $\{x : NI(x) > c_1\}$. It remains to show that F is differentiable for those x such that $I(x) = c_1$ and $I(x) = Nc_1$. Fix first x such that $I(x) = c_1$. It is easy to see that, for any d such that $d^\top S^{-1}(x-m) \geq 0$, $I(x+\epsilon d) > I(x)$ for all $\epsilon > 0$. Also, for any d such that $d^\top S^{-1}(x-m) < 0$, $I(x+\epsilon d) < I(x)$ for all sufficiently small $\epsilon > 0$. Thus, if $d^\top S^{-1}(x-m) \geq 0$, $F(x+\epsilon d) = N\sqrt{2c_1}H(x+\epsilon d) - Nc_1$, for all ϵ sufficiently small, and hence $F'_+(x;d) = N\sqrt{2c_1}d^\top \nabla H(x)$. Using now the fact that $I(x) = c_1$, we obtain that $F'_+(x;d) = Nd^\top S^{-1}(x-m)$. Consider now the case when d is such that $d^\top S^{-1}(x-m) \leq 0$. Then, by the discussion above we have that for all ϵ, $F(x+\epsilon d) = NI(x+\epsilon d)$. Hence, $F'_+(x;d) = Nd^\top \nabla I(x) = Nd^\top S^{-1}(x-m)$. Since for any x s.t. $I(x) = c_1$ and for any d we have that $F'_+(x;d) = Nd^\top \nabla I(x)$, we conclude that F is differentiable at any such x. We can in analogous manner prove differentiability of F at any x s.t. $I(x) = Nc_1$. Hence, we conclude that F is differentiable.

We now turn to proving that $F''_+(x;d) \geq 0$ for any x and d. Note that $\nabla^2 I(x) = S^{-1} \succeq 0$ and

$$\nabla^2 H(x) = \\ N\frac{\sqrt{2c_1}}{\sqrt{2I(x)}}\left(S^{-1} - \frac{1}{2I(x)}S^{-1}(x-m)(x-m)^\top S^{-1}\right),$$

for any x. To see that $\nabla^2 H(x) \succeq 0$, it suffices to observe that it can be rewritten as $\nabla^2 H(x) = N\sqrt{2c_1}/\sqrt{2I(x)}S^{-1/2}(I - qq^\top/(\|q\|^2))S^{-1/2}$, for $q = S^{-1/2}(x-m)$. Since the matrix inside the brackets is positive semidefinite, positive semidefiniteness of $\nabla^2 H(x)$ follows. Therefore, for any x in the interior of the three sets in (38), we have that $F''_+(x;d) \geq 0$. Consider now the case when x satisfies $I(x) = c_1$. Following the same steps as in the preceding paragraph, we obtain that for any d s.t. $d^\top S^{-1}(x-m) \geq 0$, $F''_+(x;d) = N^2 2c_1 d^\top \nabla^2 H(x) d \geq 0$ and for d s.t. $d^\top S^{-1}(x-m) \leq 0$, $F''_+(x;d) = Nd^\top \nabla^2 I(x) d \geq 0$. To complete the proof of 1), it only remains to consider those x that satisfy $I(x) = Nc_1$. Analogously to the preceding case, we get that for d s.t. $d^\top S^{-1}(x-m) \geq 0$, $F''_+(x;d) = d^\top \nabla^2 I(x) d \geq 0$ and for

d s.t. $d^\top S^{-1}(x-m) \leq 0$, $F''_+(x;d) = n^2 2c_1 d^\top \nabla^2 H(x) d \geq 0$. Hence, since F is differentiable and $F''_+(x;d) \geq 0$ for any x and d, we conclude that F is convex.

