Distributed Inference Over Directed Networks: Performance Limits and Optimal Design

Dragana Bajović, *Member, IEEE*, José M. F. Moura, *Fellow, IEEE*, João Xavier, *Member, IEEE*, and Bruno Sinopoli, *Member, IEEE*

***Abstract*—We find large deviations rates for consensus-based distributed inference for directed networks. When the topology is deterministic, we establish the large deviations principle and find exactly the corresponding rate function, equal at all nodes. We show that the dependence of the rate function on the stochastic weight matrix associated with the network is fully captured by its left eigenvector corresponding to the unit eigenvalue. Further, when the sensors' observations are Gaussian, the rate function admits a closed-form expression. Motivated by these observations, we formulate the optimal network design problem of finding the left eigenvector that achieves the highest value of the rate function, for a given target accuracy. This eigenvector therefore minimizes the time that the inference algorithm needs to reach the desired accuracy. For Gaussian observations, we show that the network design problem can be formulated as a semidefinite (convex) program, and hence can be solved efficiently. When observations are identically distributed across agents, the system exhibits an interesting property: the graph of the rate function always lies between the graphs of the rate function of an isolated node and the rate function of a fusion center that has access to all observations. We prove that this fundamental property holds even when the topology and the associated system matrices change randomly over time, with arbitrary distribution. Due to the generality of its assumptions, the latter result requires more subtle techniques than the standard large deviations tools, contributing to the general theory of large deviations.**

***Index Terms*—Distributed inference, large deviations analysis, rate function, large deviations principle, directed topologies, random networks, time-correlated networks, consensus algorithms.**

I. Introduction

THE field of wireless sensor networks (WSN) has significantly evolved since its beginnings about two decades ago. Starting from wildlife monitoring, smart housing, and building and infrastructure surveillance [1], the applications of WSNs have grown both in diversity and in scale. They now include monitoring and control of some highly complex large scale systems, such as vehicular networks and electric power grids. One important emerging trend in this field is networks consisting of thousands of very small and simple sensing devices, such as microrobots [2] and nano-networks [3].

Due to the increased complexity and scale of WSNs, there has been significant interest recently in algorithms that process network information using local communications only [4]–[6]. A representative of this class of algorithms is the consensus algorithm [7]–[9]. With consensus algorithms, each agent maintains over iterations an estimate of the quantity of interest and over time it communicates the estimate to its immediate neighbors. In addition, intertwined with local communications are the local agents' innovations, where agents collect new measurements and incorporate them in an iterative fashion in their current estimates. Algorithms of this form, referred to as consensus+innovations [10], [13], possess several desirable features, including scalability and simplicity of implementation. Further, they are robust to structural changes in the system, such as node failures and intermittent communications, which are typical for complex systems consisting of many structurally simple devices. In terms of applications, consensus algorithms have been applied in various different contexts: distributed Kalman filtering [11], [12], distributed detection [9], [13], [8], [14] and parameter estimation [7], [15], [10], distributed learning [16], and tracking [17].

In this paper, we study large deviations performance of consensus algorithms when the underlying network is directed. This complements the existing work that usually studies asymptotic variance or asymptotic normality [10], [18]. Our goal is to compute (or characterize—when exact computation is not possible) the rates at which the local nodes' estimates converge to the desirable values (e.g., the vector of true parameters that are being estimated). To explain the relevance of large deviations performance, consider, for example, a binary hypothesis testing problem in a WSN. In this context, the rates of large deviations correspond to error exponents, i.e., they provide answers to how fast the error probabilities—false alarm, missed detection, or total error probability decay with time. In the context of estimation, large deviation rates provide estimates of times to reach a desired accuracy region around the true parameter that the local estimates converge to. Naturally, the higher the rate of a node, the better is the decision or estimation produced by that node at a given time. One particular goal of this paper is to provide answers to questions such as: "How much faster a node in a

Manuscript received April 28, 2015; revised October 06, 2015 and January 24, 2016; accepted February 21, 2016. Date of publication March 17, 2016; date of current version May 18, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dmitry Malioutov. The work of J. M. F. Moura is supported by NSF grants CCF1011903 and CCF1513936.

D. Bajović is with the BioSense Institute, Novi Sad, Serbia, and the Department of Power, Electronic and Communication Engineering, University of Novi Sad, 21000 Novi Sad, Serbia (e-mail: dbajovic@uns.ac.rs).

J. M. F. Moura and B. Sinopoli are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213-3890 USA (e-mail: moura@ece.cmu.edu; brunos@ece.cmu.edu).

J. Xavier is with the Institute for Systems and Robotics, Instituto Superior Técnico, University of Lisbon, Lisbon 1600-011, Portugal (e-mail: jxavier@isr.ist.utl.pt).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2543209

network filters out the estimation noise compared to a node that operates alone?"

Contributions: We consider both cases when the local nodes' interactions are deterministic and when they are random, where the local interactions are captured by associated stochastic system matrices W_t[1]. For the deterministic case, when $W_t \equiv A$, we prove the large deviations principle at each node, and we find the corresponding rate function, equal at all nodes. We prove that its dependence on the (stochastic) system matrix A is fully captured by the left eigenvector a of A associated with the eigenvalue one, i.e., the left Perron vector of A. When the observations are Gaussian—independent, but non-identically distributed, we find a closed form expression for the rate function. Motivated by the fact that the rate function strongly depends on the eigenvector a, we formulate the following network design problem. For a given accuracy region, find the optimal vector a that maximizes the value of the rate function on this fixed region. We further show that for Gaussian observations with equal means (but different covariance matrices) this problem can be formulated as a semidefinite program (SDP) and thus can be solved efficiently. Simulation examples demonstrate that the optimized system significantly outperforms the system with the uniform left eigenvector a that, in the asymptotic sense, equally "weighs" all of the nodes' estimates. Finally, considering the special case when the observations are independent and identically distributed (i.i.d.), we reveal a very interesting property: the rate function, independently of the choice of A, always lies between the rate function of an isolated node I and the rate function of a fusion center NI, where N is the number of nodes in the network. Intuitively, this means that the distributed system is always better than an isolated node, and that, on the other hand, can never beat the performance of a fusion center having access to all the sensors' observations.

To provide a hint as to why the latter property holds, consider the log-moment generating function Λ of the nodes' observations and its conjugate I. Then, the (properly scaled) log-moment generating function of the estimate $X_{i,t}$ at node i has the following form: $\Lambda_t(\lambda) := 1/t \sum_{s=1}^{t} \sum_{j=1}^{N} \Lambda\left([A^{t-s}]_{ij}\lambda\right), t = 1, 2, \ldots$ (see ahead the derivations in (12)). Exploiting the fact that the powers of A are stochastic matrices together with certain nice properties of Λ (such as convexity) it can be shown that Λ_t is, for all t, "sandwiched" between the following two time-invariant functions:

$$N\Lambda\left(\frac{1}{N}\lambda\right) \leq \Lambda_t(\lambda) \leq \Lambda(\lambda). \tag{1}$$

In the deterministic case, the sequence $\Lambda_t(\lambda), t = 1, 2, \ldots$, has a point-wise limit, $\widetilde{\Lambda}(\lambda)$, and we immediately obtain by the Gärtner-Ellis theorem that the $X_{i,t}$'s satisfy the large deviations principle. The corresponding rate function $\widetilde{I}$ is given as the conjugate of $\widetilde{\Lambda}$. Then, relations (1)—through elementary properties of the conjugacy operation—yield that $\widetilde{I}$ must be sandwiched between I, the rate function of an isolated node, and NI, the rate function of a fusion center.

[1]With a stochastic matrix, rows sum to one, and all the entries are non-negative.

Proving the latter property in a general random matrices setting requires much more sophisticated techniques. Namely, we cannot use standard large deviations results like the Gärtner-Ellis theorem. Instead, proving this requires a *novel, non-trivial combination* of large deviations techniques (such as exponential Markov inequalities, exponential tilting, and Gaussian regularization), further combined with *novel intermediate results* (see Lemmas 8–10). A major reason for this is that the sequence of log moment generating functions $\Lambda_t(\lambda)$ that arises in the analysis does not have a limit and, moreover, functions $\lambda \mapsto \Lambda_t(\lambda)$ are not 1-coercive for any finite t.

Related work: Large deviations asymptotic performance of consensus+innovations algorithms has been previously studied in [9], [13], [19]–[21], and [22]. Reference [19] finds the exact large deviations rate of consensus algorithms for random topologies and therefore captures the effects of intermittent communications on large deviations performance of distributed inference. Reference [20] studies large deviations of the stochastic Riccatti equation for the distributed Kalman filter, and it provides an upper and a lower bound for the large deviations rate function. In our previous work [9], [13], we considered the case of i.i.d. networks, where each topology realization is symmetric. Under this model, [9] finds an upper and a lower bound for the rate function when the observations are Gaussian, and [13] extends the results of [9] to arbitrary distributions of sensor observations. Reference [21] extends this work to consensus based distributed detection with *constant* learning step. They show that the local decision statistics satisfy the large deviations principle and characterize the corresponding rate function. Reference [22] studies belief formations in social networks and also characterizes error exponents (Kullback-Leibler divergences) for distributed multiple hypothesis testing.

In this paper, we go beyond all these results in several important directions. First, we study here *directed* random networks, and, furthermore, we make no restrictions on the distribution of the system matrices; in particular, we allow for their arbitrary *time correlations*. Second, when the system matrices are deterministic, asymmetric, we fully characterize the rate function and show that it is amenable to optimization.

Regarding the large deviations literature, our results are related with those of [23]. This reference studies sequences of correlated random variables in $\mathbb{R}$ (scalar random variables) and, similarly to our paper, it is concerned with deriving bounds on the decay rates of the corresponding large deviations probabilities. The proposed methodology is based on transforming the random variables by an appropriate real-valued continuous function h, and then upper and lower bounding the log-moment generating functions of the transformed variables. In the special case when function h is the identity, the problem that we study in Theorem 2 and the one in [23] are similar and are essentially the following: derive bounds on the large deviations rates based on the bounds of the log-moment generating functions. However, there are several major differences between our paper and [23]. First, we study random *vector* sequences, i.e., sequences in $\mathbb{R}^d$, where the space dimension $d \geq 1$ is *arbitrary*. As a result, our upper and lower bounds on the large deviations rates are more general than those from [23]. A further important comment is that the proofs in [23] cannot be easily generalized to the $d > 1$

case[2]. Second, even if one considers the case $d = 1$, our large deviation bounds hold for much broader sets than the bounds in [23]. In particular, [23] proves the large deviations bounds for *finite* open and closed intervals, where, in addition, the open intervals are restricted to belong to a certain set, strictly smaller than $\mathbb{R}$ when the random variables have finite support. In contrast, we prove both the upper and the lower large deviations bounds with *full generality*, i.e., for arbitrary closed and open sets, respectively. This for example incurs no restrictions on the sets for the random variables with finite supports in contrast with the results in [23].

Notation: For arbitrary $d \in \mathbb{N} = \{1, 2, \ldots\}$, 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 with all entries equal $1/d$. 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 $a_i, i, j = 1, \ldots, d$. For a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we denote its domain by $\mathcal{D}_f = \{x \in \mathbb{R}^d : -\infty < f(x) < +\infty\}$; the subdifferential (gradient, when f is differentiable) of f at a point x by $\partial f(x) (\nabla f(x))$; log denotes the natural logarithm; for two sequences f_t and g_t that agree to first order in the exponent, $\lim_{t\to+\infty} \frac{1}{t} \log(f_t/g_t) = 0$, we shortly write $f_t \doteq g_t$. 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: $\Delta^{N-1} = \{\alpha \in \mathbb{R}^N : \alpha_i \geq 0, \sum_{i=1}^N \alpha_i = 1\}$. We let $\lambda_{\max}$ and λ_2, respectively, denote the maximal and the second largest (in modulus) eigenvalue of a square matrix; $\dagger$ denotes the pseudoinverse of a square matrix; and $\|\cdot\|$ denotes the spectral norm. For a matrix $S \in \mathbb{R}^{N\times N}$, we let $\mathcal{R}(S)$ denote the range of $S, \mathcal{R}(S) = \{Sx : x \in \mathbb{R}^N\}$; $\mathrm{tr}(S)$ denotes the trace of S; for N square matrices $S_1, \ldots, S_N$, we let $\mathrm{diag}\{S_1, \ldots, S_N\}$ denote the block-diagonal matrix whose ith block is S_i, for $i = 1, \ldots, N$. 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, the boundary, and the complement of an arbitrary set $D \subseteq \mathbb{R}^d$ are respectively denoted by $\overline{D}, D^o, \partial D$, and D^c; $\mathcal{B}(\mathbb{R}^d)$ denotes the Borel sigma algebra on $\mathbb{R}^d$; Ω denotes the probability space and ω denotes an element of Ω; $\mathbb{P}$ 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.

Paper Organization: In Section II we present the system model and formulate the problem that we study. In Section III we give preliminaries. Section IV presents our results for the deterministic case. Using the results of Section IV, Section V formulates the network design problem and solves it for the case of Gaussian observations with equal means. Section VI presents the fundamental bounds on the rate function for the generic case, when system matrices are random; proofs of this result are given in Subsections VI.A and VI.B. Simulation results are presented in Section VII, and the conclusion is given in Section VIII.

II. Problem Setup

This section explains the system model and the distributed inference algorithm that we study.

[2]For example, a major difficulty arises with (3.11) in [23].

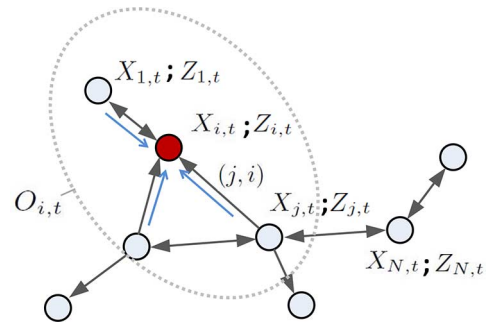


Fig. 1. Illustration of the problem setup; each node i updates its state $X_{i,t}$ according to (2).

