Large Deviations for Products of Non-I.i.d. Stochastic Matrices with Application to Distributed Detection

Dragana Bajovic
Faculty of Technical Sciences
University of Novi Sad, Serbia
Email: dbajovic@uns.ac.rs

Dusan Jakovetic
Faculty of Sciences
University of Novi Sad, Serbia
Email: djakovet@uns.ac.rs

Anit Kumar Sahu and Soummya Kar
Carnegie Mellon University
Pittsburgh, PA, USA
Email: {anits, soummyak}@andrew.cmu.edu

***Abstract*—We derive the large deviation rate for convergence in probability of products of independent but not identically distributed stochastic matrices arising in time-varying distributed consensus-type networks. More precisely, we consider the model in which there exists a baseline topology that describes all possible communications and nodes are activated sparsely. At any given time, a node is active with a certain time-dependent probability, and any two nodes communicate if they are both active at that time. Under this model, we compute the exact rate for exponential decay of probabilities that the matrix products stay bounded away from their limiting matrix. We show that the rate is given by the minimal vertex cut of the baseline topology, where the node costs are defined by their limiting activation probabilities. The computed rate has many potential applications in distributed inference with intermittent communications. We provide an application in the context of *consensus+innovations* distributed detection. Therein, we show that optimal error exponent is achievable under a very general model of sparsified activations, thus effectively constructing asymptotically optimal detectors with significant communications savings.**

I. Introduction

We study convergence in probability of products of random, independent, but not identically distributed stochastic and symmetric matrices $\mathbf{W}_t$, where the topologies that underline the matrices have time-varying distributions. Specifically, we consider the model in which there exists a baseline graph describing all feasible communication links; nodes randomly activate over time, independently one from another, such that each node is active with a certain time-dependent probability, and two nodes communicate only if they are active at the same time. A major motivation for studying the products of stochastic matrices that underlie the described randomized time varying communication protocol is the recent work [1]; therein, it is shown that incorporating the described protocol into consensus+innovations distributed estimation (see, e.g., [2]) significantly improves the estimator's communication efficiency.

In this paper, our goal is to characterize for the described model of the $\mathbf{W}_t$'s the speed at which the probabilities that the product of the $\mathbf{W}_t$'s stays bounded away from its limiting matrix. More precisely, we are interested in computing

$$\mathcal{R} = -\lim_{t\to+\infty}\frac{1}{t}\log\mathbb{P}\left(\|\mathbf{W}_t\cdot\ldots\cdot\mathbf{W}_1-\mathbf{J}\|\geq\epsilon\right), \qquad (1)$$

for an arbitrary $\epsilon\in(0,1]$, provided that the limit in (1) exists[1]. Here, N is the number of network nodes, and the limiting matrix $\mathbf{J}=\mathbf{1}\mathbf{1}^\top/N$.While prior work derives rate (1) for products of independent, identically distributed (i.i.d.) matrices, the non-i.i.d. case of independent matrices with time-varying distributions has not been studied before. Quantity (1) is an important metric that has many potential applications in consensus+innovations-based distributed inference. For example, reference [3] (see also [4]) studies error exponents for Bayes error probability of consensus+innovations distributed detection under the i.i.d. matrices model. The reference shows that rate $\mathcal{R}$ critically determines detection error exponent.

It is well-known that, if the random (doubly stochastic) matrices $\mathbf{W}_t$ are i.i.d., then the product $\mathbf{W}_t\cdot\ldots\cdot\mathbf{W}_1$ converges almost surely to the consensus matrix $\mathbf{J}$ [4], and, since almost sure convergence implies convergence in probability, we thus have that the probabilities in (1) decay to zero. In our previous work [4], we show that this convergence is *exponentially fast*, and we computed the exact rate $\mathcal{R}$ for the i.i.d. matrices. In this paper, we consider the time varying randomized activation model described above for which the weight matrices are no longer identically distributed. We show that the limit in (1) continues to exist, and moreover we compute exactly the limit $\mathcal{R}$. Specifically, we show that $\mathcal{R}$ is given by the minimal vertex cut of the baseline graph, where the nodes' associated cut costs are defined by the nodes' limiting activation probabilities.