To prove Lemma 20, it remains to prove part 2). For each unit norm $v \in \mathbb{R}^d$, $\|v\| = 1$, let $\phi_v : \mathbb{R}^d \mapsto \mathbb{R}^d$ denote the projection of F along the direction v, started at point m: $\phi_v(\rho) := F(m + \rho v)$, $\rho \in \mathbb{R}$. Then, $\text{epi}F = \cup_{v \in \mathbb{R}^d, \|v\|=1} \text{epi}\phi_v$. For each fixed v, let $[S_l]_v$ denote the projection of S_l along the line $m + \rho v$, $[S_l]_v = S_l \cap \{m + \rho v : \rho \in \mathbb{R}\}$, $l = 1, 2$. Note that $[S_1]_v = \left\{(t, m + \rho v) : t \geq N\rho^2 v^\top S^{-1} v/2,\, \rho \in \mathbb{R}\right\}$, $[S_2]_v = \left\{(t, m + \rho v) : t \geq \rho^2 v^\top S^{-1} v/2 + \mathcal{J},\, \rho \in \mathbb{R}\right\}$. Then, it is easy to see that, for each unit norm v, $\text{epi}\phi_v = \text{co}\left([S_1]_v \cup [S_2]_v\right)$. Finally, since $\text{co}\left([S_1]_v \cup [S_2]_v\right) \subseteq \text{co}\left(S_1 \cup S_2\right)$, the claim in 2) follows. This completes the proof of Lemma 20.

APPENDIX C
PROOF OF LEMMA 28

Fix an arbitrary node $i \in V$. For each $m = 1, ..., M-1$, $X^m_{i,t} = \frac{1}{t} \log \frac{b^m_{i,t}}{b^M_{i,t}}$, hence

$$\frac{1}{t} \log b^m_{i,t} = X^m_{i,t} + \frac{1}{t} \log b^M_{i,t}. \tag{91}$$

Further, the (private) beliefs by construction sum up to one: $\sum_{m=1}^M b^m_{i,t} = 1$. Dividing both sides by $b^M_{i,t}$ and exploiting the functional relation between $b^m_{i,t}$ and $X^m_{i,t}$, we obtain

$$\sum_{m=1}^{M-1} e^{tX^m_{i,t}} + 1 = \frac{1}{b^M_{i,t}}. \tag{92}$$

It follows that:

$$\frac{1}{M} e^{-t \max_{m=1,\ldots,M} X^m_{i,t}} \leq b^M_{i,t} \leq e^{-t \max_{m=1,\ldots,M} X^m_{i,t}}, \tag{93}$$

where $X^M_{i,t} \equiv 0$. From (91) and (93) we obtain

$$g_m(X_{i,t}) - \frac{1}{t} \log M \leq \frac{1}{t} \log b^m_{i,t} \leq g_m(X_{i,t}). \tag{94}$$

Consider now an arbitrary one-sided closed interval F on $\mathbb{R}$. Suppose that $F = [a, +\infty)$ (other intervals in $\mathbb{R}$ can be treated analogously). Fix $\epsilon > 0$. From (94), for all $t \geq t_0 = \log M/\epsilon$ there holds:

$$g_m(X_{i,t}) - \epsilon \leq \frac{1}{t} \log b^m_{i,t} \leq g_m(X_{i,t}), \tag{95}$$

and thus, for all $t \geq t_0$

$$\mathbb{P}(\frac{1}{t} \log b^m_{i,t} \geq a + \epsilon) \leq \mathbb{P}(g_m(X_{i,t}) \geq a) = P(X_{i,t} \in g_m^{-1}([a, +\infty)). \tag{96}$$

Taking the $\limsup$ over $t \to +\infty$,

$$\limsup_{t \to +\infty} \frac{1}{t} \log \mathbb{P}(\frac{1}{t} \log b^m_{i,t} \geq a + \epsilon) \leq \limsup_{t \to +\infty} \frac{1}{t} \log P(X_{i,t} \in g_m^{-1}([a, +\infty)). \tag{97}$$

The above inequality holds for all $\epsilon > 0$. Taking the supremum over $\epsilon > 0$ on the left hand side yields:

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}(\frac{1}{t} \log b_{i,t}^m \geq a) \leq \limsup_{t\to+\infty} \frac{1}{t} \log P(X_{i,t} \in g_m^{-1}([a, +\infty)). \tag{98}$$

Applying now the upper bound in 44, the upper bound in (49) follows. The proof of the lower bound (49) is analogous.

