Large deviations for products of non identically distributed network matrices with applications to communication-efficient distributed learning and inference

Nemanja Petrović, Dragana Bajović, *Member, IEEE*, Soummya Kar, *Fellow, IEEE*, Dušan Jakovetić, *Member, IEEE*, Anit Kumar Sahu, *Member, IEEE*

Abstract

This paper studies products of independent but non-identically distributed random network matrices that arise as weight matrices in distributed consensus-type computation and inference procedures in peer-to-peer multi-agent networks. The non-identically distributed matrices studied in this paper model various application scenarios in which the agent communication network is time-varying, either naturally or engineered to achieve communication efficiency in computational procedures. First, under broad conditions on the statistics of the network matrix sequence, the product of the sequence is shown to converge almost surely to the consensus matrix and explicit large deviations rate of convergence are obtained. Specifically, given the admissible graph of interconnections modeling the base network topology, it is shown that the large deviations rate of consensus equals the minimum limiting value

N. Petrović is with the Department of Fundamental Sciences, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, Novi Sad, Serbia. Email: nemanjab4h@gmail.com.

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

S. Kar is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA. Email: soummyak@ece.cmu.edu.

D. Jakovetić is with the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia. Email: dusan.jakovetic@dmi.uns.ac.rs.

A. K. Sahu is with Amazon.com Inc, Alexa AI, USA. Email: anit.sahu@gmail.com

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

of the fluctuating graph cuts, where the edge costs are assigned through the current probabilities of the inter-agent communications. Secondly, an application of the above large deviations principle is studied in the context of distributed detection in time-varying networks with sequential observations. By adopting a *consensus+innovations* type distributed detection algorithm, as a by-product of this result, error exponents are obtained for the performance of distributed detection. It is shown that slow starts (slow increase) of inter-agent communication probabilities yield the same asymptotic error rate – and hence the same distributed detection performance, as if the communications were at their nominal levels from the beginning. As an important special case it is shown that when all the intermittent graph cuts have a link the probability of which increases to one, the performance of distributed detection is asymptotically optimal - i.e., equivalent to a centralized setup having access to all network data at all times.

Index Terms

Distributed inference, stochastic matrices, error exponents, inaccuracy rates, large deviations, consensus.

I. Introduction

Motivated by application domains dominated by wireless networking, communication efficiency has naturally become an increasingly prominent aspect of study in the area of distributed algorithms. Some exemplary fields concerned with this aspect include distributed inference, e.g., [1], [2], [3], distributed optimization, e.g., [4], [5], [6], distributed learning, e.g., [7], [8], [9], etc. The generic setup of study involves a group of entities, often called agents, enabled with computation and potentially also sensing capabilities, that collaborate to achieve a task at hand – find the global optimum from local objectives, discover the true hypothesis/parameter from local sensor measurements, or find the model that best describes the data held by different entities. To achieve the benefits resulting from collaboration, such as increased accuracy, without affecting adversely other performance metrics, communication between the agents should be performed efficiently.

Most common approaches for efficient communication are compression and sparse transmissions. With the former, the idea is to use efficient representations of the messages to be exchanged, e.g., in the sense of the numbers of bits transmitted; see, e.g., [7], [8], for distributed optimization, [10], [11], for distributed hypothesis testing (non-Bayesian social learning). Our work belongs to the body of works that adopt the second approach – sparse transmissions where the general idea is to let the agents iteratively perform local computations while selecting times for message exchanges sparsely. Similarly as with compressed transmission, this approach has also been extensively studied in the literature, e.g., in distributed optimization [4], [5], [12], distributed learning [13], [14], [15], distributed estimation [16], [17], distributed

hypothesis testing [18], [19], [20] distributed statistical inference [21] [22], social learning [23], [24], [25] etc.

An assumption that is often made in the mentioned lines of work is that the links over which the agents communicate are fully reliable. In this paper we study communication-efficient algorithms for distributed inference over networks with random links. The class of algorithms that we study is consensus+innovations algorithms [26], [27], [28] where merging of information occurs through DeGroot averaging [29], specifically, through a sequence of the so called weight matrices $W_{ij,t}$ defining the weight that agent i assigns to the opinion of agent j at time t (see eq. (6) further ahead). Motivating applications are distributed detection, estimation and learning in modern IoT systems or social networks, where interactions are naturally intermittent (e.g., in social networks, or edge computing due to the applied communication protocol such as gossip) and/or are prone to failures (e.g., in wireless sensor networks/IoT systems).

Many previous works studied performance of DeGroot-based distributed estimation, typically in the sense of mean square error (MSE), showing that the network effect for the MSE metric is captured by the spectral properties of either the expected weight matrix $\mathbb{E}[W_t]$ or $\mathbb{E}[W_t^2]$ [28]. These results stand in sharp contrast with the results in the literature on distributed detection, where the spectrum of $\mathbb{E}[W_t^2]$ (specifically, the second largest eigenvalue in modulus) provides only a loose bound for the error exponents. A critical property that this bound is not able to capture is that, in the deterministic case, the network effect plays no role in the error exponents - i.e., all error exponents across the network nodes are equal to the error exponent of a hypothetical fusion center. As shown in the previous works, the quantity that can properly capture the network effect and, in particular, "see" the latter property is the rate of convergence in probability of products $W_t \cdots W_1$:

$$\lim_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|W_t \cdots W_1 - J\| \geq \epsilon\right), \tag{1}$$

for an arbitrary $\epsilon \in (0, 1]$. The quantity in (1) is also referred to as the *large deviations rate of consensus*. Besides distributed detection, the preceding quantity is also applicable for distributed estimation, where instead of the MSE performance, one is concerned with the *inaccuracy rates* [30], [31].

Convergence of weighted averaging (linear consensus) over random networks has also been extensively studied in the past, for the case of independent and identically distributed (i.i.d.) weight matrices, e.g., [32], [33], [34] and also for non-i.i.d., stationary matrices, e.g., [35] (stationary and ergodic), [36] (Markov-chain based switching topologies), [37] (independent, cut-balanced and strongly aperiodic). Typically addressed convergence criteria are the mean-square and the almost sure convergence. In contrast with the referenced works, assuming non-stationary matrices, we study convergence of products of weight

matrices in the sense of large deviations, considering the sequence of events that the matrix products stay away from consensus, as in (1). If almost sure convergence were known *a priori*, the probabilities of such events would be guaranteed to converge to zero. However, the exact convergence rate (1) would not be possible to establish neither from the almost sure nor from the mean-square convergence. In this paper we find the exact rate of convergence of such events. As a special case, we show, by the Borel-Cantelli lemma, that whenever the rate (1) is non-zero, almost sure convergence of consensus is implied. However, the impact and the applicability of the rate (1) is much broader, as we show in this paper.

In terms of large deviations study of distributed inference, a closely related reference to this aspect of our work here is [19]. Reference [19] studies distributed hypothesis testing or social learning where, to decide on the right hypothesis, distributed entities over time incorporate new observations and collaborate with peers. This algorithm adopts a similar weighted averaging scheme of the nodes' iterates, called beliefs, updated by incorporating the newly arrived observations, through their (log-)likelihood values, and subsequently by weighted averaging of beliefs across nodes' local neighborhoods. Assuming deterministic networks, [19] shows that the nodes' beliefs converge to the 0/1 vector corresponding to the true hypothesis and also establish that the beliefs obey the large deviations principle. The respective large deviations rate is expressed through the rank-one limit of the weight matrix powers. Effectively, the same large deviations rate obtains if the process was throughout the iterations run with the weights replaced by their corresponding limits. This is due to the fact that the averaging process induced by the matrix powers converges at a much faster, geometric rate than the process incorporating new observations. This stands in sharp contrast with the case when the underlying topologies are random where such property no longer holds (e.g., for networks with randomly occurring links, it suffices to note that the products of weight matrices can remain block-disjoint for arbitrarily long periods of time, thus preventing network-wide dissemination of observations, and hence preventing the desired performance of the employed distributed inference algorithm). This cannot be seen from the typically used generalizations of the weight matrix such as $\mathbb{E}[W_t]$ or $\mathbb{E}[W_t^2]$. The metric that is able to capture the impact of communication randomness on large deviations performance of distributed inference is the exponent of the convergence in probability of consensus (1).

Problem statement and contributions. In this paper our goal is to derive communication-efficient distributed inference algorithms that are able to achieve asymptotic optimality of deterministic networks with much less communications. Specifically, we assume that each of the links has a time varying probability of occurrence converging to a certain number, where we make no assumptions on the speed of this convergence. Under this general setup, we compute the rate of consensus in (1) and show that

it is given by the minimum edge cut of the graph underlying the weight matrices, with edge weights defined through the limiting edge probabilities. This result extends and generalizes our previous results in [38],[18] in several important directions. In [38], we consider the setup where the statistics of link occurrences are constant over time, while in [18] we consider a setup where nodes, rather than edges, appear intermittently over time. Furthermore, [18] assumes monotonously increasing probabilities, the rate of which is assumed to be sufficiently large. Here, we consider randomness in the edge sense (thus subsuming also the vertex randomness), but we make no restrictions neither on the speed nor on the monotonicity of the convergence. In the technical sense, the current paper adopts novel techniques to cope with the generality of the link probability sequences, which, in particular, induce fluctuating minimum cuts over time. The fluctuating probabilities of edge cuts in the current paper incur additional technical challenges with respect to [38],[18], as they cannot be directly connected with the probability of the maximal asymptotic cut (i.e., the solution) that corresponds to the limiting probabilities – which are exploited both in [38] and [18]. This required novel analysis techniques provided here. Furthermore, in this work we allow for the probabilities of edges to converge to one, the case that is not supported by the technique adopted in [18]. Finally, as an important special case, we show that whenever the limiting graph is connected, the rate of consensus equals $+\infty$ and hence the network effect asymptotically vanishes yielding distributed inference equivalent to the centralized one. All theoretical results are illustrated and corroborated with numerical simulations.

Paper organization. Section II describes our model and gives preliminary results. Section III states the main results of the paper. Section IV presents motivating applications. Section V provides proofs of the results from Section III, while Section VI gives results of numerical simulations for the application examples from Section IV. Section VII concludes the paper.

Notation. We denote by: A_{ij} or $[A]_{ij}$ the entry in ith row and jth column of a matrix A; A_l and A^l the l-th row and column, respectively; I and $J := (1/N)11^\top$ the identity matrix, and the ideal consensus matrix, respectively; 1 and e_i the vector with unit entries, and ith canonical vector (with the ith entry equal to 1 and the rest being zeros), respectively. Further, for a vector a, the inequality $a > 0$ is understood component wise; $\log$ denotes the natural logarithm. We denote by $\mathcal{N}(m, \sigma^2)$ Gaussian distribution with mean m and standard deviation σ. By $\|\cdot\|$ we denote the spectral norm.