We demonstrate the significance of the studied non-i.i.d. matrix model and the derived rate $\mathcal{R}$ in the context of *consensus+innovations* distributed detection. More precisely, we consider a distributed detector with a randomized and time-varying sparsified communication protocol, where neighborhood communications are probabilistically sparsified in a time varying fashion with the goal of reducing the detector's communication cost. By utilizing result (1), we first show theoretically that the detector with time-varying and sparsified protocol can be designed to achieve asymptotic optimality at all signal-to-noise ratio (SNR) regimes; this is achieved when the activation probabilities converge to unity, possibly at a very

[1]As we show later, the limit in (1) exists and is independent of ϵ.

slow rate, e.g., as $1-\Omega(1/\log(t))$. In contrast, the previously studied i.i.d.-based detectors achieve asymptotic optimality only if the SNR exceeds a threshold. Therefore, effectively, we construct a "universally" asymptotically optimal detector that makes significant communication savings. Intuitively, by slowly increasing the node activation probabilities, we achieve the "optimality range" of the static protocol, but the rate of increase of probabilities is crafted carefully to achieve a reduced communication cost. Interestingly, while in [1] it is possible to decrease the activation probabilities over time and achieve an order optimal $O(1/t)$ estimation mean square error (MSE) decay, here the probabilities need to increase in time (though possibly very slowly). The intuition for this difference is that the "baseline process" here – the decay of error probability – is much faster (it behaves like e^{-ct}, $c>0$ a constant) than the "baseline process" in [1] – the rate of MSE decay, which behaves like $1/t$.

Despite being very challenging to prove, rate $\mathcal{R}$ has a clear intuitive interpretation. The time instants when the matrix product gets a step closer to the limiting matrix J are the time instants when the union graph of the topologies gets connected. It is then easy to see that the event (1) is feasible only if the number of these improvement times is sufficiently small. We show that the event in (1) is therefore equivalent to the event that the number of improvements is sublinear in t. The latter effectively corresponds to the scenario that the activated nodes fail to form a connected graph. The most likely way in which this can happen is given by the vertex cut with participating nodes which are the "easiest" to disconnect, i.e., with smallest limiting occurrence probabilities.

Products of stochastic matrices have been studied for a long time, e.g., [5], and the problem receives a continued interest, e.g., [6], [7], [8]. The problem of computing the exact large deviations rate in (1), arising, e.g., in the analysis of distributed detection [3], see also [9], [10], has been studied before in [4] for the i.i.d. matrices and in [11] for temporally dependent matrices, where the temporal dependence is modeled through a Markov chain. This paper complements the prior work by establishing the limit (1) for a class of time varying matrix distributions that have not been studied before. As explained above, the newly studied class has a significant relevance for distributed detection.

II. Model and preliminaries

Random nodes' activation and random matrix model. The network of nodes is modeled as an undirected graph $\overline{G}=(\overline{V},\overline{E})$, where $\overline{V}=\{1,2,...,N\}$ is the set of nodes, and $\overline{E}\subseteq\binom{\overline{V}}{2}$ is the set of all communication links between nodes. We assume that $\overline{G}$ is connected. During network operation, network nodes activate at random with certain probabilities that we assume are different for different nodes. To each node $i\in\overline{V}$, we associate, for each time $t=1,2,...,$ a Bernoulli random variable $\xi_{i,t}$, which is equal to 1 if i is active at time t, and otherwise is 0. Let $p_{i,t}=\mathbb{P}(\xi_{i,t}=1)\in(0,1)$ denote the probability that i is active at time t and let V_t collect all the nodes in $\overline{V}$ that are active at time t. For arbitrary two nodes $i,j\in\overline{V}$ to communicate at time t, it is necessary and sufficient that i and j are both active at that time.

Let G_t denote the graph obtained by collecting all the nodes that are active at time t, together with their induced communication links. More precisely, $G_t=(V_t,E_t)$, where the set of active edges at time t is given by

$$E_t=\left\{\{i,j\}\in\overline{E}:i,j\in V_t\right\}. \tag{2}$$

We make the following assumption on the nodes' activations and on the weight matrices $\mathbf{W}_t$.

Assumption 1 (Nodes' activation model)**.**

1) *For each $i\in\overline{V}$, $\xi_{i,t}$ and $\xi_{i,s}$ are independent for $t\neq s$, for any $t,s\geq 1$.*
2) *For each node i, $p_{i,t}$ increases monotonically with t according to the following formula:*

$$p_{i,t}=p_i(1-\alpha_t), \tag{3}$$

where $\alpha_t\in(0,1]$ is a monotonically decreasing sequence converging to 0, equal for all nodes, and $p_i\in(0,1]$ is the limiting activation probability of node i.

It is easy to see from Assumption 1 that the topologies G_t, $t\geq 1$, are independent. We further make the following assumptions on the weight matrices $\mathbf{W}_t$.