APPENDIX D
PROOF OF LEMMA 35

Suppose that $X_t \in \mathbb{R}^d$ is a sequence of random variables for which (62) holds for some function f. Fix a compact set $F \subseteq \mathbb{R}^d$. For each $\delta > 0$, introduce the function $f^{\star,\delta} : \mathbb{R}^d \mapsto \mathbb{R}$ obtained by truncating $f^\star$ to $1/\delta$:

$$f^{\star,\delta}(x) = \inf \left\{ \frac{1}{\delta}, f^\star(x) - \delta \right\}, \text{ for } x \in \mathbb{R}^d. \tag{99}$$

The family of functions $f^{\star,\delta}$, $\delta > 0$, satisfies that, for any set D,

$$\lim_{\delta\to 0} \inf_{x\in D} f^{\star,\delta}(x) = \inf_{x\in D} f^\star(x). \tag{100}$$

To show this, let $\xi := \inf_{x\in D} f^\star(x)$ and suppose first that $\xi = +\infty$, i.e., $f^\star$ at all points $x \in D$ takes the value $+\infty$. Then, for any $\delta > 0$, $f^{\star,\delta} = 1/\delta$ for all $x \in D$, and therefore, for any $\delta > 0$, $\inf_{x\in D} f^{\star,\delta}(x) = 1/\delta$. Computing the limit $\lim_{\delta\to 0} 1/\delta = +\infty$, identity (100) follows. We next consider the case $\xi \in \mathbb{R}$. For arbitrary fixed $\delta > 0$, the quantity under the limit in the left hand side of (100) equals:

$$\begin{aligned} \inf_{x\in D} f^{\star,\delta}(x) &= \inf_{x\in D} \inf \left\{ f^\star(x) - \delta, \frac{1}{\delta} \right\} \\ &= \inf \left\{ \inf_{x\in D} \left(f^\star(x) - \delta \right), \frac{1}{\delta} \right\}. \end{aligned} \tag{101}$$

The first argument of the infimum (101) equals $\xi - \delta$ and it is finite by our assumption. Hence, for all δ sufficiently small, the infimum (101) equals $\xi - \delta$, which after taking the limit $\delta \to 0$ yields the claim. The case $\xi = -\infty$ can be proven equivalently.

Having (100), it easy to see that (63) follows if we show that the following inequality holds for any given δ:

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_t \in F\right) \leq 2\delta - \inf_{x\in F} f^{\star,\delta}(x). \tag{102}$$

Thus, in what follows we focus on proving (102). To this end, fix $\delta > 0$. For any point $y \in F$ there exists a point λ_y (which depends on δ) such that

$$\lambda_y^\top y - \Lambda^\star(\lambda_y) \geq f^{\star,\delta}(y). \tag{103}$$

Existence of such a point follows directly from the definitions of $f^\star$ and $f^{\star,\delta}$. First, since for any fixed point y $f^\star(y)$ is computed as the supremum of functions $\lambda \mapsto h_y(\lambda) := \lambda^\top y - f(\lambda)$, it follows that the value $f^\star(y)$ can be approached arbitrarily close with $h_y(\lambda)$. Second, since $f^{\star,\delta}(y)$ is the infimum of $f^\star(y) - \delta$ and $1/\delta$, it must satisfy $f^\star(y) - \delta,\, 1/\delta \geq f^{\star,\delta}(y)$. For example, if, for some y, $f^\star(y)$ is finite, then there must exist a point λ such that $h_y(\lambda) \geq f^\star(y) - \delta$, and since the latter is greater than $f^{\star,\delta}(y)$, (103) follows.