II. Model and preliminaries

Random links' activation and random matrix model. The network is modeled as an undirected graph $G = (V, E)$, where V is the set of nodes, and E is the set of communication links between nodes. We assume that G is connected. During network operation, communication links activate at random with

certain probabilities that we assume are different for different links. To each link $\{i,j\} \in E$ we associate, for each time $t = 1, 2, \ldots$, a Bernoulli random variable $\xi_{ij,t}$, which is equal to 1 if $\{i,j\}$ is active at time t, and otherwise equals 0. Let $p_{ij,t} = \mathbb{P}(\xi_{ij,t} = 1) \in (0,1)$ denote the probability that $\{i,j\}$ is active at time t and let E_t collect all the communication links in E active at time t. Let $G_t = (V, E_t)$, i.e. G_t is the subgraph of G collecting all the communication links that are active at time t.

We now state our assumptions on the network randomness and on the weight matrices W_t.

Assumption 1 (Communication links)**.**

1) *For any two links $\{i,j\}$, $\{k,l\} \in E$, $\{i,j\} \neq \{k,l\}$, and for any two time instants $t, s \geq 1$, $t \neq s$, the Bernoulli variables $\xi_{ij,t}$ and $\xi_{kl,s}$ are independent.*
2) *For each link $\{i,j\}$, for each t, $p_{ij,t} \in (0,1)$ and the respective sequence converges to $\overline{p}_{ij} \in (0,1]$, as $t \to +\infty$.*

It is easy to see from Assumption 1 that the topologies G_t, $t \geq 1$, are independent.

We associate with each graph G_t an $N \times N$ doubly stochastic matrix W_t. We make the following assumptions on the weight matrices W_t.

Assumption 2 (Weight matrices)**.**

1) *The weight matrices W_t, $t \geq 1$, are independent.*
2) *For each t, each realization of W_t is symmetric, stochastic and has positive diagonals, and it conforms to the structure of G_t, i.e., for each t, $[W_t]_{ij} = 0$ if and only if $\{i,j\} \notin E_t$, for $i \neq j$.*
3) *There exists $\delta > 0$ such that, for each t, $[W_t]_{ij} > \delta$ whenever $[W_t]_{ij} > 0$.*

We provide several examples of distributed learning scenarios where matrices W_t are symmetric. This occurs in any scenario (e.g., wireless sensor networks, multi-robot systems, etc.) where the communication between agents happens over an undirected, generic connected graph. For example, in wireless sensor networks, the graph is induced by the geographical proximity of agents, so that there is a link between a pair of agents within a distance less than a radius that relates with the protocol in question, transmit power, etc. The time-varying and random nature of graphs on top of such "backbone" graph arises, due to, e.g., movement of agents, or physical obstacles that may appear on a communication path between certain pairs of agents. Clearly, the distance and obstacle effects lead to symmetric communication links. Given an undirected graph, it is easy to construct a weight matrix that is symmetric on top of such graph. For example, one can use the Metropolis choice, or assign to all non-zero off-diagonal $W_{ij,t}$'s a pre-determined, fixed value, for example any lower bound on quantity $1/d_{\max}$ or $1/N$, where $d_{\max}$ is the maximal degree of the backbone graph. Also, on top of a "backbone" symmetric graph, one can utilize a

symmetric randomized protocol to construct randomized instances of W_t, such as, e.g., the randomized gossip protocol [39]. In addition, as explained in Section IV further ahead, a variant of federated learning with partial (or full) participation also leads symmetric matrices W_t's. We also mention that a random, symmetric model of W_t's has been extensively used in the literature (see Section IV), like with different instances of innovation + consensus algorithm where we assume that if communication exists from one agent to another, that communication back is also possible.

The requirement that the diagonal of W_t is strictly positive is essentially mild, as an agent can always associate a non-zero weight to itself. There is still a non-trivial requirement here, because $W_{ii,t}$ should also be positive (negative values are not allowed), and hence the weights that agent i assigns to neighbors should be small enough, so that "a room" is left for $W_{ii,t}$ to be positive, in view of the fact that the row-sums of W_t must be equal to one. This is however easily achieved in practice by using, e.g., the Metropolis weights, or letting each non-zero off-diagonal weight be small enough, equal to a lower bound on $1/d_{\max}$.

Theoretically, positive diagonals of W_t are important, in order to achieve "continuous information flow" and "averaging" among agents. Basically, they ensure that, if for a product $\Phi(t,1)$, an element at position (i,j) is non-zero, then, irrespective of graph G_{t+1}, the (i,j)-th element of $\Phi(t+1,1)$ is also non-zero.

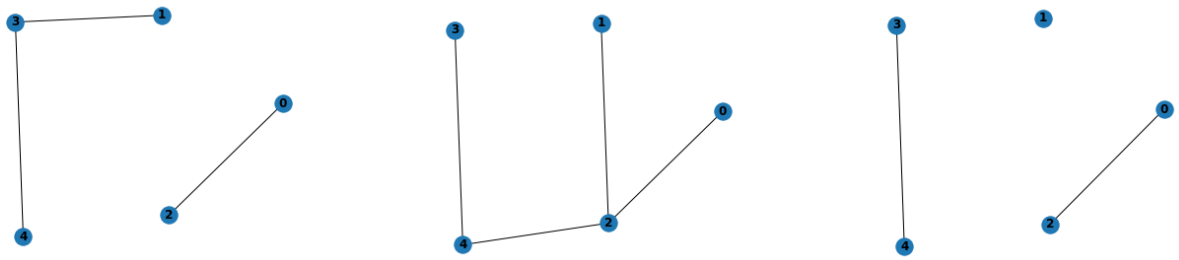
We are interested in the behavior of products $W_t \cdots W_1$. As detailed ahead in Section IV, these matrix products arise with a number of applications in distributed learning, inference and optimization. For future analysis, it will be useful to introduce the following concepts.

Union graph $\Gamma(t,1)$**.** For a collection of graphs $\mathcal{H}$ on the set of vertices V, we denote by $\Gamma(\mathcal{H})$ the graph that contains all the edges of all the graphs in $\mathcal{H}$, $\Gamma(\mathcal{H}) = (V, \cup_{H\in\mathcal{H}} E(H))$, where by $E(H)$ we denote the set of edges of a graph $H \in \mathcal{H}$. We call such a graph *union graph* (of the graphs in $\mathcal{H}$). With a slight abuse of notation, we use the same symbol Γ for the union of subsequent realizations of G_r over any given time window $s \le r \le t$:

$$\Gamma(t,s) = \left(V, \bigcup_{r=s}^{t} E(G_r)\right); \tag{2}$$

in this case we call $\Gamma(t,s)$ the union graph from time s until time t.

Similarly, we define $\Phi(t,s)$ as the product of the weight matrices that occur from time s until time t, for $1 \le s \le t$, i.e., $\Phi(t,s) = W_t \cdots W_s$. To facilitate the presentation, it is also of interest to introduce the error matrix $\widetilde{\Phi}(t,s) = \Phi(t,s) - J$, a norm of which quantifies how close the product is to the one-shot averaging matrix J.



$$W_1 = \begin{bmatrix} 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.67 & 0 & 0.33 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0 & 0 & 0.33 & 0.67 \end{bmatrix} \quad W_2 = \begin{bmatrix} 0.75 & 0 & 0.25 & 0 & 0 \\ 0 & 0.75 & 0.25 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0 & 0.25 \\ 0 & 0 & 0 & 0.67 & 0.33 \\ 0 & 0 & 0.25 & 0.33 & 0.42 \end{bmatrix} \quad W_3 = \begin{bmatrix} 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{bmatrix}$$

Fig. 2: Graphs realizations G_1, G_2, and G_3, with corresponding matrices W_1, W_2, and W_3

Here, we add some examples to illustrate the introduced concepts. In Figure 1 we can see an example of graph $G = (V, E)$. In Figure 2 we have the first three realizations of graphs G_t. Figure 3 illustrates the union graph $\Gamma(t, s)$, for $s = 1$ and $t = 3$.

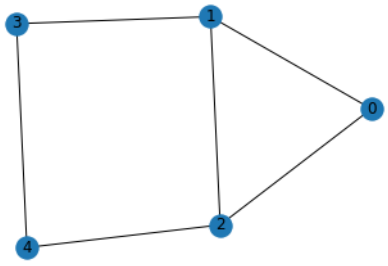
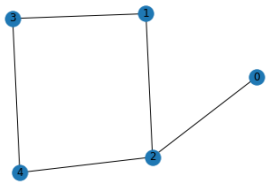


Fig. 1: Simulated graph G



$$\Phi(3,1) = \begin{bmatrix} 0.38 & 0.12 & 0.38 & 0.06 & 0.06 \\ 0.08 & 0.5 & 0.08 & 0.17 & 0.17 \\ 0.38 & 0.12 & 0.38 & 0.06 & 0.06 \\ 0.08 & 0.25 & 0.08 & 0.29 & 0.29 \\ 0.08 & 0 & 0.08 & 0.42 & 0.42 \end{bmatrix}$$

Fig. 3: Union graph $\Gamma(3, 1)$

Sequence of stopping times. Using the notion of the union graph Γ, we define the sequence of times $T_i, i = 1, 2 \ldots$, that mark time instances when Γ gets connected:

$$T_i = \min\{t \geq T_{i-1} + 1 : \Gamma(t, T_{i-1} + 1) \text{ is connected}\},$$

for $i \geq 1$, where $T_0 \equiv 0$. It is well-known that for every time window $[s, t]$ over which the occurred edges accumulate to a connected graph, the spectral norm of the error matrix $\widetilde{\Phi}(t, s)$ constructed over the same time window drops below one, see, e.g., [38]. Hence, the sequence of times $\{T_i\}_{i\geq 1}$ therefore defines the times when the averaging process makes an improvement and gets closer to matrix J.

Number of improvements. For any fixed $t \geq 1$, we introduce the random variable counting the number of improvements until time t, denoted by M_t,

$$M_t = \max\{i \geq 0 : T_i \leq t\}.$$

In other words, if the number of improvements until time t is $M_t = m$, then there exist times $t_1 < t_2 < \cdots < t_m \leq t$, such that the consecutive graphs $\Gamma(t_1, 1), \Gamma(t_2, t_1 + 1), \ldots, \Gamma(t_m, t_{m-1} + 1)$ are all connected; also, if t_m is strictly smaller than t, then $\Gamma(t, t_m + 1)$ is not connected.