Network Observations: Suppose that we have N geographically distributed agents (e.g., sensors, robots, humans) that monitor and collect observations about their environment. We denote the set of agents by $V = \{1, 2, \ldots, N\}$ such that $i \in V$ denotes the i-th agent. At each new time instant $t = 1, 2, \ldots$, each agent produces a d-dimensional observation vector. We denote by $Z_{i,t} \in \mathbb{R}^d$ the observation vector of agent i at time t, where we assume that the measurements are made synchronously across all agents. We denote by m_i the expected value of observations at node $i, m_i = \mathbb{E}[Z_{i,t}]$ (constant for all t).

Inter-Agent Communication: We assume that a direct communication is possible only between a subset of agents' pairs, e.g., the agents that are close enough to each other. (For instance, in a WSN, communication links are established only between sensors that lie within a certain, predefined distance r from each other.) We model the possible inter-agent communications via a directed graph $\widehat{G} = (V, \widehat{E})$, where the set $\widehat{E} \subseteq V \times V$ collects all possible (directed) communication links, i.e., all pairs (j, i) such that agent i can receive messages from agent j in a single hop manner. The links in $\widehat{E}$ should be understood only as potential communication channels. In other words, at a certain time t, agent j may decide whether to send or not send a message to agent i. Also, in case a message from j to i was sent, its reception at i could be unsuccessful due to imperfect channel effects (e.g., fading). For any link $(j, i) \in \widehat{E}$, we say that (j, i) is active at time t if at time t a message is sent from j and successfully received at i. We let E_t denote the set of all active links at time t. Accordingly, the neighborhood of node i at time t is $O_{i,t} = \{j : (j, i) \in E_t\}$, that is, $O_{i,t}$ is the set of all active links at time t that are pointing to i; for any $j \in O_{i,t}$, we say that j is an active neighbor of i. Finally, we denote by $G_t = (V, E_t)$ the graph realization at time t. Fig. 1 illustrates the problem setup.

Consensus+Innovations Based Distributed Inference: The distributed inference algorithm that we study operates as follows. Each node, over time, maintains a d-dimensional vector that serves as the node's estimate on the state of nature. The estimate of node i at time t is denoted by $X_{i,t}$, and we also refer to it as the state of node i. The estimates (states) are continuously improved over time twofold. First, each agent i incorporates its new observation $Z_{i,t}$ into its current state with the weight $1/t$ and forms an intermediate state update; subsequently, it transmits the intermediate state to (a subset of) its neighbors. Finally, agent i forms a convex combination (weighted average) of its

own and its active neighbors' intermediate states, with the coefficients $\{W_{ij,t} : j \in O_{i,t}\}, i \in V$. Mathematically, the state update of agent i is:

$$X_{i,t} = \sum_{j \in O_{i,t}} W_{ij,t} \left(\frac{t-1}{t} X_{j,t-1} + \frac{1}{t} Z_{j,t} \right), \qquad (2)$$

with the initialization $X_{i,0} = 0_d$. To derive a more compact representation, collect for each t the agents' weights $W_{ij,t}$ in an $N \times N$ matrix W_t as follows: for any pair $(j,i) \in \widehat{E}$ that satisfies $j \in O_{i,t}$, $[W_t]_{ij}$ is assigned the value $W_{ij,t}$, and equals zero otherwise, and for any $i \in V$, $[W_t]_{ii} = 1 - \sum_{j \in O_{i,t}} [W_t]_{ij}$. We refer to matrix W_t as the weight matrix. Due to the fact that, for any i, $\{W_{ij,t} : j \in O_{ij,t}\}$ form a convex combination, W_t is stochastic for any t. Further, let $\Phi(t,s)$, for $t \geq 1$ and $t \geq s \geq 1$, be defined as the matrix product $\Phi(t,s) = W_t \cdot \ldots \cdot W_s$, for $1 \leq s \leq t$. From (2), we then obtain:

$$X_{i,t} = \frac{1}{t} \sum_{s=1}^{t} \sum_{j=1}^{N} [\Phi(t,s)]_{ij} Z_{j,s}. \qquad (3)$$

Algorithms of the form (2) and (3) have been previously studied, e.g., in [7], [8], and [9].

We now state our assumptions on the weight matrices and the agents' observations.

Assumption 1 (Network and Observation Model):

1) Observations $Z_{i,t}, i = 1, \ldots, N, t = 1, 2, \ldots$ are independent both across nodes and over time;
2) For each agent $i, Z_{i,t}, t = 1, 2, \ldots$ are identically distributed;
3) Quantities W_t and $Z_{i,s}$ are independent for all i, s, t.

The model above is very general. In particular, in terms of the agents' interactions, it allows for directed topologies and asymmetric weight matrices, and it also allows for time dependencies between the weight matrices; directed topologies and temporal dependencies are cases that are much less studied in the literature. In terms of observations, we remark that the model above allows for non-identically distributed observations across nodes.

We next introduce the rates of large deviations and motivate their use for performance characterization of algorithm (2).

Rates of Large Deviations at Individual Agents: Suppose that, for some $i, X_{i,t}$ converges almost surely (a.s.) to a deterministic vector $\theta \in \mathbb{R}^d$, e.g., the vector of d parameters that the system wishes to estimate. In many scenarios, it is of interest to determine at what rate this convergence occurs. To explain why this is important, suppose that we wish to determine θ up to a certain accuracy defined by the accuracy region $C \subseteq \mathbb{R}^d$, where $\theta \in C$. Let T_i denote the time interval after which $X_{i,t}$ belongs to C with a prescribed, high probability, say 0.97. For convenience, define also the complement of $C, D = \mathbb{R}^d \setminus C$, usually called the deviation set. Since $X_{i,t}$ converges a.s. to θ, we know that the probability that $X_{i,t}$ remains outside of $C, \mathbb{P}(X_{i,t} \in D)$, vanishes as $t \to +\infty$. The question that we ask then is how fast this probability vanishes with time. It turns out that in many scenarios this convergence is exponential (see [13] for the scalar, $d = 1$ case). That is:

$$\mathbb{P}(X_{i,t} \in D) \stackrel{\bullet}{=} e^{-tI_i(D)}, \qquad (4)$$

for a certain function I_i, where, we recall, $\stackrel{\bullet}{=}$ means that the two functions agree to first order in the exponent. Function $I_i : \mathcal{B}(\mathbb{R}^d) \mapsto \mathbb{R}^+$ is usually called the rate function. Relating I_i with time T_i, we see that T_i can be approximately computed as

$$T_i \approx -\frac{\log(1 - 0.97)}{I_i(D)}. \qquad (5)$$

The quality of the approximation in (5) improves for higher accuracies (i.e., smaller region C around θ). In the context of, e.g., Neyman-Pearson hypothesis testing, rates I_i directly correspond to error exponents: taking, for example D to be the false alarm region $[0, +\infty)$ under H_0, $I_i(D)$ gives the error exponent of the false alarm probability at sensor i. The problem that we address in this paper is finding the rate functions $I_i, i \in V$:

$$\lim_{t \to +\infty} -\frac{1}{t} \log \mathbb{P}(X_{i,t} \in D) = I_i(D), \qquad (6)$$

whenever the limit above exists for any set $D \in \mathcal{B}(\mathbb{R}^d)$. For further details on the use of large deviations rate functions in probabilistic inference, we refer the reader to [24]–[26].

III. Preliminaries

Before we start our analysis, we first review in Subsection III.A basic large deviations concepts and tools. Subsection III.B then describes the large deviations performance of an isolated agent and a fusion node.

A. Large Deviations Preliminaries

We define the large deviations principle and introduce, for each i, the logarithmic moment generating function of observations $Z_{i,t}$. We then define the conjugate of a function and state some important properties of log-moment generating functions and their conjugates in general, and in our particular setup as well.

Large Deviations Principle: A rate function is any function that is lower semi-continuous, or equivalently, that has closed sublevel sets. A sequence of random variables $\widehat{Z}_t \in \mathbb{R}^d$ is said to satisfy the large deviations principle (LDP) with rate function $\widehat{I}$ if for any measurable set $D \in \mathcal{B}(\mathbb{R}^d)$ it holds that

$$\begin{aligned} -\inf_{x \in D^o} \widehat{I}(x) &\leq \liminf_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(\widehat{Z}_t \in D\right) \\ &\leq \limsup_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(\widehat{Z}_t \in D\right) \\ &\leq -\inf_{x \in \overline{D}} \widehat{I}(x). \end{aligned} \qquad (7)$$

Essentially, what the large deviations principle tells is that, for any (nice enough) set D, probabilities that $\widehat{Z}_t$ belongs to D decay with t exponentially, with the rate equal to $\widehat{I}(D) = \inf_{x \in D} \widehat{I}(x)$. Key objects in proving the large deviations principle and computing the rate function in general (see Cramér's and Gärtner-Ellis theorem [27], [28]) are the log-moment generating function and its conjugate, which we introduce next.

Log-Moment Generating Function of Observations $Z_{i,t}$: The log-moment generating function $\Lambda_i : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ corresponding to $Z_{i,t}$ is given by:

$$\Lambda_i(\lambda) = \log \mathbb{E}\left[e^{\lambda^\top Z_{i,t}}\right], \text{ for } \lambda \in \mathbb{R}^d. \tag{8}$$

For the special case when all the agents' observations are identically distributed, we let Λ denote the corresponding log-moment generating function, $\Lambda \equiv \Lambda_i$, for any i.

The second key object of interest in our analysis is the conjugate of a log-moment generating function. Let $\widehat{\Lambda}$ be the log-moment generating function of a d-dimensional random vector $\widehat{Z}$. Then, the conjugate, or the Fenchel-Legendre transform, of $\widehat{\Lambda}$ is given by

$$\widehat{I}(x) = \sup_{\lambda \in \mathbb{R}^d} x^\top \lambda - \widehat{\Lambda}(\lambda), \text{ for } x \in \mathbb{R}^d. \tag{9}$$

When $Z_{i,t}$ are i.i.d., we will denote by I the conjugate of Λ. We next give the Gärtner-Ellis theorem [28][3], which shows the relevance of the conjugate of a log-moment generating function for characterizing the large deviations performance. We will use this result in the next section, when considering the case of deterministic matrices.

Gärtner-Ellis Theorem: Let $\widehat{X}_t, t = 1, 2, \ldots$, be a sequence of random vectors in $\mathbb{R}^d$, and, for each t, let $\widehat{\Lambda}_t$ denote the scaled log-moment generating function of $\widehat{X}_t, \widehat{\Lambda}_t := \frac{1}{t}\log \mathbb{E}[e^{t\lambda \widehat{X}_t}], \lambda \in \mathbb{R}^d$. If, for each $\lambda \in \mathbb{R}^d$, the following limit exists:

$$\lim_{t \to +\infty} \widehat{\Lambda}_t(\lambda) =: \widehat{\Lambda}(\lambda), \tag{10}$$

and, additionally, $\mathcal{D}_{\widehat{\Lambda}} = \mathbb{R}^d$, then the sequence $\widehat{X}_t$ satisfies the large deviations principle with the rate function $\widehat{I}$, where $\widehat{I}$ is the conjugate of $\widehat{\Lambda}$.

Example 1 (Gaussian Observations): It can be shown by simple algebraic manipulations that when $Z_{i,t}$ is i.i.d., Gaussian, with mean value m and covariance matrix S, the log-moment generating function Λ and its conjugate I are both quadratic and given, respectively, by [28]:

$$\Lambda(\lambda) = m^\top \lambda + \frac{1}{2}\lambda^\top S \lambda, \; I(x) = \frac{1}{2}(x-m)^\top S^{-1}(x-m).$$

To simplify our analysis, we make the following assumption.

Assumption 2: $\mathcal{D}_{\Lambda_i} = \mathbb{R}^d$, i.e., $\Lambda_i(\lambda) < +\infty$ for all $\lambda \in \mathbb{R}^d$, for each i.

Assumption 2 holds for arbitrary Gaussian and discrete random vectors, and also for many other commonly used distributions; we refer the reader to [13] for examples of random vectors beyond the ones mentioned here that have a finite log-moment generating function.

Properties of Log-Moment Generating Functions and Their Conjugates: For future reference, we list the properties that an arbitrary log-moment generating function $\widehat{\Lambda}$ and its conjugate $\widehat{I}$ satisfy; proofs can be found in [29, p. 8] and [28, p. 27, 35].

[3]Note that we use here the variant of the Gärtner-Ellis theorem that asserts the *full* LDP under the additional assumption that the limit function, see (10), is finite on the whole space; see Exercise 2.3.20 in [28].

Lemma 1 (Properties of a Log-Moment Generating Function and Its Conjugate): Consider the log-moment generating function $\widehat{\Lambda}$ and its conjugate $\widehat{I}$, associated with an arbitrary d-dimensional random vector $\widehat{Z}$. Let $\theta = \mathbb{E}[\widehat{Z}]$. Then:

1) function $\widehat{\Lambda}$ satisfies:
   a) $\widehat{\Lambda}(0) = 0$ and $\nabla\widehat{\Lambda}(0) = \theta$, when $0 \in \mathcal{D}^{\circ}_{\widehat{\Lambda}}$;
   b) $\widehat{\Lambda}(\cdot)$ is lower semi-continuous and convex;
   c) $\widehat{\Lambda}(\cdot)$ is C^∞ on $\mathcal{D}^{\circ}_{\widehat{\Lambda}}$;
2) and function $\widehat{I}$ satisfies:
   a) $\widehat{I}$ is nonnegative and $\widehat{I}(\theta) = 0$;
   b) $\widehat{I}$ is lower semi-continuous and convex;
   c) if $0 \in \mathcal{D}^{\circ}_{\widehat{\Lambda}}$, then $\widehat{I}$ has compact level sets.
   d) $\widehat{I}$ is differentiable on $\mathcal{D}^{\circ}_{\widehat{I}}$.