Note now that (62) implies that, for any measurable set D, there exists $t_0 = t_0(\delta, D)$ such that

$$\frac{1}{t}\log \mathbb{P}\left(X_t \in D\right) \leq \delta + f(\lambda) - \inf_{x\in D} \lambda^\top x, \tag{104}$$

for all $t \geq t_0$. For any $y \in F$, let $r_y := \delta/\|\lambda_y\|$. Taking $D = \overline{B}_y(r_y)$ and $\lambda = \lambda_y$ in (104) yields for any $t \geq t_0(\delta, y)$

$$\begin{aligned}
\frac{1}{t}\log \mathbb{P}\left(X_t \in \overline{B}_y(r_y)\right) &\leq \delta + f(\lambda_y) - \inf_{\|x-y\|\leq r_y} \lambda_y^\top x && (105)\\
&\leq \delta + \Lambda^\star(\lambda_y) - \lambda_y^\top y - \inf_{\|x\|\leq r_y} \lambda_y^\top x && (106)\\
&\leq 2\delta - f^{\star,\delta}(y), && (107)
\end{aligned}$$

where the last inequality follows from (103) and the definition of r_y. Next, from the family of closed balls $\left\{\overline{B}_y(r_y) \,:\, y \in F\right\}$, a finite cover of F, $\left\{\overline{B}_{y_k}(r_{y_k}) \,:\, k = 1, ..., K\right\}$, is extracted, where, we note, $K = K(F, \delta)$. Then, by the union bound,

$$\begin{aligned}
\frac{1}{t}\log\, \mathbb{P}\left(X_t \in F\right) &\leq \frac{1}{t}\log\left(\sum_{k=1}^{K} \mathbb{P}\left(X_t \in \overline{B}_{y_k}(r_{y_k})\right)\right)\\
&\leq \frac{1}{t}\log K + \frac{1}{t}\log \max_{k=1,...,K} \mathbb{P}\left(X_t \in \overline{B}_{y_k}(r_{y_k})\right)\\
&\leq \frac{1}{t}\log K + \max_{k=1,...,K} \frac{1}{t}\log \mathbb{P}\left(X_t \in \overline{B}_{y_k}(r_{y_k})\right).
\end{aligned}$$

Combining the preceding inequality with (105) applied for every $k = 1, ..., K$, we have that for every $t \geq \max_{k=1,...K} t_0(\delta, y_k)$

$$\begin{aligned}
\frac{1}{t}\log\, \mathbb{P}\left(X_t \in F\right) &\leq \frac{1}{t}\log K + \max_{k=1,...,K} 2\delta - f^{\star,\delta}(y)\\
&\leq \frac{1}{t}\log K + 2\delta - \inf_{y\in F} f^{\star,\delta}(y). && (108)
\end{aligned}$$

Taking the limit $t \to +\infty$, and noting that K is finite, (102) follows. The last part of the claim, i.e., (102) for closed sets follows from (102) for compact sets, that we have just proved, and Lemma 1.2.18 in [28].

REFERENCES

[1] D. Bajović, "Large deviations rates for distributed inference," Ph.D. dissertation, Carnegie Mellon University, 2013.

[2] H. Chernoff, "A measure of the asymptotic efficiency of tests of a hypothesis based on a sum of observations," *The Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, Dec. 1952.

[3] A. Anandkumar and L. Tong, "Type-based random access for distributed detection over multiaccess fading channels," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 5032–5043, 2007.

[4] D. Bajović, B. Sinopoli, and J. Xavier, "Sensor selection for event detection in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4938–4953, 2011.

[5] W. P. Tay, "Whose opinion to follow in multihypothesis social learning? A large deviations perspective," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 2, pp. 344–359, 2015.

[6] P. Hu, V. Bordignon, S. Vlaski, and A. H. Sayed, "Optimal aggregation strategies for social learning over graphs," 2022. [Online]. Available: https://arxiv.org/abs/2203.07065

[7] J. A. Bucklew, *Large Deviations Techniques in Decision, Simulation and Estimation*. New York: Wiley, 1990.

[8] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications, and Computing*. New York: Chapman and Hall, 1995.