For the future analysis it will be of interest to introduce the notion of Edge cut.

Edge cut. For an arbitrary graph $G = (V, E)$, let X be a subset of V. Edge cut of G associated with X is the set of all edges of G with one end in X and the other in $V \setminus X$. By $\mathcal{C}(G)$, we denote the set of all edge cuts associated with some nonempty proper subset of G. For any $C \in \mathcal{C}(G)$, the graph $G_{\setminus C}$ is obtained from the initial graph G by removing all the edges that belong to C, i.e. $G_{\setminus C} = (V, E \setminus C)$. If each edge $\{i, j\} \in E$ is assigned a cost $c_{ij} \in \mathbb{R}$, then, *the minimal edge cut* is defined as the edge cut $C \subset E$ such that the sum of costs of edges in C is minimal among all edge cuts $C \in \mathcal{C}(G)$. We denote the associated cost by

$$MC(G, \{c_{ij}\}_{\{i,j\}\in E}) = \min_{C\in\mathcal{C}(G)} \left\{ \sum_{\{i,j\}\in C} c_{ij} \right\}. \tag{3}$$

III. Main result

This Section states and proves the main result of the paper on the large deviations rate of consensus (1).

A. Rate of consensus

Firtst, we recall that $\lim_{t\to+\infty} p_{ij,t} = \overline{p}_{ij}$ and state the main result.

Theorem 3. *Let Assumptions 1 and 2 hold. Let* $G^\star = (V, E^\star)$*, where* $E^\star = \{\{i,j\} \in E : \overline{p}_{ij} = 1\}$*. Then, for any* $\epsilon \in (0,1]$*:*

$$\lim_{t\to+\infty} -\frac{1}{t}\log \mathbb{P}\left(\|W_t \cdots W_1 - J\| \geq \epsilon\right) = MC(G, \{c_{ij}\}_{\{i,j\}\in E}), \tag{4}$$

where $\overline{q}_{ij} = 1 - \overline{p}_{ij}$ *and*

$$c_{ij} = \begin{cases} -\log \overline{q}_{ij}, & \textit{for } \{i,j\} \in E \setminus E^* \\ +\infty, & \textit{otherwise}, \end{cases}.$$

Theorem 3 characterizes analytically the rate of convergence (1) under a general model of time varying link activation probabilities defined in Assumptions 1 and 2. The theorem shows that there are two distinct cases for the rate (1), which are differed by the graph $G^\star$ that collects links that are "asymptotically certain". From a practical viewpoint, the information flow over $G^\star$ becomes increasingly reliable (without bound), and, in a sense, one can think of $G^\star$ as the (asymptotic) backbone of G. Specifically, when $G^\star$ is connected, the theorem proves that the large deviations rate of consensus (1) is infinite. This result relates to and generalizes the previous results in the literature for connected, static topologies, including static weight matrices the topology of which is connected. Intuitively, when the topology is static and connected, the information flow is guaranteed at each time instant, and under mild assumptions (e.g., Assumption 2), the products $W_t \cdots W_1$ are guaranteed to converge to J exponentially fast, e.g., [40]. Hence, the probabilities in (1) reach zero in finite time, yielding the rate (1) to be equal to infinity. The model assumed in this paper generalizes this result by allowing the static backbone graph to be reached only asymptotically, and moreover at an arbitrary rate.

Furthermore, the condition that $G^\star$ is connected is not only sufficient but also necessary for the rate of consensus to be infinite. When $G^\star$ is not connected, we prove that the rate of consensus (1) is finite and we moreover provide its analytical form as the minimum cut on G with edge costs defined through the limiting probabilities $\overline{q}_{ij}$ of edge absences.

We also relate Theorem 3 with the theory of Markov chains. Essentially, W_t acts as a probability transition matrix of a Markov chain, and the vector $\frac{1}{N}1$ acts as a stationary distribution of a Markov chain. There is a subtle difference in general, as, with Markov chains, we consider right matrix products, while in our case, we deal with left matrix products. This difference is lost when the transition matrices of the Markov chain are symmetric or doubly stochastic, as is the case here. Therefore, one can indeed draw the analogy. In this context, we deal with a Markov chain where the transition probability matrix itself is random and with a time-varying distribution of its realizations. However, we point out that,

despite the connection with Markov chains, our results are novel. Indeed, exponential rates of decay for inhomogenuous Markov chains in random environments have not been derived before.

We further provide an intuitive explnation of why J is the limit of the W_t-products, by drawing the connection with Markov chains theory. We do this only in the expected (mean) convergence sense, which is a weak result but nonetheless is useful as an illustration. Taking the expectation over the dynamics of the products $\Phi(t,1) = W_t \cdots W_1$, we obtain the following deterministic, time-varying dynamics: $p_{t+1} = A_t\, p_t$ where A_t is the expectation of W_t. The dynamics above is precisely the probability transition dynamics of a Markov chain with a deterministic, time-varying, transition matrix, that is symmetric, with strictly positive diagonal, and has a connected support graph. By standard results in Markov chains theory, this Markov chain is stochastic, indecomposable, aperiodic (SIA) [41], and therefore it has a unique stationary distribution that is also uniform.

We also note that this connection with Markov chains helps with the intuition, but only shows convergence in the mean sense. Existing results like [42] show stronger claims, namely that under the condition that $\mathbf{E}[W_t]$ is connected, the $W_t \cdots W_1 \to 1v^T$, when $t \to +\infty$, *almost surely* (not only in the mean sense). As a consequence of their results, the special case is when all the weight matrices W_t are doubly stochastic, then $v = \frac{1}{N}1$. On top of those existing results, as highlighted in the paper, we establish exact rates of exponential convergence.

Using the Theorem 3 it can be shown that, under Assumptions 1 and 2, the product of the weight matrices $\Phi(t,1)$ converges to its limit J almost surely.

Remark 4. *Let Assumptions 1 and 2 hold. Then the matrix $\Phi(t,1)$ converges to J almost surely as $t \to +\infty$, i.e.,*

$$\mathbb{P}\left(\lim_{t\to+\infty} \Phi(t,1) = J\right) = 1. \tag{5}$$

The proof can be found in the Appendix.

IV. Motivation and application examples

We demonstrate practical relevance of Theorem 3 by illustrating the relevance of matrix products $\Phi(t,1) = W_t \cdots W_1$ in several applications: consensus+innovations algorithms and social learning, and distributed learning.

Communication model. In each of the examples to follow, we consider a network of N agents connected by an arbitrary communication topology. Similarly as in Section II, the topology is represented by an undirected graph $G = (V, E)$, where V is the set of agents, and E is the set of possible communication links between agents. Realization of the communication topology at time slot t is denoted

by $G_t = (V, E_t)$, for $t = 1, 2, \dots$ where E_t is the set of links that are online at time t. For an agent i, we let $O_{i,t}$ denote the set of neighbors of i at time t, $O_{i,t} = \{j \in V : \{i, j\} \in E_t\}$.

Consensus+innovations. The agents collaborate over time in a joint detection, estimation, or learning task by intertwining innovation steps – in which new measurements are acquired and incorporated, and consensus steps – where the updated agents' states are communicated to immediate neighbors and subsequently mixed at each via De-Groot averaging [29]. Specifically, at each time t, each agent i, computes the convex combination (i.e., De-Groot averaging) between its own and the neighbors' intermediate states $\widehat{X}_{j,t}$:

$$X_{i,t} = \sum_{j \in O_{i,t} \cup \{i\}} W_{ij,t} \widehat{X}_{j,t}, \tag{6}$$

where the intermediate states are obtained by $\widehat{X}_{j,t} = \frac{t-1}{t} X_{j,t-1} + \frac{1}{t} Z_{j,t}$, $Z_{j,t}$ is the measurement (innovation) of node j at time t and $W_{ij,t}$ is the weight that agent i at time t assigns to the estimate of agent j. The preceding algorithm gives rise to the recursive form in (7) [26],[27], where the relevance of the stochastic matrix products that this work studies is evident and has been exploited, e.g., in (7) [26],[27]:

$$X_{i,t} = \frac{1}{t} \sum_{s=1}^{t} [\Phi(t, s)]_{ij} Z_{j,s}. \tag{7}$$

Specific instantiations and widely studied applications of algorithm (6) are distributed detection and distributed estimation. For example, in distributed detection based on algorithm (7), Theorem 3 is a key step to establish the error exponent for the detection error probability, see [26] for details.

Social learning. The idea of social learning is for a group of people to distinguish between M different hypotheses, through local updates and collaborative information exchange [43]. Each node i over time draws observations $Y_{i,t}$ from (the true) distribution $f_{i,M}$ (hypothesis H_M). The remaining $M-1$ candidate distributions at node i in hypothesis testing are $f_{i,m}$ (hypothesis H_m), $m = 1, \dots, M-1$. The algorithm starts at each node with initial private belief value $q_{i,t}^m > 0$, $m = 1, \dots M-1$. Upon receiving new local observation $Y_{i,t}$, the nodes compute their public belief values by:

$$b_{i,t}^m = \frac{f_{i,m}(Y_{i,t}) q_{i,t-1}^m}{\sum_{l=1}^{M} f_{i,l}(Y_{i,t}) q_{i,t-1}^l}, \tag{8}$$

for each $m = 1, \dots, M$. Each node then sends its updated public belief vector to its neighbors. Upon receiving the neighbors' (public) beliefs, the node updates its private beliefs as follows:

$$q_{i,t}^m = \frac{e^{\sum_{j \in O_{i,t}} W_{ij,t} \log b_{i,t}^m}}{\sum_{l=1}^{M} e^{\sum_{j \in O_{i,t}} W_{ij,t} \log b_{i,t}^l}}, \tag{9}$$

for each $m = 1, \dots, M$. It is easy to show (the details can be found in [44]) that the preceding algorithm admits the recursive representation of the consensus+innovations algorithm in (7), with

$$Z_{i,t}^m = \log \frac{f_{i,m}(Y_{i,t})}{f_{i,M}(Y_{i,t})}, \tag{10}$$

$$\widehat{X}_{i,t}^m = \frac{1}{t} \log \frac{q_{i,t}^m}{q_{i,t}^M}, \tag{11}$$

$$X_{i,t}^m = \frac{1}{t} \log \frac{b_{i,t}^m}{b_{i,t}^M}. \tag{12}$$

Hence, the results of this work are of direct relevance for social learning in which the interactions between the social agents are random and with rates of interactions varying over time.