We end this subsection by stating a simple but important property of the log-moment generating function that follows from its convexity and zero value at the origin. We note that the right-hand side of inequality (11) was previously proven in [13] (for the case $d = 1$).

Lemma 2: Let $\widehat{\Lambda}$ be an arbitrary log-moment generating function. For any $\alpha \in \Delta^{N-1}$ and $\lambda \in \mathbb{R}^d$,

$$N\widehat{\Lambda}\left(\frac{1}{N}\lambda\right) \leq \sum_{i=1}^{N} \widehat{\Lambda}(\alpha_i \lambda) \leq \widehat{\Lambda}(\lambda). \tag{11}$$

Proof: We first prove the right-hand side inequality in (11). (The proof is analogous to the proof of the same inequality for the special case $d = 1$ [13]; for completeness, we provide the proof here.) Fix $\varsigma \in [0, 1]$. Then, by convexity of $\widehat{\Lambda}$ and the fact that $\widehat{\Lambda}(0) = 0$, we have

$$\widehat{\Lambda}(\varsigma\lambda) = \widehat{\Lambda}(\varsigma\,\lambda + (1-\varsigma)\,0) \leq \varsigma\widehat{\Lambda}(\lambda) + (1-\varsigma)\widehat{\Lambda}(0) = \varsigma\widehat{\Lambda}(\lambda).$$

Now, fix an arbitrary $\alpha \in \Delta^{N-1}$. Applying the preceding inequality for $\varsigma = \alpha_i$, for $i = 1, \ldots, N$, yields the claim by summing out the resulting left and right hand sides.

To prove the left hand side inequality in (11), consider the function $g_\lambda : \mathbb{R}^N \mapsto \mathbb{R}, g_\lambda(\beta) = \sum_{i=1}^N \widehat{\Lambda}(\beta_i \lambda)$, for $\beta \in \mathbb{R}^N$. We prove the claim if we show that the minimum of g_λ over the unit simplex Δ^{N-1} is attained at $1/N\, 1_N = (1/N, \ldots, 1/N) \in \Delta^{N-1}$. Since g_λ is convex (being the sum of convex functions), it suffices to show that there exists a Lagrange multiplier $\nu \in \mathbb{R}$ such that the pair $(1/N\, 1_N, \nu)$ satisfies the Karush-Kuhn-Tucker (KKT) conditions [30]. To this end, define the Lagrangian $L(\beta, \nu) = g_\lambda(\beta) + \nu(1_N^\top \beta - 1)$, for some $\nu \in \mathbb{R}, \beta \in \mathbb{R}^N$. We have

$$\partial_{\beta_i} L(\beta, \nu) = \lambda^\top \nabla\widehat{\Lambda}(\beta_i \lambda) + \nu.$$

Taking $\beta_i = 1/N$ and $\nu = -\lambda^\top \nabla\widehat{\Lambda}(1/N\, \lambda)$, proves the claim. □

B. Setting the Benchmarks: Isolation and Fusion

To set benchmarks for the performance of distributed inference (2), we consider two extreme cases of the agents' cooperation: 1) complete agent's isolation, when an agent operates alone, making inferences based on its own observations only; and 2) network-wide fusion, when each agent has access to all of the observations. Mathematically, the state of agent i corresponding to these two cases are as

follows: $X_{i,t}^{\text{isol}} = 1/t \sum_{s=1}^{t} Z_{i,s}$, for $i \in V$, for the case of isolated agents (obtained when in (3) $W_t \equiv I_d$), and $X_t^{\text{cen}} = 1/(Nt) \sum_{s=1}^{t} \sum_{i=1}^{N} Z_{i,s}$, for the case of fusion center (obtained when $W_t \equiv J_d$). In Example 2 we compute the corresponding large deviation rates, and we also show that, when the observations are i.i.d., the fusion-based rate scales linearly (with constant one) with the number of participating agents.

Example 2: Suppose that $Z_{i,t}$ are i.i.d. for all i and t, and recall that I denotes the conjugate of the log moment generating function of $Z_{i,t}$. Then,

1) for each i, the sequence $X_{i,t}^{\text{isol}}$ satisfies the LDP with rate function $I_i^{\text{isol}} \equiv I$;
2) the sequence X_t^{cen} satisfies the LDP with rate function $I^{\text{cen}} \equiv NI$.

We remark that both results follow as direct applications of Cramér's theorem [27], [28, p. 36], where to prove part (1) for any given node i one uses the sequence of i.i.d. variables $Z_{i,t}$, and to prove part (2) the sequence of i.i.d. variables $1/N \sum_{i=1}^{N} Z_{i,t}$. Also, clearly, with both the isolated nodes and the fusion center cases, the corresponding states $X_{i,t}^{\text{isol}}, i = 1, \ldots, N$, and X_t^{cen} converge a.s. to $m := \mathbb{E}[Z_{i,t}]$, which follows as a direct application of the strong law of large numbers.

Example 2 asserts that the rate function of any isolated agent i is $I_i^{\text{isol}} \equiv I$, where I is the conjugate of the log-moment generating function of its observation, whereas the rate function of the network-wide fusion is N times higher, $I_i^{\text{cen}} \equiv NI$. Intuitively, for the general case of algorithm (2), we expect that the rate function of a fixed agent i should be between these two functions, I and NI. It turns out that this is indeed the case—Corollary 1 proves this for deterministic matrices, and Theorem 2 later in Section VI confirms that this is true even for arbitrary (asymmetric) random matrices.

IV. Rate Functions I_i for Deterministic Weight Matrices

This section considers deterministic weight matrices. The first result that we present, Theorem 1, computes the rate functions I_i for the case when the weight matrices at all times are equal to a stochastic matrix A such that $|\lambda_2(A)| < 1$. (This means that the underlying network has only one initial class[4], e.g., [33], [32].) We then focus on the special case when all observations are Gaussian (with possibly different parameters across agents), and we calculate the rate functions in closed form. Further, we formulate the problem of optimal network design and show that it can be efficiently solved by an SDP when the observations are Gaussian.

Theorem 1: Let $W_t \equiv A$ for each t and let Assumptions 1 and 2 hold. Suppose that $|\lambda_2(A)| < 1$ and let a denote the left Perron vector of A whose entries sum to one[5]. Then, for each $i, X_{i,t}, t = 1, 2, \ldots$ satisfies the LDP with the rate function $I_i \equiv \widetilde{I}$, where $\widetilde{I}$ is the conjugate of

$$\widetilde{\Lambda}(\lambda) := \sum_{j=1}^{N} \Lambda_j(a_j \lambda),\ \lambda \in \mathbb{R}^d.$$

Moreover, for each $i, X_{i,t}$ converges a.s. to $\widetilde{m} := \sum_{j=1}^{N} a_j m_j$.

Proof: We prove the first part of the theorem by applying the Gärtner-Ellis theorem [28]. Fix $i \in V$ and let $\Lambda_t(\lambda) := \frac{1}{t} \log \mathbb{E}\left[e^{t\lambda X_{i,t}}\right]$, for $\lambda \in \mathbb{R}^d$. In order to apply the Gärtner-Ellis theorem (see Section III), one needs to verify that the following condition is fulfilled: for every $\lambda \in \mathbb{R}^d$, the sequence $\Lambda_t(\lambda), t = 1, 2, \ldots$ has a limit in $\mathbb{R}$. Using that $Z_{i,t}$ are independent and that $\Phi(t, s) = A^{t-s+1}$ are constant, we obtain

$$\begin{aligned}
\Lambda_t(\lambda) &= \frac{1}{t} \log \mathbb{E}\left[e^{\lambda \sum_{s=1}^{t} \sum_{j=1}^{N} [\Phi(t,s)]_{ij} Z_{j,s}}\right] \\
&= \frac{1}{t} \sum_{s=1}^{t} \sum_{j=1}^{N} \log \mathbb{E}\left[e^{\lambda [\Phi(t,s)]_{ij} Z_{j,s}}\right] \\
&= \frac{1}{t} \sum_{s=1}^{t} \sum_{j=1}^{N} \Lambda_j\left([A^{t-s+1}]_{ij} \lambda\right) \\
&= \sum_{j=1}^{N} \frac{1}{t} \sum_{r=1}^{t} \Lambda_j\left([A^r]_{ij} \lambda\right). \qquad (12)
\end{aligned}$$

From $|\lambda_2(A)| < 1$ we have that $A^r \to 1_N a^\top$ as $r \to +\infty$, where a is the left Perron vector of A that satisfies $a > 0$ and $1_N^\top a = 1$ (e.g., Theorem 8.5.1 in [34]). Hence, for any i, we have that $[A^r]_{ij} \to a_j$. Consider now a fixed j. Then, by continuity of $\Lambda_j, \Lambda_j\left([A^r]_{ij} \lambda\right) \to \Lambda_j(a_j \lambda)$, and hence the Cesàro averages must converge to the same number:

$$\lim_{t \to +\infty} \frac{1}{t} \sum_{r=1}^{t} \Lambda_j\left([A^r]_{ij} \lambda\right) = \Lambda_j(a_j \lambda).$$

Going back to (12) and taking the limit yields $\lim_{t \to +\infty} \Lambda_t(\lambda) = \sum_{j=1}^{N} \Lambda(a_j \lambda)$. Since, by Assumption 2, $\mathcal{D}_\Lambda = \mathbb{R}^d$, conditions for applying the Gärtner-Ellis theorem are fulfilled, and we have that, for each $i, X_{i,t}$ satisfies the large deviations principle with the rate function equal to the conjugate of $\sum_{j=1}^{N} \Lambda_j(a_j \lambda)$.

The proof of the almost sure convergence is given in the Appendix. □

Let G denote the induced graph of A, i.e., $G = (V, E)$ where $E = \{(i, j) : A_{ji} > 0\}$, e.g., [35].

Corollary 1: When $Z_{i,t}$ are i.i.d. (identical agents), it holds

$$I \leq \widetilde{I} \leq NI, \qquad (13)$$

where I is the conjugate of an agent's log-moment generating function $\Lambda \equiv \Lambda_i$ and the inequalities in (13) hold in the pointwise sense. Moreover, the lower bound in (13) is attained whenever there exists a "leader" agent i that satisfies $A_{ii} = 1$ and for any j there is a (directed) path from i to j in the induced graph of A. The upper bound is attained when A is doubly stochastic with positive diagonals and the induced graph of A is strongly connected.

[4]An initial class of a directed graph G is any communication class of G that has no incoming edges [31]. We also note that initial classes of G correspond to essential classes of the transpose of G (the graph that results from reversing the directions of edges in G [32]).

[5]Note that since A is stochastic—hence non-negative, its left eigenvector corresponding to the maximal (unit) eigenvalue—the Perron vector, must have all entries non-negative [34].

The proof of Corollary 1 is given in an extended arxiv version of this paper [36].

Rate $\widetilde{I}$ for Gaussian Observations: Of special interest is the case when observations $Z_{i,t}$ are all Gaussian. For this case, Lemma 3 gives a closed form expression for the rate function $\widetilde{I}$. The proof follows by simple algebraic manipulations, and hence we omit it.

Lemma 3: Suppose that $Z_{j,t} \sim \mathcal{N}(m_j, S_j)$, for $j \in V$, where S_j, for each j, is a positive definite matrix. Function $\widetilde{I}$ from Theorem 1 is then given by

$$\widetilde{I}(x) = \frac{1}{2}(x-\widetilde{m})\widetilde{S}^{-1}(x-\widetilde{m}), \tag{14}$$

where $\widetilde{m} = \sum_{j=1}^{N} a_j m_j$ and $\widetilde{S} = \sum_{j=1}^{N} a_j^2 S_j$. In particular, when $m_j \equiv m$ and $S_j \equiv S$, $\widetilde{I}(x) = 1/(\sum_{j=1}^{N} a_j^2) I(x)$, where $I(x)$ is the nodes' individual rate function given in Example 1.

Remark 1: It is possible to determine $\widetilde{I}$ analytically even when matrices $S_j, j = 1, \ldots, N$, and vector a are such that $\widetilde{S}$ is not invertible. It can be shown that the expression for $\widetilde{I}$ for this case is:

$$\widetilde{I}(x) = \begin{cases} (x-\widetilde{m})^{\top}\widetilde{S}^{\dagger}(x-\widetilde{m}), & x \in \mathcal{R}(\widetilde{S}) \\ +\infty, & \text{otherwise} \end{cases}.$$

V. Network Design

In this section, we focus on the dependence of the distributed inference rate function on weight matrix A. Subsection V.A considers optimization and performance limits as regards the weight matrix A in time-asymptotic regimes. Subsection V.B discusses practical, finite time regimes and provides ways on how in principle one optimizes A, while efficient, tractable heuristics are presented in Section VII.

A. Asymptotic Regime

From Theorem 1 and Corollary 1, we can see that the large deviations (asymptotic) performance (in the sense of (6)) of algorithm (2) critically depends on the choice of the weight matrix A, where the dependence is only through the left Perron eigenvector a. We therefore pose the problem of optimizing a, for a fixed desired accuracy region C, such that the value of the corresponding large deviation rate is maximized:

$$\begin{array}{ll} \text{maximize} & \inf_{x \in \mathbb{R}^d \setminus C} \widetilde{I}(x) \\ \text{subject to} & a \in \Delta^{N-1} \end{array}. \tag{15}$$

Here, $\widetilde{I}$ is the rate function from Theorem 1. Note that optimization vector a in (15) must be constrained to belong to the simplex due to the fact that, for a given stochastic matrix A, the corresponding rate function $\widetilde{I}$ is defined by the left Perron vector of A, which is non-negative, and whose entries sum to one. We denote by $a_C^{\star}$ and $I_C^{\star}$, respectively, an optimal solution and the optimal value of problem (15).

Remark 2: In general, there is no guarantee that $a_C^{\star}$ is achievable for a given topology $\widehat{G}$. In that sense, the optimal value of (15) is only an upper bound on the achievable large deviations performance. However, as we show in Section VII, the knowledge of $a_C^{\star}$ can be instrumental in deriving efficient heuristics for optimizing the weight matrix A under given topology constraints.