[9] H. Touchette, "The large deviation approach to statistical mechanics," *Physics Reports*, vol. 478, no. 1, pp. 1 – 69, 2009.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[11] M. Arcones, "Large deviations for M-estimators," *Annals of the Institute of Statistical Mathematics*, vol. 58, no. 1, pp. 21–52, 2006.

[12] R. R. Bahadur, "On the asymptotic efficiency of tests and estimates," *Sankhya: The Indian Journal of Statistics, 1933-1960*, vol. 22, no. 3/4, pp. 229–252, 1960. [Online]. Available: http://www.jstor.org/stable/25048458

[13] D. Bajović, D. Jakovetić, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4381–4396, Sep. 2011.

[14] D. Bajović, D. Jakovetić, 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, Nov. 2012.

[15] D. Jakovetić, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: Large deviations analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4306–4320, 2012.

[16] D. Bajović, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Distributed inference over directed networks: Performance limits and optimal design," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3308–3323, July 2016.

[17] V. Matta, P. Braca, S. Marano, and A. H. Sayed, "Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime," *IEEE Trans. Information Theory*, vol. 62, no. 8, pp. 4710–4732, August 2016.

[18] ——, "Distributed detection over adaptive networks: Refined asymptotics and the role of connectivity," *IEEE Trans. Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 442–460, Dec 2016.

[19] S. Marano and A. H. Sayed, "Detection under one-bit messaging over adaptive networks," *IEEE Trans. Information Theory*, vol. 65, no. 10, pp. 6519–6538, October 2019.

[20] 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: http://www.sciencedirect.com/science/article/pii/S0899825612000851

[21] 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, 2016.

[22] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161–6179, 2018.

[23] A. Mitra, J. A. Richards, and S. Sundaram, "A new approach to distributed hypothesis testing and non-bayesian learning: Improved learning rate and byzantine resilience," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4084–4100, 2021.

[24] M. H. DeGroot, "Reaching a consensus," *Journal of American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[25] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-Bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.

[26] D. Bajović, J. Xavier, J. M. F. Moura, and B. Sinopoli, "Consensus and products of random stochastic matrices: Exact rate for convergence in probability," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2557–2571, May 2013.

[27] R. Parasnis, M. Franceschetti, and B. Touri, "Non-Bayesian social learning on random digraphs with aperiodically varying network connectivity," 2020. [Online]. Available: https://arxiv.org/abs/2010.06695

[28] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Boston, MA: Jones and Barlett, 1993.

[29] D. Li, S. Kar, J. M. F. Moura, H. V. Poor, and S. Cui, "Distributed Kalman filtering over massive data sets: Analysis through large deviations of random Riccati equations," *IEEE Transactions on Information Theory*, vol. 61, no. 3, pp. 1351–1372, March 2015.

[30] F. den Hollander, *Large Deviations*. Fields Institute Monographs, American Mathematical Society, 2000.

[31] V. Vysotsky, "When is the rate function of a random vector strictly convex?" *Electronic Communications in Probability*, vol. 26, no. none, pp. 1 – 11, 2021. [Online]. Available: https://doi.org/10.1214/21-ECP409

[32] J.-B. Hiriart-Urruty and C. Lemarechal, *Fundamentals of Convex Analysis*, ser. Grundlehren Text Editions. Berlin, Germany: Springer-Verlag, 2004.

[33] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 2015. [Online]. Available: https://doi.org/10.1515/9781400873173

[34] A. F. Karr, *Probability*, ser. Springer Texts in Statistics. New York: Springer-Verlag, 1993.

[35] D. Bajović, D. Jakovetić, J. M. F. Moura, J. Xavier, , and B. Sinopoli, "Large deviations analysis of consensus+innovations detection in random networks," in *Allerton'11, 49th Allerton Conference on Communication, Control, and Computing*, Monticello, Il, October 2011.

[36] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176, March 1958.

[37] I. Ginchev and V. I. Ivanov, "Second-order characterizations of convex and pseudoconvex functions," *Journal of Applied Analysis*, vol. 9, no. 2, pp. 261–273, June 2010.