Distributed multiagent optimization and learning. We now consider standard multi-agent consensus optimization problem, where the agents in a connected network collaboratively solve the following unconstrained problem:

$$\text{minimize } f(x) := \sum_{i=1}^{N} f_i(x). \tag{13}$$

Here, $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is a local agent i's loss function, for example, an empirical loss for the considered machine learning model (linear regression, logistic loss, hinge loss, etc.) based on agent i's training data.

While a number of distributed multi-agent optimization algorithms have been proposed, a simple widely considered method is the decentralized (sub)gradient descent (DGD) [45]. For $t = 0, 1, ...$, DGD works as follows:

$$X_{i,t+1} = \sum_{j \in O_{i,t}} W_{ij,t} \, X_{j,t} - \alpha_t \, \nabla f_i(X_{i,t}). \tag{14}$$

Here, $X_{i,t} \in \mathbb{R}^d$ is agent i's solution estimate at iteration t, $\nabla f_i(X_{i,t})$ is the gradient (or an arbitrary sub-gradient) of f_i evaluated at $X_{i,t}$, $\alpha_t > 0$ is the step-size, and the $W_{ij,t}$'s and $O_{i,t}$'s are the averaging weights and agents' neighborhoods, as before.

In the context of the analysis of (14) and similar methods, the analysis of products $\Phi(t, s)$ can play an important role. Specifically, they determine the dynamics of the *disagreement estimates* $\widetilde{X}_{i,t} = X_{i,t} - \overline{X}_t$, where $\overline{X}_t = \frac{1}{N} \sum_{i=1}^{N} X_{i,t}$ is the (hypothetical) global average of the agents' solution estimates. That is, the analysis of the dynamics of $\widetilde{X}_{i,t}$ is an important intermediate step when establishing convergence of (14), see, e.g., [45], [46]. Namely, the analysis of (14) will typically consist of two steps: 1) show that $\widetilde{X}_{i,t} \to 0$, almost surely (consensus); and 2) show that $f(\overline{X}_t) - f^\star \to 0$, almost surely, where $f^\star = \inf_x f(x)$ (convergence of function values). In other words, the analysis considers that the following two conditions need to be fulfilled: 1) the agents need to reach a consensus; and 2) the point at which

the consensus is reached is optimal in terms of the objective function value. For the first, consensus condition, the matrix products $\Phi(t,s)$ play an important role. Namely, it can be shown that [45], [46][1]

$$\widetilde{X}_{t+1} = (W_t - J)\widetilde{X}_t + \alpha_t\, g_t, \tag{15}$$

where $\widetilde{X}_t = (\widetilde{X}_{t,1}, ..., \widetilde{X}_{t,N})^\top$, $g_t = (I-J)h_t$, and $h_t = (\nabla f_1(X_{1,t}), ..., \nabla f_N(X_{N,t}))^\top$. By unwinding the recursion (15), we obtain:

$$\widetilde{X}_t = \sum_{s=0}^{t-1} (\Phi(t,s) - J)\, \alpha_s\, g_s. \tag{16}$$

Under a common uniformly bounded gradients assumption, (16) implies:

$$\|\widetilde{X}_t\| \leq G\sqrt{N} \sum_{s=0}^{t-1} \|\Phi(t,s) - J\|\, \alpha_s. \tag{17}$$

where $G = \max_{i=1,...,N} \sup_x \|\nabla f_i(x)\|$. From (17), we can see that, by studying the behavior of quantities $\|\Phi(t,s) - J\|$ (available thanks to this paper's results), we can provide estimates of the disagreement size $\|\widetilde{X}_t\|$. Detailed derivations are left for future work.

Error exponent (distributed detection). When large number of observations is needed to reach the desired accuracy, detection performance is typically measured through error exponents; for example, with the Bayesian hypothesis testing one is interested in the exponential decay rate R of the expected error probability, and similarly, for the Neyman-Pearson hypothesis testing, the decay rates of the probability of false alarm and the probability of missed detection. It is well known that, under Bayesian hypothesis testing, the error exponent R for the centralized hypothesis testing is given by the Chernoff information computed from the distributions of the network-wide vector measurements under the two hypotheses. However, this optimal rate is not guaranteed with distributed algorithms. Specifically, when the communication links are intermittent, the information flow, enabled through local averaging in (6), can be cut for arbitrary long periods. This would prevent the new measurements to propagate, and disable averaging out the measurement noise. To understand the practical importance of the rate R for distributed detection, adopting, for large T, the approximate expression for the Bayesian error[2] (say, at an arbitrary network node) $P_{\rm e} \approx e^{-TR}$. We obtain that the minimal number of observations T needed to achieve a given target accuracy (e.g., $P_{\rm e} = 0.05$) equals

$$T^\star \approx -\log(0.05)/R. \tag{18}$$

Hence, the higher the exponent R, the faster is the target accuracy achieved.

[1]For notational simplicity, we let here $d = 1$, while the analysis can be extended for any $d > 1$.

[2]The Bayesian error probability $P_{\rm e}$ can be expressed as $P_{\rm e} = \kappa_T e^{-TR}$, where the first factor $\frac{1}{T}\log\kappa_T$ vanishes with T. Hence, $\frac{1}{T}\log\kappa_T/(RT) \to 1$, as $T \to +\infty$, motivating the above approximation for large T.

Previous works have investigated analytically the exponent R for distributed detection (6) for both deterministic and stochastic networks, showing that the large deviations rate of consensus (4) plays a critical role in determining the value of R. Specifically, for deterministic networks, it has been shown that the error exponent R is always equal to the optimal, centralized error exponent, independent of the topology and the averaging coefficients (weight matrices) used. This is in accordance with the result of Theorem 3, which asserts the infinite rate of consensus in the deterministic case hence guaranteeing infinitely fast information flow in the error exponent sense. For stochastic networks, the above works provided analytical conditions for asymptotic optimality of distributed detection, defined in terms of system parameters, such as the number of nodes N and local Chernoff information $\mathcal{C}h$. Specifically, for Gaussian observations with distributions $\mathcal{N}(m_0, \sigma^2)$ and $\mathcal{N}(m_1, \sigma^2)$, the condition has the following form

$$MC \geq N(N-1)\mathcal{C}h, \tag{19}$$

where $\mathcal{C}h = (m_1 - m_0)^2/(8\sigma^2)$ is the nodes' individual Chernoff information and MC is the minimum cut value from Theorem 3 – when (19) holds, algorithm (6) achieves the optimal error exponent, equal to $R = N\mathcal{C}h$, at each node in the network. Relating with (18), when (19) holds, the desired detection accuracy is reached at the earliest possible time.

Inaccuracy rates (distributed learning). Similarly as with error exponents, to characterize the speed of convergence of the solution estimates $X_{i,t}$ to $x^\star$, one can consider a confidence interval around $x^\star$, $\|X_{i,t} - x^\star\| < \epsilon$, seeking for the earliest possible time when the probability that the parameter estimates $X_{i,t}$ belong to this region of $\mathbb{R}^d$, reaches the desired value. It can be shown that the probabilities of the complement (large deviation) event $\|X_{i,t} - x^\star\| \geq \epsilon$ decay exponentially fast, e.g., [44]. The corresponding exponents are known in the literature as inaccuracy rates [30], [31]. For De-Groot based distributed estimation, as shown in Chapter 4.6 of [47], they are, similarly to distributed detection, critically determined by the large deviations rate of consensus (1), albeit with a more complex relation. We omit here the detailed treatment as it is out of scope of the current paper. For large deviations analysis of learning algorithms, we point to recent reference [48].

Federated learning. Interestingly, matrix products $\Phi(t, 1)$ also arise in other applications where an explicit form of graph sequence does not appear. An example is a server-clients federated learning system with either full or partial clients participation [49]. Therein, each client has a loss function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$, and the goal is to minimize the sum of the clients' losses. To be specific, we consider the FedAvg method that works as follows. At each iteration t, a subset of K out of N clients is selected uniformly at random. We denote by X_t the global model available at the server, and by $X_{i,t}$ the local model at

client i, $i = 1, ..., N$. The selected clients i receive a global model X_t from the server, and then make the following update:

$$X_{i,t+1} = X_t - \alpha \nabla f_i(X_t), \tag{20}$$

where $\alpha > 0$ is a step-size and ∇ denotes the gradient operator.[3] The remaining clients j stay idle, meaning that $X_{j,t+1} = X_{j,t}$. All active clients sent their $X_{i,t+1}$'s to the server that subsequently averages the received vectors. This procedure can be modeled via graphs G_t introduced in Section II in the following way. Each client is a node in G_t, while the server is formally not a part of the graph. The graph G_t has as a subgraph a K-node complete graph that includes precisely the clients that participate in communication with the server at iteration t, and the graph G_t has no other links. The weight matrix W_t has its $K \times K$ submatrix at the positions that correspond to the active clients equal to J_K, while the remaining off-diagonal blocks are zero, and the remaining diagonal blocks equal the identity. Letting $d = 1$ for simplicity, and collecting all the $X_{i,t}$'s in a $N \times 1$ vector $\mathbf{X}_t = (X_{1,t}, ..., X_{N,t})^\top$, the FedAvg update rule can be written as $\mathbf{X}_{t+1} = W_t (\mathbf{X}_t - \alpha Z_t \odot g_t)$, where $g_t = (\nabla f_1(X_{1,t}), ..., \nabla f_N(X_{N,t}))^\top$, $\odot$ is the Hadamard product, and Z_t is a random zero-one vector that has non-zero elements precisely at the positions of active clients. The example in Figure 4 illustrates G_t and W_t for a toy example of a five-node federated learning setup where nodes 2-4 active, while nodes 1 and 5 are idle.

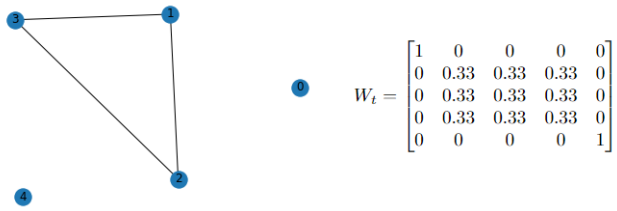


Fig. 4: G_t and W_t in an example of federated learning setup.

Spatial autoregression. Finally, the model that we consider is related to spatial autoregression. Therein, $X_{i,t}$ corresponds to a spatial field's value at time t for a specific location, while matrix W_t captures the random variable dependencies at time t among the $X_{i,t}$'s, e.g., according to geographical proximity [50].