For networks of small to moderate sizes and generic (non-Gaussian) sensor observations, optimization problem (15) can be in principle solved by brute force (grid search). We exploit here the analytical expression (14) for the rate function from Lemma 3, to show that, for Gaussian observations, problem (15) can be solved efficiently. We assume that all nodes are observing the same set of physical quantities $\theta = (\theta_1, \ldots, \theta_d)^{\top}$, embedded in local sensor noises. Hence, the observations $Z_{i,t}$ have the same expected value $\theta =: m \equiv m_i$ across all nodes. We show in Lemma 4 that, when C is a ball, (15) can be formulated as an SDP (a convex problem). The proof of this result is given in an extended arxiv version of this paper [36].

Lemma 4: Consider the setup of Lemma 3 when $m_i \equiv m$. When the confidence set C is an Euclidean ball of some arbitrary radius $\zeta > 0$ centered at m, $B_m(\zeta)$, the optimal solution of (15) is obtained by solving:

$$\begin{array}{ll} \text{minimize} & \gamma \\ \text{subject to} & \begin{bmatrix} \gamma I_d & \mathcal{I}\widetilde{\mathcal{S}} \\ \widetilde{\mathcal{S}}\mathcal{I}^{\top} & I_{Nd} \end{bmatrix} \succeq 0, \\ & a \in \Delta^{N-1} \end{array} \tag{16}$$

where $\widetilde{\mathcal{S}} \in \mathbb{R}^{Nd \times Nd}$ is a block diagonal matrix given by $\widetilde{\mathcal{S}} = \text{diag}\left\{a_1 S_1^{1/2}, \ldots, a_N S_N^{1/2}\right\}$, and $\mathcal{I} = [I_d \ldots I_d] \in \mathbb{R}^{d \times Nd}$, where I_d repeats N times. Furthermore, $I_C^{\star} = \zeta^2/(2\gamma^{\star})$, where $\gamma^{\star}$ is the optimum of (16).

Remark 3: Although problem (15) involves the expected value of the observations m (which we do not know), it is clear from the equivalent reformulation (16) that, under the stated assumption, the knowledge of m is not needed for discovering the optimal a in (15). We also remark that, under the same assumptions, the solution of (15) does not depend on the particular accuracy ζ: once (16) is solved, the same vector $a_C^{\star}$ applies for all $C = B_m(\zeta), \zeta > 0$.

Remark 4: When the observations are one-dimensional ($d = 1$), it can be shown that the SDP in (16) reduces to a quadratic program (QP).

B. Finite Time Regime Considerations

We now examine more closely the practical, finite time regime. This will also present certain aspects of the performance of distributed inference (2) (in the sense of probabilities of the form (4)) that are not captured by the large deviations rates. To this end, pick an arbitrary node i and represent its error probability $P_{i,t} := \mathbb{P}(X_{i,t} \notin C)$ in the following form:

$$P_{i,t} = \rho_{i,t} e^{-t\widetilde{I}_{i,t}(C)}. \tag{17}$$

By Theorem 1 it holds that:

$$\lim_{t \to +\infty} \widetilde{I}_{i,t}(C) = \widetilde{I}(C) := \inf_{x \in \mathbb{R}^d \setminus C} \widetilde{I}(x), \tag{18}$$

and $\lim_{t \to +\infty} 1/t \log \rho_{i,t} = 0$. Note that, even though $\widetilde{I}(C)$ is equal across nodes (i.e., all nodes have equal asymptotic performance), quantities $\rho_{i,t}$ and $\widetilde{I}_{i,t}(C)$ may be different across different nodes; therefore, finite time performance of different nodes may be different, as confirmed in subsequent simulations.

Naturally, finite time performance (and hence, the effects of $\rho_{i,t}$ and $I_{i,t}(C)$) is to a large extent determined by how fast the

weight matrix A approaches its asymptotic limit $1_N a^\top$. This is in turn determined by the modulus of the second largest eigenvalue of A: the smaller $|\lambda_2(A)|$ is, the faster this convergence is. Therefore, besides optimizing rate $\widetilde{I}(C)$, one also wants to make $|\lambda_2(A)|$ small. This yields the following optimization problem:

$$\begin{array}{ll} \text{minimize} & |\lambda_2(A)| - \tau \inf_{x\in\mathbb{R}^d\setminus C} \widetilde{I}(x) \\ \text{subject to} & A1_N = 1_N \\ & a^\top A = a^\top \\ & a \in \Delta_{N-1} \\ & A \in \mathcal{A} \end{array} . \qquad (19)$$

Here, the optimization variable is the $N \times N$ matrix A and $a = a(A)$, a function of A, is its left Perron eigenvector. Further, $\mathcal{A} = \{A \in \mathbb{R}_+^{N\times N} : A_{ij} = 0, \text{if } (j,i) \notin \widehat{G}\}$ is the set of matrices with a given sparsity pattern dictated by the network topology, and $\tau > 0$ is the weight that balances the trade-off between the two objectives in (19). Problem (19) is very difficult to solve optimally (e.g., $|\lambda_2(A)|$ is not a convex function, and also the problem involves simultaneous optimization of matrix A and its left eigenvector a). In Section VII, we present a convex programming alternative to (19), whose efficiency we demonstrate by simulations.

VI. Universal Bounds on the Rate Functions for General, Random Weight Matrices

We have seen in the previous section (Corollary 1) that, when the weight matrices W_t are deterministic and constant, the states exhibit a very interesting and fundamental property: their large deviation probabilities $\mathbb{P}(X_{i,t} \in D)$ have rates that are always lower than the corresponding rate of the fusion center, and always higher than the corresponding rate of a node working in isolation. Theorem 2 that we present next asserts that this property in fact holds, not only for deterministic, but for arbitrary sequences of random weight matrices.

Theorem 2: Consider the distributed inference algorithm (2) under Assumptions 1 and 2, when $Z_{i,t}$ are i.i.d. (identical agents). For any measurable set $G \subseteq \mathbb{R}^d$, for each i:

$$-\inf_{x\in G^o} NI(x) \le \liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(X_{i,t} \in G\right) \qquad (20)$$

$$\le \limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}(X_{i,t} \in G) \le -\inf_{x\in\overline{G}} I(x). \qquad (21)$$

Theorem 2 asserts that, no matter how we design the agents' interactions (represented by the weight matrices), in terms of large deviations performance, algorithm (2) can never be worse than when a node is working in isolation, but it also can never beat the fusion center. This result is important as it provides fundamental bounds for large deviations performance of *any* algorithm of the form (2) that satisfies Assumptions 1 and 2 and processes i.i.d. observations. In the next two subsections we state our proofs of Theorem 2.

A. Proof of the Upper Bound

Fix an arbitrary $i \in V$. To prove (21) for node i, it suffices to show that, for any closed set F,

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}(X_{i,t} \in F) \le -\inf_{x\in F} I(x). \qquad (22)$$

To see why this is true, note that, for an arbitrary measurable set D, there holds $\mathbb{P}(X_{i,t} \in D) \le \mathbb{P}\left(X_{i,t} \in \overline{D}\right)$. Applying (22) to the closed set $F = \overline{D}$ yields (21).

The proof of (22) consists of the following three steps.

Step 1: We use the exponential Markov inequality, together with conditioning on the matrices $W_1, \ldots, W_t$, to show that, for any measurable set $D \subseteq \mathbb{R}^d$,

$$\frac{1}{t} \log \mathbb{P}(X_{i,t} \in D) \le -\inf_{x\in D} \lambda^\top x - \Lambda(\lambda). \qquad (23)$$

Step 2: In the second step, we show that (23) is a sufficient condition for (22) to hold for all *compact* sets F. Lemma 5 formalizes this statement.

Lemma 5: Suppose that (23) holds for any measurable set $D \subseteq \mathbb{R}^d$. Then the inequality (22) holds for all compact sets F.

The proof of Lemma 5 uses the standard "finite cover" argument: for a compact set F, a finite number of balls forming a cover of F is constructed, and then (23) is applied to each of the balls. The details of this derivation are given in Appendix A of an extended arxiv version of this paper [36].

Step 3: So far, *Steps 1* and *2* together imply that (22) holds for all compact sets. To extend (22) to all *closed* sets F, by a well known result from large deviations theory, Lemma 1.2.18 from [28], it suffices to show that the sequence of measures $\mu_{i,t} : \mathcal{B}(\mathbb{R}^d) \mapsto [0,1], \mu_{i,t}(D) := \mathbb{P}(X_{i,t} \in D)$ is exponentially tight. We prove this by considering the family of compact sets $H_\rho := [-\rho, \rho]^d$, with ρ increasing to infinity. The result is given in Lemma 6, and the proof can be found in Appendix B of an extended arxiv version of this paper [36].

Lemma 6: For every $i \in V$,

$$\lim_{\rho\to+\infty} \limsup_{t\to+\infty} \mu_{i,t}\left(H_\rho^c\right) = -\infty. \qquad (24)$$

Hence, the sequence $\{\mu_{i,t}\}_{t=1,2,\ldots}$ is exponentially tight.

We now provide the details of *Step 1*.

Step 1. The proof of (23) is based on two key arguments: exponential Markov inequality [37] and the right hand side inequality of Lemma 2. For any measurable set $D \subseteq \mathbb{R}^d$ and any $\lambda \in \mathbb{R}^d$, by the exponential Markov inequality, we have

$$1_{\{X_{i,t}\in D\}} \le e^{t\lambda^\top X_{i,t} - t\inf_{x\in D} \lambda^\top x}, \qquad (25)$$

which, after computing the expectation, yields

$$\mathbb{P}(X_{i,t} \in D) \le e^{-t\inf_{x\in D} \lambda^\top x} \mathbb{E}\left[e^{t\lambda^\top X_{i,t}}\right]. \qquad (26)$$

We now focus on the right hand side of (26). Conditioning on $W_1, \ldots, W_t$, the summands in (3) become independent, and using the fact that the $Z_{i,t}$'s are i.i.d. with the same log-moment generating function Λ, we obtain

$$\mathbb{E}\left[e^{t\lambda^\top X_{i,t}} \middle| W_1, \ldots, W_t\right] = e^{\sum_{s=1}^t \sum_{j=1}^N \Lambda([\Phi(t,s)]_{ij}\lambda)}. \qquad (27)$$

Applying now the right-hand side inequality from Lemma 2 to $\sum_{j=1}^N \Lambda\left([\Phi(t,s)]_{ij}\lambda\right)$ for each fixed s (note that, for a fixed s, $[[\Phi(t,s)]_{i1}, \ldots, [\Phi(t,s)]_{iN}] \in \Delta^{N-1}$), it follows that the conditional expectation above is upper bounded by $e^{t\Lambda(\lambda)}$, i.e.,

$$\mathbb{E}\left[e^{t\lambda^\top X_{i,t}} \middle| W_1, \ldots, W_t\right] \le e^{t\Lambda(\lambda)}, \qquad (28)$$

for any $\lambda \in \mathbb{R}^d$. Since in (28) $W_1, \ldots, W_t$ were arbitrary, taking the expectation, we get $\mathbb{E}[e^{t\lambda^\top X_{i,t}}] \leq e^{t\Lambda(\lambda)}$. Combining this with (26), we finally obtain

$$\frac{1}{t}\log \mathbb{P}(X_{i,t} \in D) \leq -\inf_{x\in D} \lambda^\top x + \Lambda(\lambda). \tag{29}$$

B. Proof of the Lower Bound

We prove (20) following the general lines of the proof of the Gärtner-Ellis theorem lower bound, see [28]. However, as we will see later in this proof, we encounter several difficulties along the way that force us to depart from the standard Gärtner-Ellis method and use finer arguments. The main reason for this is that, in contrast with the setup of the Gärtner-Ellis theorem, the sequence of the (scaled) log-moment generating functions of $X_{i,t}$ (see ahead (31)) need not have a limit. Nevertheless, with the help of Lemma 2, we will be able to "sandwich" each member of this sequence between $\Lambda(\cdot)$ and $N\Lambda(1/N\cdot)$. This is the key ingredient that allows us to derive (20). The proof is organized in the following four steps.

Step 1: In this step, we derive a sufficient condition, given in Lemma 7, for (20) to hold. Namely, to prove (20) for a given set D, it suffices to confine $X_{i,t}$ to a smaller region $B_x(\delta)$ within D, and show that, conditioned on any realization of the matrices $W_1, \ldots, W_t$, the rate of this event is at most $NI(x)$. This implication is proven by applying Fatou's lemma [37] to the sequence of random variables $R_t := \frac{1}{t}\log \mathbb{P}\left(X_{i,t} \in D | W_1, \ldots, W_t\right)$, and then combining the obtained result with the simple fact that, for every $x \in D^o$ and all δ sufficiently small, $B_x(\delta) \subseteq D$. The proof is given in the Appendix. We remark that the sufficient condition (30) below is with respect to all possible realizations of sequences of random matrices $W_1, W_2, \ldots$ (i.e., for all possible $\omega \in \Omega$).

Lemma 7: If for every $x \in \mathbb{R}^d$ and $\omega \in \Omega$,

$$\lim_{\delta\to 0} \liminf_{t\to+\infty} \frac{1}{t}\log \mathbb{P}\left(X_{i,t} \in B_x(\delta) | W_1, \ldots, W_t\right) \geq -NI(x), \tag{30}$$

then (20) holds for all measurable sets D.