Assumption 2 (Weight matrices model)**.**

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

The rest of the section gives preliminaries needed to state and prove the main result on rate $\mathcal{R}$ calculation.

Union graph of topologies. We denote by $\Gamma(t,s)$ the random graph that collects the edges from all the graphs G_r that appeared from time $r=s+1$ to $r=t$, $s<t$, i.e.,

$$\Gamma(t,s):=\Gamma(\{G_t,G_{t-1},\ldots,G_{s+1}\}).$$

Similarly as with Γ, for any $s<t$ we analogously define the product matrix over time window $r=s+1$ to $r=t$,

$$\mathbf{\Phi}(t,s)=\mathbf{W}_t\mathbf{W}_{t-1}\cdot\ldots\mathbf{W}_{s+1}. \tag{4}$$

To simplify the notation, it is also of interest to introduce the error matrix $\widetilde{\mathbf{\Phi}}(t,s)=\mathbf{\Phi}(t,s)-\mathbf{J}$, a norm of which quantifies how close the product is to its limit $\mathbf{J}$.

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

$$T_i=\min\{t\geq T_{i-1}+1:\ \Gamma(t,T_{i-1})\text{ is connected}\},\ \text{for } i\geq 1, \tag{5}$$

with $T_0=0$. 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 $\widetilde{\mathbf{\Phi}}(t,s)$ drops below one (see, e.g., Lemma 11 in [4]). 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. Finally, for any fixed $t \geq 1$, we introduce the number of improvements until time t, denoted by M_t,

$$M_t = \max\{i \geq 0 : T_i \leq t\}. \tag{6}$$

Vertex cut. For an arbitrary graph $G = (V, E)$ the vertex cut is defined as any subset of graph nodes C such that the remaining graph $G_{\setminus C} := \left(V \setminus C, E_{\setminus C}\right)$ is not connected, where $E_{\setminus C} := \{\{i,j\} \in E : i, j \in V \setminus C\}$; or, in words, $G_{\setminus C}$ is the graph obtained from the initial graph G by removing from G all the vertices that belong to C and all the edges connected to these vertices. We denote the set of all vertex cuts of G by $\mathcal{C}(G)$.
If each node $i \in V$ is assigned a cost $c_i \in \mathbb{R}$, then, the minimal vertex cut is defined as a vertex cut $C \subseteq V$ such that the sum of costs of nodes in C is minimal among all vertex cuts $C \in \mathcal{C}(G)$. We denote the associated cost by

$$VC\left(G, \{c_i\}_{i\in V}\right) = \min_{C\in\mathcal{C}(G)} \left\{\sum_{i\in C} c_i\right\}. \tag{7}$$

III. Main result

We now state the main result of the paper – existence and characterization of the rate $\mathcal{R}$ via vertex cut.

Theorem 3. *Let Assumptions 1 and 2 hold. The rate of consensus $\mathcal{R}$ in (1) is then, for any $\epsilon \in (0, 1]$, given by*

$$\mathcal{R} = VC\left(G, \{-\log q_i\}_{i\in V}\right), \tag{8}$$

where $q_i = (1 - p_i)$.

Proof. We start with the following result from [4], which asserts that, if the number of improvements until time t scale linearly with t, for $t \geq 1$, then, starting from some finite time t, the events in (1) have zero probabilities (see Lemma 14, part 1 in [4]).

Lemma 4. *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*

$$\mathbb{P}\left(\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\| \geq \epsilon,\ M_t \geq \beta t\right) = 0, \quad \forall t \geq t_0(\beta, \epsilon). \tag{9}$$

Using the preceding result, it is easy to see that, for any fixed $\beta \in (0, 1)$, a necessary condition for $\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\| \geq \epsilon$ is that $M_t \leq \beta t$ (as otherwise the probability of this event is 0, which is asserted by Lemma 4). On the other hand, it is easy to see that the sufficient condition for this event to occur is that $M_t = 0$ (as in this case $\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\| = 1$, see Lemma 11 in [4]). Thus, we have that for each $\beta \in (0, 1)$, all $t \geq t_0(\beta, \epsilon)$,

$$\mathbb{P}\left(M_t = 0\right) \leq \mathbb{P}\left(\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\|\right) \leq \mathbb{P}\left(M_t < \beta t\right). \tag{10}$$

We prove the theorem by proving first that the left hand-side has an exponential decay rate equal to $\mathcal{R}$, as given in (8). We then show that the right hand-side probability in (10) decays with an β-dependent rate that gets closer to $\mathcal{R}$ as β decreases to 0, and in the limit equals $\mathcal{R}$.