A. Application example: Sparsifying communication

In this section, we are concerned with the distributed detection application and whether the optimality of error exponents can be achieved with fewer communications. To motivate the approach, we consider

[3]For simplicity, we let the number of local gradient updates equal to one.

the following example. Suppose we are given a network $G = (V, E)$, with N nodes, which run the algorithm (6) and where each link $\{i, j\} \in E$ occurs with probability $\overline{p}_{ij}$, e.g., due to imperfect communication channels. We assume that the measurements are Gaussian and the minimum cut $MC = MC(\{\overline{p}_{ij}\}_{\{i,j\}\in E})$ satisfies the condition (19) so that the error exponent at each node equals $R = N\mathcal{C}h$. To reach the target accuracy $P_{\rm e} = \eta$, it is sufficient to perform $T = -\log(\eta)/R$ rounds of iterations. Each iteration incurs the expected number of communications equal to $\mathbb{E}\left[\sum_{\{i,j\}\in E} 1_{\{i,j\}}\right] = \sum_{\{i,j\}\in E} \overline{p}_{ij}$, leading to the expected cumulative number of communications at iteration T equal to:

$$C_{\rm const} = T \sum_{\{i,j\}\in E} \overline{p}_{ij}. \tag{21}$$

Theorem 3 on the other hand asserts that the same asymptotic rate of consensus MC and the same error exponent $N\mathcal{C}h$, are guaranteed as long as the probability of each link $p_{ij,t}$ converges to the same limit $\overline{p}_{ij}$. Specifically, the same rates, MC and $N\mathcal{C}h$, will be achieved for the case when $p_{ij,t} = \overline{p}_{ij}(1 - o_t)$, where o_t is an arbitrary function that decays to zero (e.g., $o_t = 1/t^2, t \geq 1$). The term $(1 - o_t)$ decreasing the communication probability can be practically realized by a random variable η_{ij}, generated independently at each link, which censors transmissions across the respective link with probability o_t. The total number of communications until time T for the sparsifying scheme is then given by:

$$C_{\rm sparse} = \sum_{t=1}^{T} \sum_{\{i,j\}\in E} \overline{p}_{ij}(1 - o_t) \tag{22}$$

$$= C_{\rm const}\left(1 - \frac{\sum_{t=1}^{T} o_t}{T}\right), \tag{23}$$

leading to the savings potentially equal to $C_{\rm const}\frac{\sum_{t=1}^{T} o_t}{T}$. In Subsection VI-B we compare numerically the number of communications for the constant and the sparsifying communication protocol needed to achieve target error probability.

V. Proof of the main result

In this Section we prove Theorem 3. We first show that the rate of consensus (1) equals $-MC$ for the case when G^* is not connected, in Subsection V-A. In Subsection V-B we show that the rate of consensus (1) diverges to negative infinity, when G^* is connected.

A. Finite rate case

In this Subsection we prove Theorem 3 for the case when G^* is not connected by showing the lower and the upper large deviation bound:

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \geq -MC, \tag{24}$$

$$\limsup_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \leq -MC, \tag{25}$$

where $MC = MC(G, \{c_{ij}\}_{\{i,j\}\in E})$.

Lower bound.

Observe that, since $\epsilon \in (0,1]$:

$$\mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \geq \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| = 1\right).$$

We have that sufficient condition for event $\{\|\widetilde{\Phi}(t,1)\| = 1\}$ is that $\Gamma(t,1)$ is not connected (see e.g. Lemma 5 in [38]). Using the fact that $G^\star$ is not connected, we conclude that there must be at least one edge cut with all edges whose limiting probability $\overline{p}_{ij}$ is strictly less than 1, i.e., they all belong to $E \setminus E^\star$. Let C_μ denote the minimal edge cut of G, where the link costs are assigned as in the claim of the theorem. Sufficient condition for $\Gamma(t,1)$ not to be connected is that all the links in C_μ were inactive over the time interval from time 1 to t. Thus,

$$\begin{aligned}\mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) &\geq \mathbb{P}\left(E_r \cap C_\mu = \emptyset, r = 1,\ldots,t\right) \\ &= \prod_{r=1}^{t} \prod_{\{i,j\}\in C_\mu} q_{ij,r},\end{aligned} \tag{26}$$

where $q_{ij,r} = 1 - p_{ij,r}$ and the equality follows by the first and the second part of the Assumption 1. Computing the logarithm and dividing by t, we obtain

$$\begin{aligned}&\frac{1}{t} \log \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \\ &\geq \frac{1}{t} \sum_{\{i,j\}\in C_\mu} \sum_{r=1}^{t} \log q_{ij,r} = \sum_{\{i,j\}\in C_\mu} \frac{1}{t} \sum_{r=1}^{t} \log q_{ij,r}.\end{aligned} \tag{27}$$

Recall that, for each $\{i,j\} \in E$, $q_{ij,r}$ converges to $\overline{q}_{ij}$, when $r \to +\infty$. By the Cesàro means theorem (see Theorem 50 in [51]), we thus have that, as $t \to +\infty$, $\frac{1}{t}\sum_{r=1}^{t} \log q_{ij,r}$ converges to $\log \overline{q}_{ij}$ for each $\{i,j\} \in C_\mu$. Thus, taking the limit in (27) completes the proof of the lower bound (24).

Upper bound.

We start with Lemma 5 borrowed from [38], which asserts that, if the number of improvements until time t scales linearly with t, then, starting from some finite time t_0, the events $\left\{\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right\}$ have zero probabilities.

Lemma 5. *Consider the sequence of events* $\{M_t \geq \beta t\}$*, where* $\beta \in (0,1]$*,* $t = 1, 2, \ldots$*. For every* $\beta, \epsilon \in (0,1]$*, there exists sufficiently large* $t_0 = t_0(\beta, \epsilon)$ *such that*

$$\mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon, M_t \geq \beta t\right) = 0, \ \forall t \geq t_0(\beta, \epsilon). \tag{28}$$

Using the preceding result, it is easy to see that, when t is large enough, for any fixed $\beta \in (0,1)$, a necessary condition for $\|\widetilde{\Phi}(t,1)\| \geq \epsilon$ is that $M_t < \beta t$. Thus, we have that for each $\beta \in (0,1)$, there holds for all $t \geq t_0(\beta, \epsilon)$:

$$\mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) = \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon, M_t < \beta t\right) \tag{29}$$

$$\leq \mathbb{P}\left(M_t < \beta t\right). \tag{30}$$

Now,

$$\mathbb{P}\left(M_t < \beta t\right) = \sum_{m=0}^{\lceil \beta t \rceil - 1} \mathbb{P}(M_t = m)$$

$$= \sum_{m=0}^{\lceil \beta t \rceil - 1} \sum_{1 \leq t_1 \leq \cdots \leq t_m \leq t} \mathbb{P}(T_l = t_l, \text{ for } 1 \leq l \leq m, T_{m+1} > t), \tag{31}$$

where in the second equality we consider all possible realizations of improvement times $T_l, \ l \leq m$. We focus on one arbitrary allocation $T_l = t_l, \ 1 \leq l \leq m, \ T_{m+1} > t$, and the respective probability of which is $\mathbb{P}(T_l = t_l, \text{ for } l \leq m, T_{m+1} > t)$. By the construction of the sequence T_l, for each $l \leq m$, we know that the uniongraph $\Gamma(t_l - 1, t_{l-1})$ is not connected. Also, the condition $T_{m+1} > t$ implies that $\Gamma(t, T_m)$ is not connected. Denoting $t_0 = 0$ and $t_{m+1} = t + 1$ for compact representation, we have

$$\mathbb{P}(T_l = t_l, \text{ for } l \leq m, T_{m+1} > t)$$

$$\leq \prod_{l=1}^{m+1} \mathbb{P}(\Gamma(t_l - 1, t_{l-1} + 1) \text{ not connected}), \tag{32}$$

where we used Assumption 1, i.e. the independence of the graph realizations. Note that, for arbitrary $t_a > t_b$, the event that supergraph $\Gamma(t_a, t_b)$ is not connected can be represented as the union of events that all edges from an arbitrary edge cut of G were inactive over time window $t_a \leq r \leq t_b$, i.e.

$$\{\Gamma(t_b, t_a) \text{ not connected}\}$$

$$= \cup_{C \in \mathcal{C}(G)} \left\{E_r \cap C = \emptyset, t_a \leq r \leq t_b\right\}. \tag{33}$$

Applying (33) to each of the intervals $t_{l-1}+1 \leq r \leq t_l - 1$ and computing the probabilities, we get by the union bound

$$\begin{aligned}
&\mathbb{P}(\Gamma(t_l - 1, t_{l-1} + 1) \text{ not connected}) \\
&\leq \sum_{C \in \mathcal{C}(G)} \mathbb{P}(E_r \cap C = \emptyset, t_{l-1} + 1 \leq r \leq t_l - 1) \\
&= \sum_{C \in \mathcal{C}(G)} \prod_{t_{l-1}+1 \leq r \leq t_l - 1} \prod_{\{i,j\} \in C} q_{ij,r} \qquad (34) \\
&\leq \sum_{C \in \mathcal{C}(G)} \prod_{t_{l-1}+1 \leq r \leq t_l - 1} Q_{C,r}, \qquad (35)
\end{aligned}$$

where $Q_{C,r} = \prod_{\{i,j\} \in C} q_{ij,r}$, for all $r \geq 1$. Introducing now $Q_r = \max_{C \in \mathcal{C}(G)} Q_{C,r}$ for $r \geq 1$, we have,

$$\begin{aligned}
&\mathbb{P}(\Gamma(t_l - 1, t_{l-1} + 1) \text{ not connected}) \\
&\leq |\mathcal{C}(G)| \prod_{t_{l-1}+1 \leq r \leq t_l - 1} Q_r.
\end{aligned}$$

The preceding bound applies to each of the terms in (32), hence we obtain

$$\begin{aligned}
&\mathbb{P}(T_l = t_l, \text{ for } l \leq m, T_{m+1} > t) \\
&\leq |\mathcal{C}(G)|^{m+1} \prod_{l=1}^{m+1} \prod_{t_{l-1}+1 \leq r \leq t_l - 1} Q_r. \qquad (36)
\end{aligned}$$

Now, since for each link $\{i,j\}$, $q_{ij,r}$ converges to $\overline{q}_{ij} = 1 - \overline{p}_{ij}$, we know that for each $C \in \mathcal{C}(G)$, $Q_{C,r}$ converges to

$$Q_C := \prod_{\{i,j\} \in C} \overline{q}_{ij}. \qquad (37)$$

The latter quantity is either greater than 0, if all the links in the respective cut are from $E \setminus E^\star$, or it is equal to 0 if there is a link in the cut that belongs to $E^\star$. We note further that, since C_μ is the minimal edge cut of G with the link cost assignment as in the claim of the theorem, we have that $Q_C \leq Q_{C_\mu}$, for all $C \in \mathcal{C}(G)$. By the convergence of $Q_{C,r}$ to Q_C, for each cut C (including Q_{C_μ}), it must be that there exists $r_1 \in \mathbb{N}$ such that for all $r \geq r_1$, $Q_{C,r} \leq Q_{C_\mu,r}$ for all $C \in \mathcal{C}(G)$. This implies that for all $r \geq r_1$, $Q_r = \max_{C \in \mathcal{C}(G)} Q_{C,r} = Q_{C_\mu,r}$. Also, for all $0 < \xi < 1 - Q_{C_\mu}$, there exists $r_2 \in \mathbb{N}$ such that for all $r \geq r_2$, $Q_{C_\mu,r} \leq Q_{C_\mu} + \xi$. Thus for all $r \geq r_0 := \max\{r_1, r_2\}$ we have,

$$Q_r = Q_{C_\mu,r} \leq Q_{C_\mu} + \xi. \qquad (38)$$

Maximal product, Q_r, over all cuts, after a certain iteration r_1 is always equal to the product $Q_{C_\mu,r}$ associated with the cut C_μ, and after r_2, that product is in ξ vicinity of its limit. Taking maximum of those two times, we get the upper bound (38) for Q_r.