Step 2: To prove (30), we introduce the scaled log-moment generating function of $X_{i,t}$, under the conditioning on $W_1, \ldots, W_t$,

$$\Lambda_t(\lambda) := \frac{1}{t}\log \mathbb{E}\left[e^{t\lambda^\top X_{i,t}} \middle| W_1, \ldots, W_t \right]. \tag{31}$$

It can be shown (similarly as in *Step 1* of the proof of the upper bound) that, for any $\lambda \in \mathbb{R}^d$,

$$\Lambda_t(\lambda) = \frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{N} \Lambda\left([\Phi(t,s)]_{ij}\lambda\right), \tag{32}$$

where, we recall, $\Phi(t,s) = W_t \cdot \ldots \cdot W_s$. Note that Λ_t is convex and differentiable. However, Λ_t is not necessarily 1-coercive[6], which is needed to show (30) for all points[7] $x \in \mathbb{R}^d$. To overcome this, we introduce a small Gaussian noise to the states $X_{i,t}$ and define, for each t, $Y_{i,t} = X_{i,t} + V/\sqrt{Mt}$, where V has the standard multivariate Gaussian distribution $\mathcal{N}(0_d, I_d)$, and, we assume, is independent of $Z_{j,t}$ and W_t, for all j and t (hence, V is independent of $X_{i,t}$, for all t). The parameter $M > 0$ controls the magnitude of the noise, and the factor $1/\sqrt{t}$ adjusts the noise variance to the same level of the variance of $X_{i,t}$.

For each fixed M, let $\Lambda_{t,M}$ denote the log-moment generating function associated with the corresponding $Y_{i,t}$, under the conditioning on $W_1, \ldots, W_t$. It can be shown, using the independence of V and $X_{i,t}$, that

$$\Lambda_{t,M}(\lambda) = \Lambda_t(\lambda) + \frac{\|\lambda\|^2}{2M},\ \lambda \in \mathbb{R}^d. \tag{33}$$

Hence, the noise adds a (strictly) quadratic function to Λ_t, thus making $\Lambda_{t,M}$ 1-coercive, as proved in the following lemma. Lemma 8 gives the properties of $\Lambda_{t,M}$ that we use in the sequel; the proof is given in the Appendix.

Lemma 8:

1) Function $\Lambda_{t,M}$ is strictly convex, differentiable, and 1-coercive. Thus, for any $x \in \mathbb{R}^d$, there exists $\eta_t = \eta_t(x)$ such that $\nabla\Lambda_{t,M}(\eta_t) = x$.
2) Let $\theta = \mathbb{E}[Z_{i,t}]$. For any x, the corresponding sequence $\eta_t, t = 1, 2, \ldots$, is uniformly bounded, i.e.,

$$\|\eta_t\| \leq M\|x - \theta\|, \text{ for all } t. \tag{34}$$

Using the results of Lemma 8, we prove in *Step 3* the counterpart of (30) for the sequence $Y_{i,t}$—(35), and in *Step 4* we complete the proof of (20) by showing that (30) (a sufficient condition for (20)) is implied by (35).

Step 3: We show that, for any fixed $x \in \mathbb{R}^d$, $M > 0$, and $\omega \in \Omega$,

$$\lim_{\delta\to 0} \liminf_{t\to+\infty} \frac{1}{t}\log \nu_{t,M}\left(B_x(\delta)\right) \geq -NI(x). \tag{35}$$

where $\nu_{t,M}$ is the conditional probability measure induced by $Y_{i,t}$, $\nu_{t,M}(D) := \mathbb{P}\left(Y_{i,t} \in D | W_1, \ldots, W_t\right), D \in \mathcal{B}(\mathbb{R}^d)$.

To this end, fix arbitrary $x \in \mathbb{R}^d$, $\delta > 0$, $M > 0$, and $\omega \in \Omega$. We prove (35) by the change of measure argument. For any $t \geq 1$, we use the point η_t from Lemma 8 to change the measure on $\mathbb{R}^d$ from $\nu_{t,M}$ to $\widetilde{\nu}_{t,M}$ by:

$$\frac{d\widetilde{\nu}_{t,M}}{d\nu_{t,M}}(z) = e^{\,t\,\eta_t^\top z - t\,\Lambda_{t,M}(\eta_t)},\ z \in \mathbb{R}^d. \tag{36}$$

Note that, in contrast with the standard method of Gärtner-Ellis Theorem where the change of measure is fixed (once x is given), here we have a different change of measure[8,9] for each t. Expressing the probability $\nu_{t,M}\left(B_x(\delta)\right)$ through $\widetilde{\nu}_{t,M}$, for each t, we get:

$$\begin{aligned}
&\frac{1}{t}\log \nu_{t,M}\left(B_x(\delta)\right) \\
&\quad = \Lambda_{t,M}(\eta_t) - \eta_t^\top x \\
&\qquad + \frac{1}{t}\log \int_{z\in B_x(\delta)} e^{t\eta_t^\top (x-z)} d\widetilde{\nu}_{t,M}(z) \\
&\quad \geq \Lambda_{t,M}(\eta_t) - \eta_t^\top x - \delta\|\eta_t\| + \frac{1}{t}\log \widetilde{\nu}_{t,M}\left(B_x(\delta)\right).
\end{aligned} \tag{37}$$

[6]A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is 1-coercive if $\lim_{\|x\|\to+\infty} \frac{f(x)}{\|x\|} = +\infty$, e.g., [38]. A notable property of 1-coercive functions, which we exploit in this paper, is that their conjugates are finite at all points.

[7]More precisely, the problem arises when x is not an exposed point of the conjugate I_t of Λ_t, as will be clear from later parts of the proof (see also Exercise 2.3.20 in [28]).

[8]The reason for this alteration of the standard method is the fact that our sequence of functions $\Lambda_{t,M}$ does not have a limit.

[9]It can be shown that all distributions $\widetilde{\nu}_{t,M}, t \geq 1$, have the same expected value x; we do not pursue this result here, as it is not crucial for our goals.

We analyze separately each of the terms in (37). First, since η_t is uniformly bounded, by Lemma 8, we immediately obtain that the third term vanishes:

$$\lim_{\delta\to+0}\liminf_{t\to+\infty} -\delta\,\|\eta_t\| \geq -\lim_{\delta\to 0}\delta M\|x-\theta\| = 0. \qquad (38)$$

We consider next the sum of the first two terms. Let $I_{t,M}$ denote the conjugate of $\Lambda_{t,M}$. By Lemma 8, we have that η_t is the maximizer of $\lambda \mapsto \lambda^\top x - \Lambda_{t,M}(\lambda)$. Thus, the sum of the first two terms in (37) equals $-I_{t,M}(x) = \Lambda_{t,M}(\eta_t) - \eta_t^\top x$. Further, starting from the fact that $\Lambda_{t,M} \geq \Lambda_t$ and then invoking Lemma 2 (lower bound), we obtain:

$$\begin{aligned} I_{t,M}(x) &\leq \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - \Lambda_t(\lambda) \leq \sup_{\lambda\in\mathbb{R}^d} \lambda^\top x - N\Lambda(\lambda/N) \\ &= NI(x), \end{aligned} \qquad (39)$$

which holds for all $t \geq 1$ and all $M > 0$. Comparing with (35), we see that it only remains to show that the lim inf as $t \to +\infty$ of the last term in (37) vanishes with δ.

It is easy to show that the log-moment generating function associated with $\widetilde{\nu}_{t,M}$ is $\widetilde{\Lambda}_{t,M} := \Lambda_{t,M}(\lambda+\eta_t) - \Lambda_{t,M}(\eta_t)$. Let $\widetilde{I}_{t,M}$ denote the conjugate of $\widetilde{\Lambda}_{t,M}$. Similarly as in the proof of the upper bound in Section VI.A, it can be shown that

$$\frac{1}{t}\log\widetilde{\nu}_{t,M}\left(B_x^c(\delta)\right) \leq -\inf_{w\in B_x^c(\delta)} \widetilde{I}_{t,M}(w). \qquad (40)$$

The next lemma asserts that the right-hand side of (40) is strictly negative[10], and uniformly bounded away from zero. The proof is given in the Appendix.

Lemma 9: For any t, there exists a minimizer $w_t = w_t(x,\delta)$ of the optimization problem $\inf_{w\in B_x^c(\delta)} \widetilde{I}_{t,M}(w)$. Furthermore, there exists $\xi = \xi(x,\delta) > 0$ such that

$$\widetilde{I}_{t,M}(w_t) \geq \xi, \quad \text{for all } t. \qquad (41)$$

Combining (40) and (41), we get

$$\widetilde{\nu}_{t,M}\left(B_x(\delta)\right) \geq 1 - e^{-\xi t}, \text{ for all } t.$$

which, together with the fact that $\widetilde{\nu}_{t,M}$ is a probability measure (and hence $\widetilde{\nu}_{t,M}\left(B_x(\delta)\right) \leq 1$), yields

$$\lim_{t\to+\infty}\frac{1}{t}\log\widetilde{\nu}_{t,M}\left(B_x(\delta)\right) = 0. \qquad (42)$$

Since (42) holds for all $\delta > 0$, we conclude that the last term in (37) vanishes after the appropriate limits have been taken. Summarizing (38), (39), and (42) finally proves (35).

Step 4*:* To complete the proof of (20), it only remains to show that (35) implies (30). Since $X_{i,t} = Y_{i,t} - V/\sqrt{tM}$, we have

$$\begin{aligned} &\mathbb{P}(X_{i,t}\in B_x(2\delta)|W_1,\ldots,W_t) \\ &\quad\geq \mathbb{P}(Y_{i,t}\in B_x(\delta), V/\sqrt{tM}\in B_x(\delta)|W_1,\ldots,W_t) \\ &\quad\geq \nu_{t,M}(B_x(\delta)) - \mathbb{P}(V/\sqrt{tM}\notin B_x(\delta)). \end{aligned} \qquad (43)$$

From (35), the rate for the probability of the first term in (43) is at most $NI(x)$. On the other hand, the probability that the norm of V is greater than $\sqrt{tM}\delta$ decays exponentially with t at the rate $M\delta^2/2$,

$$\lim_{t\to+\infty}\frac{1}{t}\log\mathbb{P}(V/\sqrt{tM}\in B_x(\delta)) = -\frac{M\delta^2}{2}. \qquad (44)$$

Observe now that, for any fixed δ, for all M large enough so that $NI(x) < \frac{M\delta^2}{2}$, the exponential decay of the difference in (43) is determined by the rate of the first term, $NI(x)$. This finally establishes (30), which combined with Lemma 7 proves (20).

VII. Simulation Results

This section presents our simulation results for the performance of algorithm (2) for both deterministic and random weight matrices. In the deterministic case, we present two solutions for designing the weight matrix A, one that builds on the optimized left eigenvector a (see ahead (45)), and the other based on a simple heuristic, that uses local information only (see ahead (46)). Simulations show that both the optimized system based on the knowledge of a and the heuristic solution significantly outperform the system with the uniform left Perron vector, that, asymptotically, weighs equally all the observations. When the number of estimated parameters d increases, the optimized system outperforms both the heuristic and the uniform solution, hence proving the benefit of network design. We then consider randomly time-varying weight matrices and verify by simulations Theorem 2 for the following cases: 1) W_t are i.i.d. in time, with i.i.d. link failures; and 2) link failures of each link in the network, independently from other links, are governed by a Markov chain.

Simulation Setup: The number of nodes is $N = 10$. We form the communication graph by placing the nodes uniformly at random in a unit square and forming the links between the node pairs that lie within distance $r = 0.4$. The resulting graph $\widehat{G} = (V, \widehat{E})$ used in simulations is connected and also contains all the self-loops. For both deterministic and random cases, observations $Z_{i,t}$ are Gaussian, with equal expected values across all nodes. In the deterministic case, we consider $d = 1$ and $d = 3$. When $d = 1$, the expected value $m \in \mathbb{R}$ (equal at all nodes) is chosen uniformly at random from the $[0,1]$ interval, whereas the variances $\sigma_i^2, i = 1,\ldots,N$, are different across the nodes and are chosen uniformly at random, independently from each other, from $[0,1]$. When $d = 3$, each of the nodes has the same vector of expected values $M \in \mathbb{R}^d$, with the components chosen uniformly at random from $[0,1]$, and independently from each other. The covariance matrices $S_i, i = 1,\ldots,N$, are generated as follows: 1) for each i, we generate from the standard Gaussian distribution a 3 by 3 matrix B_i, form $B_i' = \frac{1}{2}(B_i + B_i^\top)$, and compute the matrix of eigenvectors Q_i of B_i'; 2) we then assign $S_i = 10\,Q_i Q_i^\top$. In the random case, we only consider the case of i.i.d. one-dimensional observations, where the mean value is m and the variance σ^2 is chosen uniformly at random from $[0,1]$.

In the case of one-dimensional observations, since all the nodes were assigned the same expected value m and because the left Perron vector a of A from Theorem 1 belongs to the simplex, it must be that $\widetilde{m} = m$. By Theorem 1, we therefore

[10]In the proof of the lower bound of the Gärtner-Ellis theorem, the sequence $\widetilde{\Lambda}_t$ (our $\widetilde{\Lambda}_{t,M}$) has a limit $\widetilde{\Lambda}$, and, because of this, it is sufficient to show that $\inf_{w\in B_x^c(\delta)} \widetilde{I}(w)$ is strictly negative, where $\widetilde{I}$ is the conjugate of $\widetilde{\Lambda}$. Here, since we do not have the limit of the $\widetilde{\Lambda}_{t,M}$'s, we need to prove that the latter holds for each function of the sequence $\widetilde{I}_{t,M}, t \geq 1$, and, moreover, that the strict negativity does not "fade out" with t.

obtain that all the states $X_{i,t}$ converge almost surely to m. The accuracy region that we target is $C = [m-\zeta, m+\zeta]$, where we set $\zeta = 0.035$. In the case when $d = 3$, it can similarly be shown by Theorem 1 that all the states converge to vector $M \in \mathbb{R}^d$, and the accuracy region that we tested in this case is $B_M(\zeta)$ (with the same $\zeta = 0.035$ as with $d = 1$).

A. Network Design for the Deterministic Case

In this section, we consider the problem of designing the (constant) weight matrix A. We present two heuristic design choices for the weight matrix A, and we show by simulations that they both perform well in practice.