We start by noting that the event $M_t = 0$ is equivalent to the event that $\Gamma(t, 0)$ is disconnected. Let $C^\star$ denote the minimal vertex cut of $\overline{G}$, where the node costs are assigned as in the claim of the theorem. It is easy to see that a sufficient condition for $\Gamma(t, 0)$ to be disconnected is that each of the nodes in the set $C^\star$ was inactive over the time interval from time 1 to time t. Thus,

$$\begin{aligned}\mathbb{P}\left(\Gamma(t,0)\,; \text{not connected}\right) &= \mathbb{P}\left(V_t \cap C^\star = \emptyset, k = 1, ..., t\right)\\ &= \prod_{k=1}^{t}\prod_{i\in C^\star}\left(1 - p_{i,k}\right)\\ &\geq \prod_{i\in C^\star}\left(1 - p_i\right)^t,\end{aligned} \tag{11}$$

where the second equality follows by the Assumption 1.1 and last inequality follows by the monotonic increase of the probabilities of nodes' activations, Assumption 1.2. Computing the logarithm, dividing by t and computing the limit $t \to +\infty$, we obtain from (11) and (10)

$$\liminf_{t\to+\infty} \frac{1}{t} \log \mathbb{P}\left(\left\|\widetilde{\mathbf{\Phi}}(t, 0)\right\| \geq \epsilon\right) \geq -\mathcal{R}. \tag{12}$$

We next turn to computing the exponential rate of the right hand-side in the inequality (10). We start by noting that

$$\begin{aligned}&\mathbb{P}\left(M_t < \beta t\right) = \sum_{m=0}^{\lceil\beta t\rceil - 1} \mathbb{P}\left(M_t = m\right)\\ &= \sum_{m=0}^{\lceil\beta t\rceil - 1} \sum_{1\leq t_1\leq\ldots\leq t_m\leq k} \mathbb{P}\left(T_l = t_l, \text{for } 1 \leq l \leq m,\ T_{m+1} > t\right),\end{aligned} \tag{13}$$

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

$$\mathbb{P}\left(T_l = t_l, \text{for } l \leq m,\ T_{m+1} > t\right) \tag{14}$$

$$\leq \mathbb{P}\left(\Gamma(t_l - 1, t_{l-1}) \text{ not connected, for } l \leq m+1\right)$$

$$= \prod_{l=1}^{m+1} \mathbb{P}(\Gamma(t_l - 1, t_{l-1}) \text{ not connected}) \tag{15}$$

where the last equality follows by the independence of the graph realizations. Note that, for arbitrary $t > s$, the event that the supergraph $\Gamma(t, s)$ is not connected can be represented as the union of events that an arbitrary vertex cut of $\overline{G}$ was absent from the random graphs G_k over time window $s < k \leq t$, i.e.,

$$\{\Gamma(t, s) \text{ not connected}\} = \cup_{C\in\mathcal{C}(G)} \left\{V_k \cap C = \emptyset,\ s < k \leq t\right\}. \tag{16}$$

Applying (16) to each of the intervals $t_{l-1} < k \leq t_l - 1$ and computing the probabilities, we get by the union bound

$$
\begin{aligned}
&\{\Gamma(t_l - 1, t_{l-1}) \text{ not connected}\} \\
&\leq \sum_{C \in \mathcal{C}(\overline{G})} \mathbb{P}\left(V_k \cap C = \emptyset,\ t_{l-1} \leq k \leq t_l - 1\right) \\
&= \sum_{C \in \mathcal{C}(\overline{G})} \prod_{t_{l-1} \leq k \leq t_l - 1} \prod_{i \in C} (1 - p_{i,k})\,. \qquad (17)
\end{aligned}
$$

Expressing $1 - p_{i,k} = (1 - p_i)\left(1 + \frac{p_i}{1 - p_i}\alpha_k\right) \leq (1 - p_i)(1 + \kappa\alpha_k)$, where $\kappa = \max_{i \in V} p_i/(1 - p_i)$, and using the fact that $1 + x \leq e^x$, we obtain from (17):

$$
\begin{aligned}
&\{\Gamma(t_l - 1, t_{l-1}) \text{ not connected}\} \\
&\leq \sum_{C \in \mathcal{C}(\overline{G})} e^{N\kappa \sum_{t_{l-1} < k \leq t_l - 1} \alpha_k} \prod_{i \in C} (1 - p_i)^(t_l - 1 - t_{l-1}) \\
&\leq \left|\mathcal{C}(\overline{G})\right| e^{N\kappa \sum_{t_{l-1} < k \leq t_l - 1} \alpha_k} e^{-(t_l - 1 - t_{l-1}) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)}. \qquad (18)
\end{aligned}
$$