Using the fact that $Q_r \leq 1$ for all $r \leq r_0$, applying (38) for all $r \geq r_0$, and using fact that at most m factors were excluded between r_0 and t from (36) we have:

$$\begin{aligned}&\mathbb{P}(T_l = t_l, \text{ for } l \leq m, T_{m+1} > t)\\ &\leq |\mathcal{C}(G)|^{m+1}(Q_{C_\mu} + \xi)^{t-r_0-m}. \end{aligned} \tag{39}$$

We know that for fixed m, equation (39) holds for any possible realisations of times $t_1, \ldots, t_m$ and by Stirlings approximation $\binom{t}{m} \leq (\frac{te}{m})^m$, combining (39) and (31) we obtain:

$$\begin{aligned}&\mathbb{P}\left(M_t < \beta t\right)\\ &\leq \sum_{m=0}^{\lceil \beta t \rceil - 1} \binom{t}{m} |\mathcal{C}(G)|^{m+1}(Q_{C_\mu} + \xi)^{t-r_0-m}\\ &\leq \sum_{m=0}^{\lceil \beta t \rceil - 1} \left(\frac{te}{m}\right)^m |\mathcal{C}(G)|^{m+1}(Q_{C_\mu} + \xi)^{t-r_0-m}.\end{aligned}$$

We claim that the function $f(x) = (\frac{te}{x})^x$ is increasing on the interval $(0, t)$. Specifically, let $g(x) = \log f(x) = x \log t + x - x \log x$. We have that $g'(x) = \log t - \log x \geq 0$, for $x \in (0, t)$. Using the fact that composition of the increasing fuctions is increasing, the claim follows. Further, since $|\mathcal{C}(G)|$ is integer, we have $|\mathcal{C}(G)|^{m+1} \leq |\mathcal{C}(G)|^{\beta t+1}$ and since for ξ sufficiently small, $Q_{C_\mu} + \xi \in (0, 1)$, we have $(Q_{C_\mu} + \xi)^{t-r_0-m} \leq (Q_{C_\mu} + \xi)^{t-r_0-\beta t}$. Summarising, we have

$$\mathbb{P}\left(M_t < \beta t\right) \leq (\beta t) \left(\frac{te}{\beta t}\right)^{\beta t} |\mathcal{C}(G)|^{\beta t+1}(Q_{C_\mu} + \xi)^{t-r_0-\beta t}.$$

Computing the logarithm and dividing by t, we have

$$\begin{aligned}&\frac{1}{t}\log \mathbb{P}\left(M_t < \beta t\right) \leq \frac{1}{t}\log(\beta t) + \beta \log(\frac{e}{\beta})\\ &+ (\beta + \frac{1}{t}) \log |\mathcal{C}(G)| + (1 - \frac{r_0 - \beta t}{t}) \log \left(Q_{C_\mu} + \xi\right).\end{aligned}$$

Taking the limit $t \to +\infty$ we have

$$\begin{aligned}&\lim_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(M_t < \beta t\right)\\ &\leq \beta \log(\frac{e}{\beta}) + \beta \log |\mathcal{C}(G)| + (1 + \beta) \log \left(Q_{C_\mu} + \xi\right).\end{aligned} \tag{40}$$

We note that (40) holds for each $\beta \in (0, 1]$ and any ξ suficiently small. Taking the infimum of both sides with respect to β we have

$$\inf_\beta \lim_{t \to +\infty} \frac{1}{t} \log \mathbb{P}\left(M_t < \beta t\right) \leq \log \left(Q_{C_\mu} + \xi\right).$$

Finally, taking the infimum with respect to $\xi > 0$, the upper bound (25) follows. This concludes the case of Theorem 3, when G^* is not connected.

B. Infinite rate case

Now we prove Theorem 3 for the case when $G^\star$ is connected, for which the rate is infinite. Reapplying the steps in (29)-(33) we focus on:

$$\begin{aligned}&\mathbb{P}(\Gamma(t_l-1,t_{l-1}+1) \text{ not connected})\\ &\leq \sum_{C\in\mathcal{C}(G)}\prod_{t_{l-1}+1\leq r\leq t_l-1}\prod_{\{i,j\}\in C} q_{ij,r}. \end{aligned} \tag{41}$$

Since $G^\star$ is connected, every $C \in \mathcal{C}(G)$ contains at least one link $\{i^\star, j^\star\} \in E^\star$. Note also that, from the fact that $q_{ij,r} \leq 1$, which holds for all $r \geq 1$ and all $\{i,j\} \in E$, we have $\prod_{\{i,j\}\in C} q_{ij,r} \leq q_{i^*j^*,r}$. Introducing now $q_r := \max_{\{i,j\}\in E^\star} q_{ij,r}$, from (41) we obtain

$$\begin{aligned}&\mathbb{P}(\Gamma(t_l-1,t_{l-1}+1) \text{ not connected})\\ &\leq \sum_{C\in\mathcal{C}(G)}\prod_{t_{l-1}+1\leq r\leq t_l-1} q_{i^*j^*,r}\\ &\leq |\mathcal{C}(G)|\prod_{t_{l-1}+1\leq r\leq t_l-1} q_r.\end{aligned}$$

The preceding inequality holds for each of the terms in (32), hence we obtain

$$\begin{aligned}&\mathbb{P}(T_l = t_l, \text{ for } l\leq m, T_{m+1} > t)\\ &\leq |\mathcal{C}(G)|^{m+1}\prod_{l=1}^{m+1}\prod_{t_{l-1}+1\leq r\leq t_l-1} q_r.\end{aligned}$$

Now, let $L_t := \lceil \log t \rceil$ and define $\widehat{q}_t := \max_{L_t\leq r\leq t} q_r$, for all $t \geq 1$. Using the same arguments as in the case when G^* is not connected (eq. (39)), we have

$$\mathbb{P}(T_l = t_l, \text{ for } l\leq m, T_{m+1} > t) \leq |\mathcal{C}(G)|^{m+1}\widehat{q}_t^{\,t-L_t-m}.$$

Hence, in (31), we obtain

$$\begin{aligned}\mathbb{P}\left(M_t < \beta t\right) &\leq \sum_{m=0}^{\lceil\beta t\rceil-1}\binom{t}{m}|\mathcal{C}(G)|^{m+1}\widehat{q}_t^{\,t-L_t-m}\\ &\leq \sum_{m=0}^{\lceil\beta t\rceil-1}\left(\frac{te}{m}\right)^m|\mathcal{C}(G)|^{m+1}\widehat{q}_t^{\,t-L_t-m}\\ &\leq (\beta t)\left(\frac{te}{\beta t}\right)^{\beta t}|\mathcal{C}(G)|^{\beta t+1}\widehat{q}_t^{\,t-L_t-\beta t}.\end{aligned}$$

Further, computing the logarithm and dividing by t, we have

$$\frac{1}{t}\log\mathbb{P}\left(M_t<\beta t\right)\leq\frac{1}{t}\log(\beta t)+\beta\log(\frac{e}{\beta})$$
$$+(\beta+\frac{1}{t})\log|\mathcal{C}(G)|+(1-\frac{L_t+\beta t}{t})\log\widehat{q}_t. \tag{42}$$

Before taking the limit we prove the Lemma 6 which regards the convergence of the sequence $\widehat{q}_t$.

Lemma 6. *Sequence $\widehat{q}_t$ converges to 0, when $t\to+\infty$.*

Proof By the definition of E^*, for all $\{i,j\}\in E^*$, $q_{ij,r}$ converges to 0, hence it is easy to see that q_r converges to 0 as well. Now, we define $\widehat{\widehat{q}}_t:=\max_{r\geq L_t}q_r$. For each $\varepsilon>0$ there is r_0 such that for all $r\geq r_0$, $q_r<\varepsilon$. Hence, for $t\geq t_0=\lceil e^{r_0}\rceil$ we have:

$$L_t=\lceil\log t\rceil\geq\lceil\log\lceil e^{r_0}\rceil\rceil\geq\lceil\log e^{r_0}\rceil=r_0.$$

Furthermore, for $t\geq t_0$

$$\widehat{\widehat{q}}_t\leq\max_{r\geq r_0}\{q_r\}\leq\varepsilon.$$

Now, using the convergence of $\widehat{\widehat{q}}_r$ and the fact that $0\leq\widehat{q}_t\leq\widehat{\widehat{q}}_t$ we prove the lemma. □

Taking the limit $t\to+\infty$ and infimum with respect to $\beta>0$ and using the result of Lemma 6 in (42), we conclude the proof of the second case of Theorem 3. □

VI. Numerical results

We use Monte Carlo simulations for two sets of numerical experiments. First, we estimate in Subsection VI-A the probability $\mathbb{P}(\|W_t\cdots W_1-J\|\geq\epsilon)$ through time t, for both the constant and the time-varying model.

Then in Subsection VI-B, we estimate the probability of error for distributed detection algorithm described in Section IV. Here, we also make a comparison between the constant and the time-varying model.