Tuning A to $a_{\rm opt}$: We now present an optimization method for selecting matrix A in the Gaussian case, given the optimal Perron vector $a_{\rm opt}$ (obtained, e.g., by solving (16)). This method is, in a certain sense, a tractable substitute of (19). The method consists of the following: under the given communication constraints, we optimize A such that we force its left Perron vector to be as close as possible to the optimal Perron vector $a_{\rm opt}$ and, at the same time, achieve that the powers of A converge quickly. Motivated by this idea, we formulate the following optimization problem for finding $A = A_{\rm opt}$:

$$\begin{array}{ll} \text{minimize} & \left\|A - 1_N a_{\rm opt}^\top\right\| + \mu \left\|A^\top a_{\rm opt} - a_{\rm opt}\right\| \\ \text{subject to} & A 1_N = 1_N \\ & A \in \mathcal{A}_\epsilon \end{array} . \quad (45)$$

Here, $\mathcal{A}_\epsilon := \Big\{A \in \mathbb{R}_+^{N\times N} : A_{ij} = 0, \text{ if } (j,i) \notin \widehat{G}, i,j \in V, \text{ and } A_{ij} \geq \epsilon, \text{ otherwise}\Big\}$; $\mu > 0$ is a large penalty parameter (e.g., $\mu = 10{-}100$); and $\epsilon > 0$ is a small constant (e.g., $\epsilon = 10^{-6}{-}10^{-4}$). The second summand in the objective function of (45) forces A to tune its left Perron vector to $a_{\rm opt}$; the first term in the objective function is a tractable heuristic replacement for the non-convex function $|\lambda_2(A)|$ from (19), which one would ideally wish to optimize in order to achieve the fastest mixing of A. Finally, for any $\epsilon > 0$ and a strongly connected topology $\widehat{G}$, the constraint $A \in \mathcal{A}_\epsilon$ ensures that any solution of (45) has the second largest eigenvalue strictly smaller than 1 in modulus. For the simulation results shown in Fig. 2, we solved (16) and (45) via CVX [39], [40], where in (45) we used $\epsilon = 10^{-5}$ and $\mu = 100$. We remark that, due to the quadratic increase of complexity of (45) in N, for larger networks one might need to consider more scalable solutions than CVX, such as projected subgradient methods or alternating direction method of multipliers (ADMM), in order to solve (45).

Heuristic Based on Sensors' Relative Variances: To address the design of A when the dimension of the network is very large, such that solving (16) and (45) is impractical, we provide a simple but efficient heuristic. The heuristic is based on the sensor variances of the node's local neighborhoods. In particular, the heuristic solution $A = A_{\rm loc}$ is achieved by the following weight assignment:

$$A_{{\rm loc},ij} = \begin{cases} \frac{\frac{1}{\mathrm{tr}(S_j)}}{\sum_{l\in O_i}\frac{1}{\mathrm{tr}(S_l)}}, & \text{if } j \in O_i \\ 0, & \text{otherwise} \end{cases} . \quad (46)$$

That is, for each pair of nodes (j,i) such that i can receive messages from j, the corresponding weight A_{ij} is set to be the relative inverse of the trace of the covariance matrix of node j, where the scaling is taken with respect to the neighborhood of j, and takes value 0 otherwise.

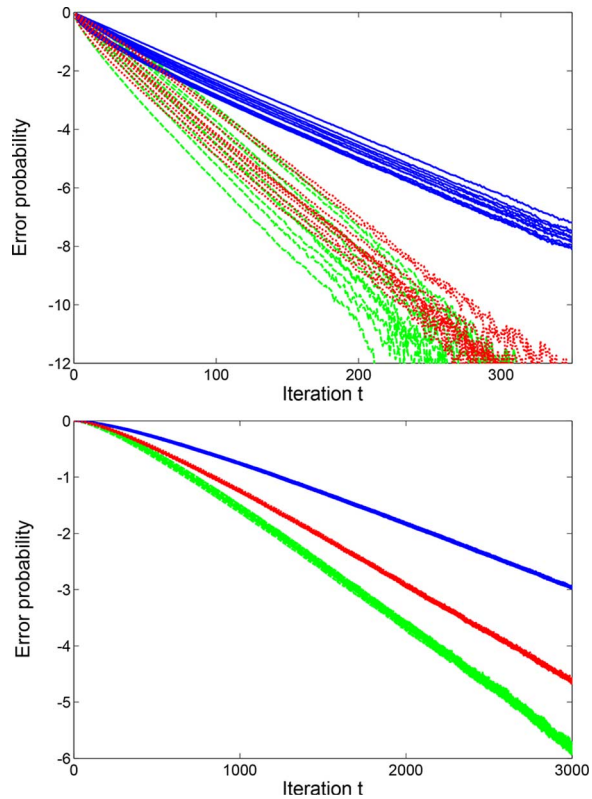


Fig. 2. Estimated error probabilities $\widehat{P}_{i,t}$ vs. number of iterations t, for each i, for the deterministic model. Green dashed lines correspond to $A_{\rm opt}$; red dotted lines to $A_{\rm loc}$; and blue full lines to $A_{\rm unif}$. Top: $d = 1$. Bottom: $d = 3$.

Fastest Averaging With Uniform Perron Vector: For the purpose of comparison, we present a solution where we seek the fastest possible averaging as in (45), but now with $a_{\rm opt}$ replaced by $a_{\rm unif} = (1/N)1_N$—the vector which is oblivious to the differences in the distributions (i.e., qualities) of different sensors. We denote the corresponding solution by $A_{\rm unif}$.

We compare the performance of distributed inference algorithm (2) running with, respectively, $A_{\rm opt}$, $A_{\rm loc}$, and $A_{\rm unif}$. For the case $d = 1$, at each node i and each time t, we estimate the probability of error $\widehat{P}_{i,t}$, by Monte Carlo simulations: we count the number of times that the state of node i at time t, $X_{i,t}$, falls outside of the accuracy region $C = [m-\zeta, m+\zeta]$, and then we divide this number by the number of simulation runs $K = 1000000$, $\widehat{P}_{i,t} = \frac{1}{K}\sum_{k=1}^{K} 1\left\{\left|X_{i,t}^k - m\right| \geq \zeta\right\}$. Similarly, in the case $d = 3$, we estimate the error probabilities by $\widehat{P}_{i,t} = \frac{1}{K'}\sum_{k=1}^{K'} 1\left\{\left\|X_{i,t}^k - M\right\| \geq \zeta\right\}$, where we used $K' = 100000$.

The plots in Fig. 2 show the evolution of the error probability over iterations, in the log-scale (we take the natural logarithm), for each node i. The top figure corresponds to the described setup for $d = 1$, and the bottom figure to the described setup for $d = 3$. In both plots, green dashed lines correspond to $A_{\rm opt}$, dotted red lines correspond to $A_{\rm loc}$, while blue full lines correspond to $A_{\rm unif}$. We can see from Fig. 2 (top) that for each of the three systems, $A_{\rm opt}$, $A_{\rm loc}$, and $A_{\rm unif}$, the curves at all nodes have the same slope, equal to the value of the corresponding rate

function over the set C. For the same weight matrix, the vertical shift in different curves (that correspond to different nodes) is due to the difference in the observations parameters (intuitively, nodes with higher variances σ_i^2 need more time to filter out the noise—and thus their error probability curves are shifted upwards), and the placement in the network (nodes with more central location in the network converge faster). We can see that the system with matrix $A_{\rm opt}$ achieves much higher large deviations rate than the one with the uniform eigenvector, as predicted by our theory. For example, for the target error probability of $e^{-5} \approx 0.007$, the optimized system needs around 140 iterations on average (across nodes), while the system with the uniform vector a requires almost as twice as much iterations for the same accuracy. The reason for this behavior is quite intuitive: optimizing the vector a corresponds to choosing different weights for different sensors depending on their local variances (i.e., covariance matrices, when $d > 1$). What is also interesting is that the system with the heuristic matrix $A_{\rm loc}$, obtained through the simple, local variances based rule, performs almost as equally well as the system with the optimized matrix $A_{\rm opt}$, which requires optimization at the network scale. We also report that, in certain cases, the heuristic even beats the optimized system in the finite regime. This happens when the left Perron vector of $A_{\rm loc}$ is close to $a_{\rm opt}$ and $A_{\rm loc}$ has faster convergence than $A_{\rm opt}$ (smaller second in modulus eigenvalue). However, when we move to higher dimensions, the difference between the optimized system and the heuristic one starts to show, as can be seen from Fig. 2 (bottom), which shows the simulation results for $d = 3$. This is to be expected, as the heuristic cannot "see" (and thus account for) the correlations in the observations of different parameters.

B. Random Weight Matrices

This subsection considers random weight matrices W_t for two cases: i.i.d. link failures and Markov chain link failures. With the i.i.d. model, each directed link $(i, j) \in \widehat{E}$ can fail with probability $1 - p$ at any given time t; this occurs independently from other link failures and independently from past times. With the Markov chain model, each link $(i, j) \in \widehat{G}$ behaves as a Markov chain, independent from the Markov chains of other links, such that with probability q_1 the link stays online, if it was online in the previous time slot, and with probability q_2 stays offline. (For example, if at time t a link is online, then at time $t+1$ this link stays online with probability q_1 and fails with probability $1-q_1$). With both i.i.d. and the Markov chain model, the weight matrix at time t equals $W_t = I_N - \alpha L_t$, where L_t is the Laplacian of the (directed) topology realization at time t, $\alpha = 1/(d_{\max} + 1)$, and $d_{\max}$ is the maximal degree in $\widehat{G}$.

The top and the bottom plot in Fig. 3 show the estimated error probabilities versus the number of iterations for both the i.i.d. and the Markov chain model, for two different sets of parameters: $p = 0.1, q_1 = q_2 = 0.3$ (top) and $p = 0.5, q_1 = 0.7, q_2 = 0.1$ (bottom). The error probability curves shown are computed by Monte Carlo, similarly as in the deterministic $d = 1$ case, based on $K'' = 5\ 000\ 000$ Monte Carlo runs. Both simulations are obtained for the same value of accuracy $\zeta = 0.1$, and one-dimensional Gaussian observations with parameters m and σ^2 chosen uniformly at random from the $[0, 1]$ interval. The results for the i.i.d. model are plotted in dashed lines, while the results for the Markov chain model are plotted in dotted lines. For reference, we also plot the estimated error probabilities for perfect fusion and isolation (full lines), see Section III.B; the lower curve corresponds to fusion. We can see from the plots that, under both models, the rate at which the error probability at each node decays is between the respective decay rates of the isolated node and the fusion center curves, as predicted by Theorem 2. We can also see that the agents' decay rates for the Markov chain model are faster than the ones for the i.i.d. model. This is expected since, for both sets of parameters, links in the i.i.d. model are online less frequently than the links in the Markov chain model, once the system reaches a stationary regime. Also, we see that improvements in the system parameters (higher p, in the i.i.d. model, and higher q_1 and lower q_2 in the Markov chain model) significantly affect the large deviations rates: in the bottom plot, the rates at each node got closer to the optimal, fusion center rate.

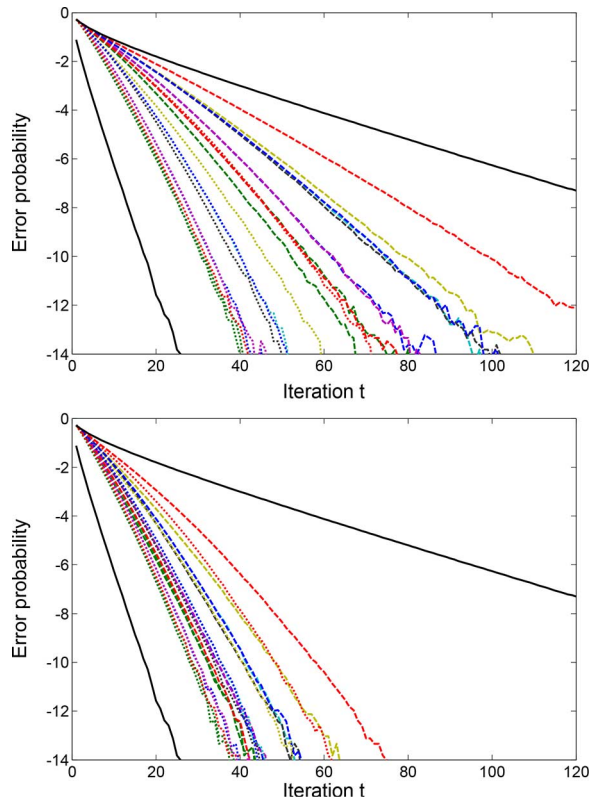


Fig. 3. Estimated error probabilities $\widehat{P}_{i,t}$ vs. number of iterations t, for each i, for the random model. Dashed curves correspond to the i.i.d. model, dotted curves to the Markov chain model, and full curves to an isolated node (upper) and the fusion center (lower). Top: $p = 0.1, q_1 = q_2 = 0.3$. Bottom: $p = 0.5, q_1 = 0.7, q_2 = 0.1$.

VIII. Conclusion

We studied large deviations rates for consensus based distributed inference, for deterministic and random asymmetric weight matrices. For the deterministic case, we characterized the corresponding large deviations rate function, and we showed that it depends on the weight matrix only through its left eigenvector that corresponds to its unit eigenvalue. When the observations are Gaussian (not necessarily identically distributed

across agents), the rate function has a closed form expression. Motivated by these insights, we formulate the optimal weight matrix design problem and show that, in the Gaussian case, it can be formulated as an SDP and hence efficiently solved. When the weight matrices are random, we prove that the rate functions of any node in the network lie between the rate functions corresponding to a fusion node, that processes all observations, and a node in isolation. The bounds hold for any random model of weight matrices, with the single condition that the weight matrices are independent from the agents' observations.

APPENDIX

A. Proof of the Almost Sure Convergence of $X_{i,t}$

We break the sum in (3) according to the limiting values of the elements of A^{t-s}:

$$X_{i,t} = \frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{N} a_j Z_{j,s} + \frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{N}\left([A^{t-s}]_{ij} - a_j\right) Z_{j,s}. \tag{47}$$

By the strong law of large numbers [37] applied to the sequence of i.i.d. random vectors $\sum_{j=1}^{N} a_j Z_{j,s}$, we know that the first term on the right hand side of (47) converges almost surely to $\sum_{j=1}^{N} a_j m_j$. Hence, to complete the proof of the claim, it remains to show that the second term of (47) converges almost surely to 0. We do this by showing that, for every $j = 1, \ldots, N$,

$$S_{j,t} := \frac{1}{t}\sum_{s=1}^{t}\left([A^{t-s}]_{ij} - a_j\right) Z_{j,s} \xrightarrow{\text{a.s.}} 0. \tag{48}$$

Fix j, and, to ease the exposition, denote $\zeta(t,s) := [A^{t-s}]_{ij} - a_j$. Note that, since $A1 = 1$ and $a^\top A = a^\top$, there holds that $A^{t-s} - 1a^\top = (A - 1a^\top)^{t-s}$, for any $t \geq s \geq 1$. Recalling that $|\lambda_2(A)| < 1$, it can be shown that there exists $\varsigma \in (|\lambda_2(A)|, 1)$ and a constant $K = K(A, \varsigma)$ such that, for all t, s such that $t \geq s \geq 1$:

$$|\zeta(t,s)| \leq K\varsigma^{t-s} \tag{49}$$

(see, e.g., Corollary 5.6.13 in [34]). Having (49), the proof of (48) follows the standard path. By Chebyshev's inequality [37], for arbitrary $\epsilon > 0$, the following holds:

$$\begin{aligned}\mathbb{P}\left(\|S_{j,t}\| \geq \epsilon\right) &\leq \frac{1}{\epsilon^2}\mathbb{E}\left[\|S_{j,t}\|^2\right] \\ &= \frac{1}{\epsilon^2 t^2}\mathbb{E}\left[\left(\sum_{s=1}^{t}\zeta(t,s)Z_{j,s}\right)^\top\left(\sum_{r=1}^{t}\zeta(t,r)Z_{j,r}\right)\right] \\ &= \frac{1}{\epsilon^2 t^2}\sum_{s=1}^{t}\zeta(t,s)^2 C_1 + \frac{2}{\epsilon^2 t^2}\sum_{s=1}^{t}\sum_{r=s+1}^{t}\zeta(t,s)\zeta(t,r)C_2,\end{aligned} \tag{50}$$

where $C_1 = \mathbb{E}\left[Z_{j,s}^\top Z_{j,s}\right]$, and $C_2 = m_j^\top m_j$. Using part 1c of Lemma 1 together with Assumption 2, yields that both the expected value $m_j = \nabla\Lambda_j(0)$ and the covariance matrix $\mathbb{E}\left[Z_{j,s}Z_{j,s}^\top\right] = \nabla^2\Lambda_j(0)$ are finite—hence, C_1 and C_2 are finite (and also non-negative). Considering separately the sums in the left and the right hand side of (50), and applying (49) to each of the terms $\zeta(t,s)$, we obtain:

$$\sum_{s=1}^{t}\zeta(t,s)^2 \leq \sum_{s=1}^{t}K^2\varsigma^{2(t-s)} \leq K^2\frac{1}{1-\varsigma^2}, \tag{51}$$

and

$$\sum_{s=1}^{t}\sum_{r=s+1}^{t}\zeta(t,s)\zeta(t,r) \leq K^2\frac{1}{(1-\varsigma)^2}, \tag{52}$$

which when combined in (50) yields

$$\mathbb{P}\left(\|S_{j,t}\| \geq \epsilon\right) \leq \frac{1}{t^2}\frac{K^2}{\epsilon^2}\left(\frac{C_1}{1-\varsigma^2} + \frac{C_2}{(1-\varsigma)^2}\right). \tag{53}$$

From (53) we obtain that, for any fixed $\epsilon > 0$, $\sum_{t=1}^{+\infty}\mathbb{P}\left(\|S_{j,t}\| \geq \epsilon\right) < +\infty$. This by the Borel-Cantelli lemma [37] implies $\mathbb{P}\left(\|S_{j,t}\| \geq \epsilon, \text{ inf. often}\right) = 0$, finally proving (48). Since j was arbitrary, this completes the proof of almost sure convergence of the $X_{i,t}$'s.

B. Proof of Lemma 7

Fix a measurable set D. We first show that if (30) holds for any $x \in D^{\text{o}}$ and any $\omega \in \Omega$, then for any $x \in D^{\text{o}}$

$$\lim_{\delta\to 0}\liminf_{t\to+\infty}\frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in B_x(\delta)\right) \geq -NI(x). \tag{54}$$

To this end, fix $x \in D^{\text{o}}$ and fix $\omega \in \Omega$.

Applying Fatou's lemma [37] to the sequence of random variables $R_t := \frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in D|W_1, \ldots, W_t\right), t = 1, 2, \ldots$, we get

$$\liminf_{t\to+\infty}\mathbb{E}\left[\frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in D\|W_1, \ldots, W_t\right)\right] \geq \mathbb{E}[R^\star], \tag{55}$$

where $R^\star(\omega) := \liminf_{t\to+\infty} R_t(\omega), \omega \in \Omega$. Consider the left-hand side of (55). By linearity of the expectation and concavity of the logarithmic function, we have

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in D|W_1, \ldots, W_t\right)\right] \\ &\quad\leq \frac{1}{t}\log\mathbb{E}\left[\mathbb{P}\left(X_{i,t} \in D|W_1, \ldots, W_t\right)\right] \\ &\quad= \frac{1}{t}\log\mathbb{P}(X_{i,t} \in D).\end{aligned}$$

Taking the lim inf as $t \to +\infty$ on both sides of the preceding inequality and combining the result with (55), yields:

$$\liminf_{t\to+\infty}\frac{1}{t}\log\mathbb{P}(X_{i,t} \in D) \geq \mathbb{E}[R^\star]. \tag{56}$$

We now focus on the random variable R_t. Note that we assumed that D^{o} is nonempty (if the interior of D is empty, the lower bound (20) holds trivially). Since D^{o} is open, for any $x \in D^{\text{o}}$, we can find a small neighborhood $B_x(\delta_0)$ that is fully contained in D^{o} (where, we note, $\delta_0 = \delta_0(x)$). Hence, for all $\delta \leq \delta_0$, we have $B_x(\delta) \subseteq D^{\text{o}} \subseteq D$, and thus, for any fixed $\omega \in \Omega$

$$R_t \geq \frac{1}{t}\log\mathbb{P}\left(X_{i,t} \in B_x(\delta)|W_1, \ldots, W_t\right) \tag{57}$$

(we used here that the logarithmic function is non-decreasing). Since (57) holds for all t and all δ sufficiently small, taking the corresponding limits yields

$$R^\star \geq \lim_{\delta\to 0}\liminf_{t\to+\infty}\frac{1}{t}\log\mathbb{P}\left(X_{i,t}\in B_x(\delta)|W_1,\ldots,W_t\right).$$

Using now the assumption (30) of the lemma to bound the right-hand side of the preceding inequality, we obtain $R^\star \geq -NI(x)$, which, we note, holds for every point x in $D^{\rm o}$. Taking the supremum over all $x\in D^{\rm o}$, we obtain that for every $\omega\in\Omega$,

$$R^\star \geq -\inf_{x\in D^{\rm o}} NI(x). \tag{58}$$

Taking the expectation in the left-hand side, and combining with (56), we finally obtain the lower bound (20):

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

Since D was arbitrary, the claim of Lemma 7 is proven.

C. Proof of Lemma 8

Being the sum of Λ_t and a (convex) quadratic function, $\Lambda_{t,M}$ inherits convexity and differentiability from Λ_t; $\Lambda_{t,M}$ is strictly convex due to strict convexity of $\|\lambda\|^2/(2M)$. To prove 1-coercivity, by convexity of Λ_t, we have that $\Lambda_t(\lambda)\geq\lambda^\top\theta$. Hence,

$$\Lambda_{t,M}(\lambda)\geq\lambda^\top\theta+\frac{\|\lambda\|^2}{2M}.$$

Dividing both sides by $\|\lambda\|$ and using in the right hand side that $\lambda^\top\theta\geq-\|\lambda\|\|\theta\|$, we obtain

$$\frac{\Lambda_{t,M}(\lambda)}{\|\lambda\|}\geq-\|\theta\|+\frac{\|\lambda\|}{2M}\to+\infty,$$

when $\|\lambda\|\to+\infty$, proving that $\Lambda_{t,M}$ is 1-coercive. Strict convexity, differentiability, and 1-coercivity of $\Lambda_{t,M}$ imply that the gradient map $\nabla\Lambda_{t,M}$ is a bijection, see, e.g., Corollary 4.1.3 in [38, p. 239]. This proves part 1.

We now prove part 2. Fix x and fix $t\geq 1$. Note that η_t is the maximizer in $I_{t,M}(x)=\sup_{\lambda\in\mathbb{R}^d}\lambda^\top x-\Lambda_{t,M}(\lambda)$, and thus it holds that $I_{t,M}(x)=\eta_t^\top x-\Lambda_{t,M}(\eta_t)$. Since Λ_t is convex (and differentiable), its gradient map is monotone. Hence,

$$\left(\nabla\Lambda_t(\eta_t)-\nabla\Lambda_t(0)\right)^\top(\eta_t-0)\geq 0. \tag{59}$$

We next show that the value of the gradient of Λ_t at 0 equals θ. From (32), we have

$$\nabla\Lambda_t(\lambda)=\frac{1}{t}\sum_{s=1}^{t}\sum_{j=1}^{N}[\Phi(t,s)]_{ij}\Lambda([\Phi(t,s)]_{ij}\lambda). \tag{60}$$

The gradient of Λ at $\lambda=0$ equals θ, see Lemma 1. Using the fact that, for each fixed s, $\sum_{j=1}^{N}[\Phi(t,s)]_{ij}=1$, we obtain that $\nabla\Lambda_t(0)=\theta$. Thus, from (59) we have

$$\left(\nabla\Lambda_t(\eta_t)-\theta\right)^\top\eta_t\geq 0. \tag{61}$$

Now, note from (33) that $\nabla\Lambda_t(\lambda)=\nabla\Lambda_{t,M}(\lambda)-\lambda/M$, for arbitrary λ. Using now the fact $\nabla\Lambda_{t,M}(\lambda)=x$, (61) implies $(x-1/M\eta_t-\theta)^\top\eta_t\geq 0$. Thus, $(x-\theta)^\top\eta_t\geq\eta_t^\top\eta_t/M$, and using further the fact that $\|x-\theta\|\|\eta_t\|\geq(x-\theta)^\top\eta_t$ proves the claim of the lemma for this fixed x. Since x was arbitrary, the proof of the lemma is complete.

D. Proof of Lemma 9

From the fact that $\mathcal{D}_{\widetilde{\Lambda}_{t,M}}=\mathbb{R}^d$, one can show that $\widetilde{I}_{t,M}$ has compact level sets (note that $\widetilde{I}_{t,M}$ is lower semi-continuous). Thus, the infimum in (40) has a solution. Denote this solution by w_t and let ζ_t denote a point for which $w_t=\nabla\widetilde{\Lambda}_{t,M}(\zeta_t)\,(=\nabla\Lambda_{t,M}(\zeta_t+\eta_t))$ (such a point exists by Lemma 8). We now show that $\|w_t\|$ is uniformly bounded for all t, which, combined with part 2 of Lemma 8, in turn implies that $\eta_t+\zeta_t$ is uniformly bounded.

Lemma 10: For any fixed $\delta>0$ and $M>0$, there exists $R=R(x,\delta,M)<+\infty$ such that for all t:

1) $\|w_t\|\leq R$, and
2) $\|\zeta_t+\eta_t\|\leq M(R+\|\theta\|)$.

Proof: Fix $M>0,\delta>0$. Define $\overline{f}_M,\underline{f}_M:\mathbb{R}^d\mapsto\mathbb{R}$ as: $\overline{f}_M(z)=\sup_{\lambda\in\mathbb{R}^d}\lambda^\top z-N\Lambda\left(1/N\lambda\right)-\frac{\|\lambda\|^2}{2M}$, $\underline{f}_M(z)=\sup_{\lambda\in\mathbb{R}^d}\lambda^\top z-\Lambda(\lambda)-\frac{\|\lambda\|^2}{2M}$, for $z\in\mathbb{R}^d$. Note that both $\overline{f}_M,\underline{f}_M$ are lower semi-continuous, finite for every z, and have compact level sets. Let $c=\inf_{z\in B_x^{\rm c}(\delta)}\overline{f}_M(z)<+\infty$, and define $S_c=\{z\in\mathbb{R}^d:\underline{f}_M(z)\leq c\}$.

Fix arbitrary $t\geq 1$. One can show, with the help of Lemma 2, that, for any $z\in\mathbb{R}^d$,

$$\underline{f}_M(z)\leq I_{t,M}(z)\leq\overline{f}_M(z). \tag{62}$$

Observe now that $I_{t,M}(w_t)=\inf_{z\in B_x^{\rm c}(\delta)}I_{t,M}(z)\leq\inf_{z\in B_x^{\rm c}(\delta)}f_M(z)\leq c$. On the other hand, taking in (62) $z=w_t$, yields $\underline{f}_M(w_t)\leq I_{t,M}(w_t)$, and it thus follows that w_t belongs to S_c.

Finally, as S_c is compact, we can find a ball of some radius $R=R(x,M,\delta)>0$ that covers S_c, implying $w_t\in B_0(R)$. Since t was arbitrary, the claim in part 1 follows.