A. Estimating the probability $\mathbb{P}(\|W_t\cdots W_1-J\|\geq\epsilon)$

We first describe the simulation setup. We consider a geometric network with $N=7$ sensors, shown in Figure 5. The total number of (undirected) links is seven. Each of these seven links is active, at time t, with probability $p_{ij,t}$. There are two types of links, the ones for which the limiting probability of activation is $\overline{p}_{ij}=1$, shown in Figure 5 in green; for the remaining links, shown in Figure 5 in red, the limiting probability of activation is $\overline{p}_{ij}=0.2$.

We simulated numerically two different models: the time varying and the constant model. With the time-varying model, the sequence of probabilities $p_{ij,t}$ is converging to its limit $\overline{p}_{ij}$ with speed $\frac{1}{t^2}$, $p_{ij,t} = \overline{p}_{ij} - \frac{1}{(t+2)^2}$, where $\overline{p}_{ij} = 1$ for the green links and $\overline{p}_{ij} = 0.2$ for the red links. For the constant model, the sequence of link failure probabilities is constant throughout iterations, and equals $\overline{p}_{ij}$.

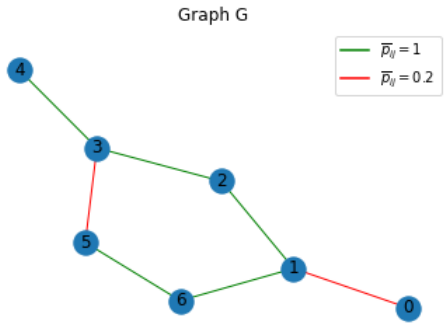


Fig. 5: Graph G with 7 nodes and 7 links; The links in $G^\star$ are colored in green ($\overline{p}_{ij} = 1$), while the remaining links in G are colored in red ($\overline{p}_{ij} = 0.2$)

For the averaging weights, we use Metropolis weights, i.e., if link $\{i, j\} \in E$ is online, we assign $W_{ij,t} = 1/(1 + \max\{d_{i,t}, d_{j,t}\})$, where $d_{i,t}$ is the degree of node i at time t and $W_{ij,t} = 0$ otherwise. Also, $W_{ii,t} = 1 - \sum_j W_{ij,t}$.

We show the results in Figure 6. The green curve plots a straight line $\mathrm{MC}t$ with the theoretical rate from Theorem 3, equal to the minimum cut $\mathrm{MC} = \log 0.8 = -0.223$. The orange and the blue curves plot $\log \mathbb{P}(\|W_t \cdots W_1 - J\| \geq \epsilon)$, for the time-varying and constant models, respectively. It can be seen from the figure that the slopes of the blue and the orange curve approach their theoretical limit MC. We can also see that the orange curve is slightly worse (higher) than the blue one. This result is expected as the activation probabilities start slowly, hence the interactions are sparser at the beginning, and they gradually increase to their limiting values as the iterations progress, as can be observed from the figure.

B. Estimating the probability of error

For the distributed detection problem in Section IV we consider a geometric network with $N = 10$ sensors. We place the sensors uniformly over a unit square, and connect those sensors whose Euclidean distance is less than a radius ($r = 0.5$). The total number of (undirected) links is 23. The network is shown in Figure 7. There are two types of links, with limiting probabilities of activation equal to $\overline{p}_{ij} = 1$ shown in green, and $\overline{p}_{ij} = 0.5$ shown in red. With the time-varying model, the sequence of probabilities

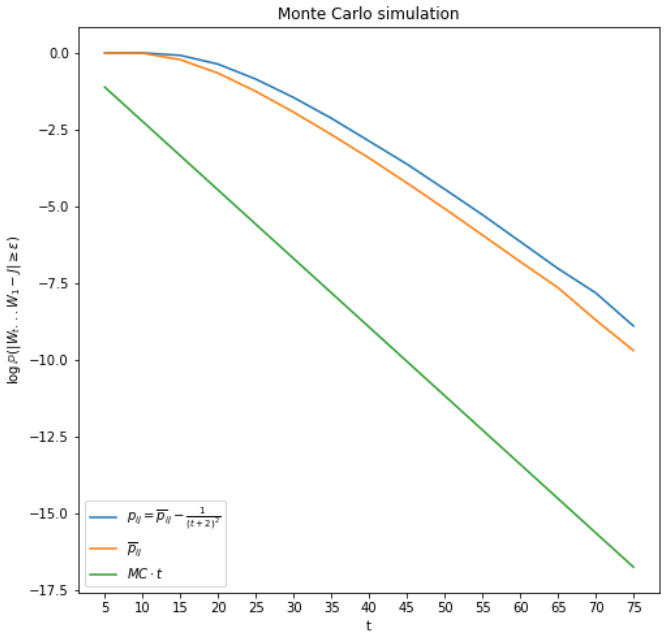


Fig. 6: Estimated probability (in the log scale), $\log \mathbb{P}(\|W_t \cdots W_1 - J\| \geq \epsilon)$, with constant activation probabilities (blue), time-varying activation probabilities (orange) and $\mathrm{MC} \cdot t$ (green) versus time t.

$p_{ij,t}$ is converging to its limit $\overline{p}_{ij}$ with the speed $\frac{1}{\log(t)}$, i.e., $p_{ij,t} = \overline{p}_{ij}(1 - 1/\log(t+3))$. For the constant model, the sequence of link activation probabilities is constant throughout iterations, and for a given link $\{i, j\} \in E$, equals $\overline{p}_{ij}$.

To define the weight matrices W_t, we use the Metropolis weights, which were described in previous subsection. For the distributed detection problem we consider two Gaussian distributions $\mathcal{N}_0(-0.1, 1)$ and $\mathcal{N}_1(0.1, 1)$ and we estimate the probability of detection error of each sensor.

We plot the results in Figures 8 and 9. The curves plot the probability of error estimated through Monte Carlo simulations, and averaged over all sensors ($\widehat{P}^{av}_{e,t} = \frac{1}{N}\sum_i \widehat{P}^i_{e,t}$), versus the number of iterations in Figure 8, and the expected number of communications in Figure 9. The orange curve represents the time-varying and the blue one the constant model. We can see that the probability of error for the time-varying model is approaching the constant model when the number of iterations is increasing, but it is better (lower) plotted versus the number of expected communications. For any given number of expected communications, the time-varying model has a lower probability of error. For example, for target detection accuracy $10^{-}6$, the proposed scheme reduces the expected number of communications from about 1300 to 900, i.e., by approximately 30%, while incurring a negligible overhead iteration-wise.

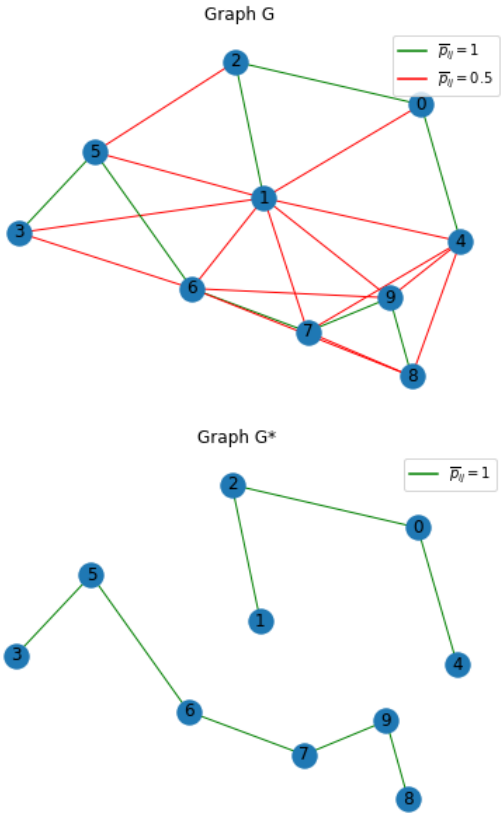


Fig. 7: Simulated graph G with 10 nodes and 23 links; The links with $\overline{p}_{ij} = 1$ colored in green, and links with $\overline{p}_{ij} = 0.5$ colored in red.

VII. Conclusion

We studied the large deviations rate of products of stochastic matrices, or the rate of consensus, for time-varying probabilities of the underlying topologies. We showed that, if the graph induced by the limiting probabilities contains a connected subgraph, then the rate of consensus is the best possible, and equals infinity. In the opposite case, the rate is given by the minimum edge cut of the limiting graph with costs defined through the limiting link failure probabilities. As a corollary to this result, we proved that the product of the weight matrices $\Phi(t, 1)$ converges to consensus matrix J almost surely. Theoretical findings are corroborated numerically by showing that the estimated exponential rate of probabilities in (4), obtained by Monte Carlo simulations, approaches the theoretical asymptotic limit.

References

[1] M. I. Jordan, J. D. Lee, and Y. Yang, "Communication-efficient distributed statistical inference," *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 668–681, 2019. [Online]. Available:

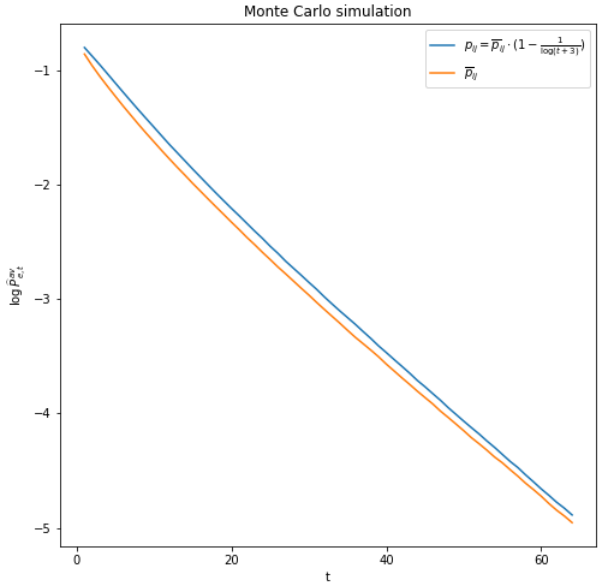


Fig. 8: Estimated probability (in the log scale) of error with constant activation probabilities (orange), time-varying activation probabilities (blue), versus the number of iterations t.

https://doi.org/10.1080/01621459.2018.1429274

[2] J. Fan, Y. Guo, and K. Wang, "Communication-efficient accurate statistical estimation," *Journal of the American Statistical Association*, vol. 0, no. 0, pp. 1–11, 2021. [Online]. Available: https://doi.org/10.1080/01621459.2021.1969238

[3] Y. Bao and W. Xiong, "One-round communication efficient distributed m-estimation," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 46–54. [Online]. Available: https://proceedings.mlr.press/v130/bao21a.html

[4] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Communication/computation tradeoffs in consensus-based distributed optimization," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'12. Red Hook, NY, USA: Curran Associates Inc., 2012, p. 1943–1951.