We now prove part 2. Recall that, for any t, w_t and $\zeta_t+\eta_t$ satisfy $w_t=\nabla\Lambda_{t_0,M}\left(\zeta_t+\eta_t\right)$. Applying part 2 of Lemma 8 for $z=w_t$, we have that $\|\zeta_t+\eta_t\|\leq M\|w_t-\theta\|$. Combining this with part 1 of this lemma yields

$$\|\zeta_t+\eta_t\|\leq M\|w_t-\theta\|\leq M\sup_{w\in B_0(R)}\|w-\theta\|\leq M(R+\|\theta\|).$$

This completes the proof of part 2 and the proof of Lemma 10. □

Fix x,δ and M and define $r_1=M\|z-\theta\|,r_2=M(R+\|\theta\|)$, where R is the constant that verifies Lemma 10. Fix now $t\geq 1$ and recall that η_t,ζ_t, and w_t are chosen such that $x=\nabla\Lambda_{k,M}(\eta_t),\widetilde{I}_{t,M}(w_t)=\inf_{z\in B_x^{\rm c}(\delta)}I_{t,M}(z)$, and $w_t=\nabla\Lambda_{t,M}(\eta_t+\zeta_t)$. By part 2 of Lemma 8 and part 2 of Lemma 10 we have for η_t and ζ_t, $\|\eta_t\|\leq r_1,\|\eta_t+\zeta_t\|\leq r_2$. To prove Lemma 9, we first show that there exists some positive constant r_3, independent of t, such that $\|\zeta_t\|\geq r_3$ for all t. To this end, consider the gradient map $\lambda\mapsto\nabla\Lambda_{t,M}(\lambda)$, and note that $\nabla\Lambda_{t,M}$ is continuous, and hence uniformly continuous on every compact set. Note also that $\|\eta_t\|,\|\eta_t+\zeta_t\|\leq\max\{r_1,r_2\}$; that is, points η_t and $\eta_t+\zeta_t$ are uniformly bounded for all t.

Suppose now, for the sake of contradiction, that for some sequence of times $t_k, k = 1, 2, \ldots, \|\zeta_{t_k}\| \to 0$, as $k \to +\infty$. Then, $\|(\eta_{t_k} + \zeta_{t_k}) - \eta_{t_k}\| \to 0$, and hence, by the uniform continuity of $\nabla\Lambda_{t,M}(\cdot)$ on $\overline{B}_0(\max\{r_1, r_2\})$ we have

$$\|\nabla\Lambda_{t,M}(\eta_{t_k}) - \nabla\Lambda_{t,M}(\eta_{t_k} + \zeta_{t_k})\| \to 0, \text{ as } t \to \infty.$$

Recalling that $x = \nabla\Lambda_{t,M}(\eta_{t_k}), w_{t_k} = \nabla\Lambda_{t,M}(\eta_{t_k})$, yields

$$\|w_{t_k} - x\| \to 0.$$

This contradicts with the fact that, for all $t, w_t \in B_x^c(\delta)$. Thus, we proved the existence of r_3 independent of t such that $\|\zeta_t\| \geq r_3$, for all t.

Now, let

$$\Upsilon = \left\{(\eta, \zeta) \in \mathbb{R}^d \times \mathbb{R}^d : \|\eta\| \leq r_1, \|\eta + \zeta\| \leq r_2, \|\zeta\| \geq r_3\right\},$$

and introduce $g : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$,

$$g(\zeta, \eta) = \Lambda_{t,M}(\eta) - \Lambda_{t,M}(\zeta + \eta) + \nabla\Lambda_{t,M}(\zeta + \eta)^\top \zeta. \quad (63)$$

By strict convexity of $\Lambda_{t,M}$, we see that, for any η and $\zeta \neq 0$, the value $g(\eta, \zeta)$ is strictly positive. Further, note that since $\Lambda_{t,M}$ and $\nabla\Lambda_{t,M}$ are continuous, function g is also continuous. Consider now

$$\xi := \inf_{(\eta,\zeta)\in\Upsilon} g(\eta, \zeta). \quad (64)$$

Because Υ is compact, by the Weierstrass theorem, the problem in (64) has a solution, that is, there exists $(\eta_0, \zeta_0) \in \Upsilon$, such that $g(\eta_0, \zeta_0) = \xi$. Finally, because g is strictly positive at each point in Υ (note that $\zeta \neq 0$ in Υ), we conclude that $\xi = g(\eta_0, \zeta_0) > 0$. Returning to the claim of Lemma 9, by Lemma 10, $(\eta_t, \eta_t + \zeta_t)$ belongs to Υ, and, thus,

$$\begin{aligned}\widetilde{I}_{t,M}(w_t) &= \Lambda_{t,M}(\eta_t) - \Lambda_{t,M}(\zeta_t + \eta_t) + \nabla\Lambda_{t,M}(\zeta_t + \eta_t)^\top \zeta_t \\ &= g(\eta_t, \zeta_t) \geq \xi.\end{aligned}$$

This completes the proof of Lemma 9.

REFERENCES

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Commun. Mag.*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[2] J. Abbott, Z. Nagy, F. Beyeler, and B. Nelson, "Robotics in the small, Part I: Microbotics," *IEEE Robot. Autom. Mag.*, vol. 14, no. 2, pp. 92–103, June 2007.

[3] I. F. Akyildiz and J. M. Jornet, "Electromagnetic wireless nanosensor networks," *Nano Commun. Netw.*, vol. 1, no. 1, pp. 3–19, 2010.

[4] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Distributed detection and estimation in wireless sensor networks," Jul. 2013 [Online]. Available: http://arxiv.org/abs/1307.1448

[5] N. E. Leonard and A. Olshevsky, "Cooperative learning in multiagent systems from intermittent measurements," *SIAM J. Control Optim.*, vol. 53, no. 1, pp. 1–29, 2015.

[6] M. Çetin, L. Chen, J. W. Fisher, III, A. T. Ihler, R. L. Moses, M. J. Wainwright, and A. S. Willsky, "Distributed fusion in sensor networks—A graphical models perspective," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 42–55, 2006.

[7] S. Stanković, M. S. Stanković, and D. M. Stipanović, "Consensus based overlapping decentralized estimator," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 410–415, Feb. 2009.

[8] P. Braca, S. Marano, and V. Matta, "Enforcing consensus while monitoring the environment in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3375–3380, Jul. 2008.

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

[10] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, Jun. 2012.

[11] R. Olfati-Saber, "Distributed Kalman filter with embedded consensus filters," in *Proc. 44th IEEE Conf. Decision Control/Eur. Control Conf. (CDC-ECC)*, Dec. 2005, pp. 8179–8184.

[12] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering based on consensus strategies," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 622–633, May 2008.

[13] 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 Trans. Signal Process.*, vol. 60, no. 11, pp. 5987–6002, Nov. 2012.

[14] S. S. Stanković, N. Ilić, M. S. Stanković, and K. H. Johansson, "Distributed change detection based on a consensus algorithm," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5686–5697, Dec. 2011.

[15] G. Mateos, I. D. Schizas, and G. B. Giannakis, "Performance analysis of the consensus-based distributed LMS algorithm," *EURASIP J. Adv. Signal Process*, vol. 68, Jan. 2009 [Online]. Available: http://dx.doi.org/10.1155/2009/981030

[16] P. Di Lorenzo and A. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, Mar. 2013.

[17] R. Rahman, M. Alanyali, and V. Saligrama, "Distributed tracking in multihop sensor networks with communication delays," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4656–4668, Sep. 2007.

[18] S. Kar and J. M. F. Moura, "Asymptotically efficient distributed estimation with exponential family statistics," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4811–4831, Aug. 2014.

[19] 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 Trans. Signal Process.*, vol. 61, no. 10, pp. 2557–2571, May 2013.

[20] 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 Trans. Inf. Theory*, vol. 61, no. 3, pp. 1351–1372, Mar. 2015.

[21] P. Braca, S. Marano, V. Matta, and A. H. Sayed, "Asymptotic performance of adaptive distributed detection over networks," Jan. 2014 [Online]. Available: http://arxiv.org/abs/1401.5742

[22] A. Lalitha, A. Sarwate, and T. Javidi, "Social learning and distributed hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2014, pp. 551–555.

[23] P.-N. Chen, "Generalization of Gärtner-Ellis theorem," *IEEE Trans. Inf. Theory*, vol. 46, no. 7, pp. 2752–2760, Nov. 2000.

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

[25] H. Chernoff, "A measure of the asymptotic efficiency of tests of a hypothesis based on a sum of observations," *Ann. Math. Statist.*, vol. 23, no. 4, pp. 493–507, Dec. 1952.

[26] M. Arcones, "Large deviations for M-estimators," *Ann. Inst. Statist. Math.*, vol. 58, no. 1, pp. 21–52, 2006.

[27] H. Cramér, "Sur un nouveau théorème-limite de la théorie des probabilités," (in French) *Actualités Scientifiques et Industrielles*, vol. 736, pp. 5–23, 1938, Paris.

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

[29] F. den Hollander, *Large Deviations*, ser. Fields Institute Monographs. Providence, RI, USA: Amer. Math. Soc., 2000.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[31] A. Tahbaz-Salehi and A. Jadbabaie, "A necessary and sufficient condition for consensus over random networks," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 791–795, Apr. 2008.

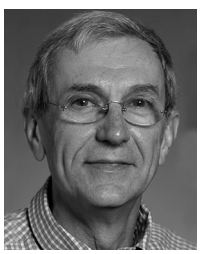
[32] E. Seneta, *Nonnegative Matrices and Markov Chains*. New York, NY, USA: Springer, 1981.

[33] S. Kirkland, "Subdominant eigenvalues for stochastic matrices with given column sums," *Electron. J. Linear Algebra*, vol. 18, pp. 784–800, 2009 [Online]. Available: http://eudml.org/doc/232271

[34] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[35] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Autom. Control*, vol. 56, no. 6, pp. 1291–1306, Jun. 2011.
[36] D. Bajović, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Distributed inference over directed networks: Performance limits and optimal design," Apr. 2015 [Online]. Available: http://arxiv.org/abs/1504.07526
[37] A. F. Karr, *Probability*. New York, NY, USA: Springer-Verlag, 1993.
[38] J.-B. Hiriart-Urruty and C. Lemarechal, *Fundamentals of Convex Analysis*, ser. ser. Grundlehren Text Editions. Berlin, Germany: Springer-Verlag, 2004.
[39] M. Grant and S. Boyd, CVX: Matlab Software for Disciplined Convex Programming, v. 2.1, Mar. 2014.
[40] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex programs," in *Recent Advances in Learning and Control*, ser. Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, Eds. New York, NY, USA: Springer-Verlag, 2008, pp. 95–110.

Dragana Bajović (M'13) received the Dipl.Ing. degree from the School of Electrical Engineering, University of Belgrade, Serbia, in August 2007, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, and Institute of Systems and Robotics (ISR), Instituto Superior Técnico (IST), Lisbon, Portugal, in May 2013. Currently, she is an Assistant Professor at the Department of Power, Electronics and Communication Engineering, University of Novi Sad, Serbia, and a researcher at the BioSense Institute, Novi Sad, Serbia. Her research interests include statistical signal processing and large deviations analysis for sensor networks.



José M. F. Moura (S'71–M'75–SM'90–F'94) received the engenheiro electrotécnico degree from Instituto Superior Técnico (IST), Lisbon, Portugal, and the M.Sc., E.E., and D.Sc. degrees in EECS from MIT, Cambridge, MA.

He is the Philip L. and Marsha Dowd University Professor at Carnegie Mellon University (CMU). He was on the faculty at IST and has held visiting faculty appointments at MIT. He founded and directs a large education and research program between CMU and Portugal, www.icti.cmu.edu.

His research interests are on statistical and algebraic signal and image processing, signal processing on graphs, and data science. He has published over 550 papers and holds thirteen patents issued by the US Patent Office. The technology of two of his patents (co-inventor A. Kavčić) are in about three billion disk drives read channel chips of 60% of all computers sold in the last 13 years worldwide and were, in 2016, the subject of the largest university verdict/settlement in the information technologies area.

Dr. Moura is the IEEE Technical Activities Vice-President (2016) and member of the IEEE Board of Directors. He served in several other capacities including IEEE Division IX Director, President of the IEEE Signal Processing Society(SPS), Editor in Chief for the IEEE TRANSACTIONS IN SIGNAL PROCESSING, interim Editor in Chief for the IEEE SIGNAL PROCESSING LETTERS.

Dr. Moura has received several awards, including the Technical Achievement Award and the Society Award from the IEEE Signal Processing Society. He is a Fellow of the IEEE, a Fellow of the American Association for the Advancement of Science (AAAS), a corresponding member of the Academy of Sciences of Portugal, Fellow of the U.S. National Academy of Inventors, and a member of the U.S. National Academy of Engineering.

João Xavier (S'97–M'03) received the Ph.D. degree in electrical and computer engineering from Instituto Superior Técnico (IST), Lisbon, Portugal, in 2002. Currently, he is an Assistant Professor in the Department of Electrical and Computer Engineering, IST. He is also a Researcher at the Institute of Systems and Robotics (ISR), Lisbon, Portugal. His current research interests are in the area of optimization and statistical inference for distributed systems.

Bruno Sinopoli (M'03) received the Dr. Eng. degree from the University of Padova in 1998 and his M.S. and Ph.D. in Electrical Engineering from the University of California at Berkeley, in 2003 and 2005 respectively. After a postdoctoral position at Stanford University, Dr. Sinopoli joined the faculty at Carnegie Mellon University where he is an associate professor in the Department of Electrical and Computer Engineering with courtesy appointments in Mechanical Engineering and in the Robotics Institute and co-director of the Smart Infrastructure Institute, a research center aimed at advancing innovation in the modeling analysis and design of smart infrastructure. Dr. Sinopoli was awarded the 2006 Eli Jury Award for outstanding research achievement in the areas of systems, communications, control and signal processing at U.C. Berkeley, the 2010 George Tallman Ladd Research Award from Carnegie Mellon University and the NSF Career award in 2010. His research interests include the modeling, analysis and design of Secure by Design Cyber-Physical Systems with application to interdependent infrastructures, Internet of Things and Data-driven Networking.