[5] O. Shamir, N. Srebro, and T. Zhang, "Communication-efficient distributed optimization using an approximate newton-type method," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1000–1008. [Online]. Available: https://proceedings.mlr.press/v32/shamir14.html

[6] X. Mao, K. Yuan, Y. Hu, Y. Gu, A. H. Sayed, and W. Yin, "Walkman: A communication-efficient random-walk algorithm for decentralized optimization," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2513–2528, 2020.

[7] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 1707–1718.

[8] N. Agarwal, A. T. Suresh, F. Yu, S. Kumar, and H. B. McMahan, "Cpsgd: Communication-efficient and differentially-private distributed sgd," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 7575–7586.

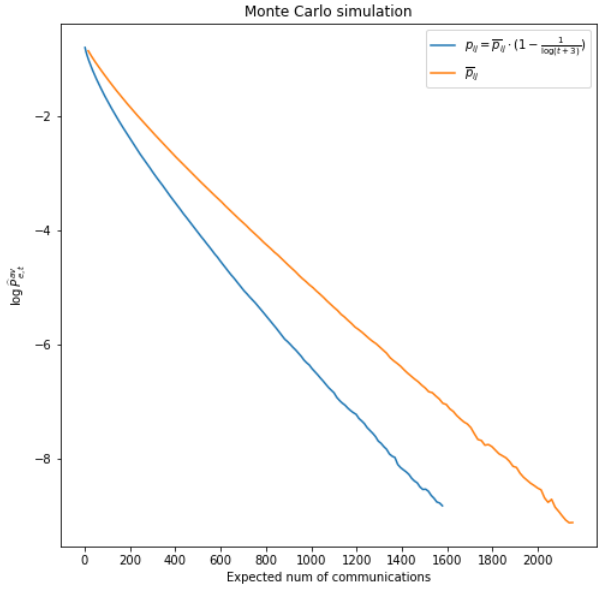


Fig. 9: Estimated probability of error (in the log scale) with constant activation probabilities (orange), time-varying activation probabilities (blue), versus the expected number of communications.

[9] B. Li, S. Cen, Y. Chen, and Y. Chi, "Communication-efficient distributed optimization in networks with gradient tracking and variance reduction," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 1662–1672.

[10] A. Mitra, J. Richards, S. Bagchi, and S. Sundaram, "Distributed inference with sparse and quantized communication," April 2020, arXiv, vol. abs/2004.01302v3.

[11] M. T. Toghani and C. A. Uribe, "Communication-efficient distributed cooperative learning with compressed beliefs," 2021. [Online]. Available: https://arxiv.org/pdf/2102.07767.pdf

[12] F. Alimisis, P. Davies, and D. Alistarh, "Communication-efficient distributed optimization with quantized preconditioners," in *International Conference on Machine Learning*. PMLR, 2021, pp. 196–206.

[13] T. Chen, G. B. Giannakis, T. Sun, and W. Yin, "Lag: Lazily aggregated gradient for communication-efficient distributed learning," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 5055–5065.

[14] S. Vargaftik, R. B. Basat, A. Portnoy, G. Mendelson, Y. Ben-Itzhak, and M. Mitzenmacher, "Communication-efficient federated learning via robust distributed mean estimation," *arXiv preprint arXiv:2108.08842*, 2021.

[15] A. Nedić, A. Olshevsky, and C. A. Uribe, "Distributed learning for cooperative inference," *arXiv preprint arXiv:1704.02718*, 2017.

[16] A. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Communication efficient distributed weighted non-linear least squares estimation," *EURASIP J. Adv. Signal Process.*, vol. 66, 2018.

[17] P. Braca, S. Marano, V. Matta, and A. H. Sayed, "Large deviations analysis of adaptive distributed detection," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 6112–6116.

[18] D. Bajovic, D. Jakovetic, A. K. Sahu, and S. Kar, "Large deviations for products of non-i.i.d. stochastic matrices with application to distributed detection," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1061–1065.

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

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

[21] Y. Gao, W. Liu, H. Wang, X. Wang, Y. Yan, and R. Zhang, "A review of distributed statistical inference," *Statistical Theory and Related Fields*, pp. 1–11, 2021.

[22] Y. Zhang, M. J. Wainwright, and J. C. Duchi, "Communication-efficient algorithms for statistical optimization," *Advances in neural information processing systems*, vol. 25, 2012.

[23] A. Mitra, J. A. Richards, and S. Sundaram, "A communication-efficient algorithm for exponentially fast non-bayesian learning in networks," in *2019 IEEE 58th Conference on Decision and Control (CDC)*, 2019, pp. 8347–8352.

[24] A. D. Sarwate and T. Javidi, "Distributed learning of distributions via social sampling," *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 34–45, 2015.

[25] A. Lalitha, X. Wang, O. C. Kilinc, Y. Lu, T. Javidi, and F. Koushanfar, "Decentralized bayesian learning over graphs," *ArXiv*, vol. abs/1905.10466, 2019.

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

[27] D. Bajović, D. Jakovetić, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Large deviations performance of consensus+innovations distributed detection with non-Gaussian observations," *IEEE Transactions on Signal Processing*, vol. 60, no. 11, pp. 5987–6002, Nov. 2012.

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

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

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

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

[32] Y. Hatano and M. Mesbahi, "Agreement over random networks," *IEEE Transactions on Automatic Control*, vol. 50, no. 11, pp. 1867–1872, 2005.

[33] M. Porfiri and D. J. Stilwell, "Consensus seeking over random weighted directed graphs," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1767–1773, 2007.

[34] M. T. Hale and M. Egerstedt, "Convergence rate estimates for consensus over random graphs," in *2017 American Control Conference (ACC)*, 2017, pp. 1024–1029.

[35] A. Tahbaz-Salehi and A. Jadbabaie, "Consensus over ergodic stationary graph processes," *IEEE Transactions on Automatic Control*, vol. 55, no. 1, pp. 225–230, 2010.

[36] I. Matei, J. S. Baras, and C. Somarakis, "Convergence results for the linear consensus problem under markovian random graphs," *SIAM Journal on Control and Optimization*, vol. 51, no. 2, pp. 1574–1591, 2013. [Online]. Available: https://doi.org/10.1137/100816870

[37] B. Touri and A. Nedić, "Product of random stochastic matrices," *IEEE Transactions on Automatic Control*, vol. 59, no. 2, pp. 437–448, 2014.

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

[39] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

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

[41] J. Wolfowitz, "Products of indecomposable, aperiodic, stochastic matrices," *Proceedings of the American Mathematical Society*, vol. 14, no. 5, pp. 733–737, 1963.

[42] A. Tahbaz-Salehi and A. Jadbabaie, "A necessary and sufficient condition for consensus over random networks," *IEEE Transactions on Automatic Control*, vol. 53, no. 3, pp. 791–795, 2008.

[43] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210 – 225, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0899825612000851

[44] D. Bajovic, "Inaccuracy rates for distributed inference over random networks with applications to social learning," 2022. [Online]. Available: https://arxiv.org/abs/2208.05236

[45] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[46] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 4238–4245.

[47] D. Bajovic, "Large deviations rates for distributed inference," PhD, Carnegie Mellon University, Pittsburgh, PA 15213, USA, May 2013.

[48] D. Bajovic, D. Jakovetic, and S. Kar, "Large deviations rates for stochastic gradient descent with strongly convex functions," 2022. [Online]. Available: https://arxiv.org/abs/2211.00969

[49] P. e. a. Kairouz, 2021.

[50] J. M. Ver Hoef, E. E. Peterson, M. B. Hooten, E. M. Hanks, and M.-J. Fortin, "Spatial autoregressive models for statistical inference from ecological data," *Ecological Monographs*, vol. 88, no. 1, pp. 36–59, 2018. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1283

[51] G. H. Hardy, *Divergent series*. American Mathematical Soc., 2000, vol. 334.

APPENDIX

Proof of the Remark 4: We will prove (5) by showing that the error matrix $\widetilde{\Phi}(t,1) = \Phi(t,1) - J$ converges almost surely to the matrix of all zeros $O_{N\times N}$. We do this by considering separately the cases when $G^\star$ is connected and when $G^\star$ is not connected.

Assume first that $G^\star$ is not connected. From Theorem 3, we have that for each $\zeta > 0$, there exists $t_0 = t_0(\zeta) \in \mathbb{N}$, such that for all $t \geq t_0$

$$\frac{1}{t}\ \log \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \leq -MC + \zeta,$$
$$\mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \leq e^{-t(MC-\zeta)}.$$

Summing over all $t \geq 1$, we get,

$$\sum_{t=1}^{+\infty} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right)$$
$$\leq \sum_{t=1}^{t_0-1} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) + \sum_{t=t_0}^{+\infty} e^{-t(MC-\zeta)}.$$

Since ζ can be chosen arbitrarily, we choose $\zeta = MC/2$, hence

$$\sum_{t=1}^{+\infty} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right)$$
$$\leq \sum_{t=1}^{t_0-1} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) + \sum_{t=t_0}^{+\infty} e^{-tMC/2}$$
$$\leq \sum_{t=1}^{t_0-1} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) + e^{-t_0 MC/2} \sum_{t=0}^{+\infty} e^{-tMC/2}$$
$$\leq \sum_{t=1}^{t_0-1} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) + \frac{e^{-t_0 MC/2}}{1-e^{-MC/2}} < +\infty.$$

Now, using Borel–Cantelli lemma, from the boundedness of $\sum_{t=1}^{+\infty} \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right)$, we have that the probability that only finitely many of the events $A_n^{(\epsilon)} = \{\|\widetilde{\Phi}(t,1)\| \geq \epsilon\}$ occur, is 1 for all $\epsilon \in (0,1]$. Hence $\mathbb{P}\{\lim_{t\to+\infty} \|\widetilde{\Phi}(t,1)\| = 0\} = 1$, and thus we have (5).

Suppose now that $G^\star$ is connected. From Theorem 3, we know that $\lim_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) = -\infty$. Thus, for each $M > 0$, there is $t_0 \in \mathbb{N}$, such that for all $t \geq t_0$

$$\frac{1}{t} \log \mathbb{P}\left(\|\widetilde{\Phi}(t,1)\| \geq \epsilon\right) \leq -M.$$

The rest of the proof is analogous to the proof in the case when $G^\star$ is not connected. □