Applying the preceding inequality for each time interval $t_{l-1} < k \leq t_l - 1$ yields in (14)

$$
\begin{aligned}
&\mathbb{P}\left(T_l = t_l, \text{for}\, l \leq m,\ T_{m+1} > t\right) \\
&\leq \left|\mathcal{C}(\overline{G})\right|^m e^{N\kappa \sum_{1 \leq k \leq t,\ k \neq t_i} \alpha_k} e^{-(k - m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} \\
&\leq \left|\mathcal{C}(\overline{G})\right|^m e^{N\kappa \sum_{k=1}^t \alpha_k} e^{-(k - m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)}. \qquad (19)
\end{aligned}
$$

The preceding bound holds for each of the terms in (13) that correspond to a fixed number of improvements $M_t = m$, and since there are $\binom{t}{m}$ possible allocations of times $T_1, ..., T_m$, we obtain

$$
\begin{aligned}
&\mathbb{P}\left(M_t = m\right) \\
&\leq \binom{t}{m} \left|\mathcal{C}(\overline{G})\right|^m e^{-(k - m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} e^{N\kappa \sum_{k=1}^t \alpha_k} \\
&\leq \left(\frac{te}{m}\right)^m \left|\mathcal{C}(\overline{G})\right|^m e^{-(k - m) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} e^{N\kappa \sum_{k=1}^t \alpha_k}. \qquad (20)
\end{aligned}
$$

It is easy to see that, for any $\beta < 1/2$, the preceding bound is maximal for $m = \beta t$. Therefore,

$$
\begin{aligned}
\mathbb{P}\left(M_t < \beta t\right) &\leq \beta t \left(\frac{te}{\beta t}\right)^{\beta t} \left|\mathcal{C}(\overline{G})\right|^{\beta t} \\
&\times e^{-(k - \beta t) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right)} e^{N\kappa \sum_{k=1}^t \alpha_k}. \qquad (21)
\end{aligned}
$$

Computing the logarithm, dividing by t, and taking the limit $t \to +\infty$ yields

$$
\begin{aligned}
&\limsup_{t \to +\infty} \mathbb{P}\left(M_t < \beta t\right) \leq \beta \log \frac{e}{\beta} + \beta \log |\mathcal{C}(\overline{G})| \\
&- (1 - \beta) VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right) + \kappa \lim_{t \to +\infty} \frac{\sum_{k=1}^t \alpha_k}{t}. \qquad (22)
\end{aligned}
$$

By Assumption 1.2, the last term vanishes, and taking the infimum with respect to $\beta > 0$, we finally obtain

$$
\limsup_{t \to +\infty} \mathbb{P}\left(M_t < \beta t\right) \leq -VC\left(\overline{G}, \{-\log q_i\}_{i \in V}\right). \qquad (23)
$$

Combining with (12), the claim of Theorem 3 follows. $\square$

IV. Application to distributed detection

We now demonstrate the usefulness of Theorem 3 by applying it to consensus+innovations distributed detection; see also [3], [12], [13]. The detection problem is as follows. Sensors in a N-node network cooperate to decide on a binary hypothesis, H_1 versus H_0. Each sensor i, at each time step t, $t = 1, 2, ...$, performs a measurement $Y_i(t)$; the measurements are i.i.d., both in time and across sensors, where under hypothesis H_l, $Y_i(t)$ has the density function f_l, $l = 0, 1$, for $i = 1, \ldots, N$ and $t = 1, 2, \ldots$. In the aforementioned detector, each sensor i maintains its local decision statistic $x_{i,t}$ and compares it with the zero threshold; if $x_{i,t} > 0$, sensor i accepts H_1; otherwise, it accepts H_0. At each time t, a sensor i updates its decision statistic $x_{i,t}$ by exchanging its decision statistic in the neighborhood and by assimilation of decision statistics from its neighborhood and its latest sensed information through a log-likelihood ratio $L_{i,t} = \log \frac{f_1(Y_{i,t})}{f_0(Y_{i,t})}$:

$$
x_{i,t} = \sum_{j \in O_{i,t}} W_{ij,t} \left(\frac{t-1}{t} x_{j,t-1} + \frac{1}{t} L_{j,t}\right), \qquad (24)
$$

with $x_{i,0} = 0$. Here $O_{i,t}$ is the (random) neighborhood of sensor i at time t (including i), and $W_{ij,t}$ is the (random) averaging weight that sensor i assigns to sensor j at time t. We let the $N \times N$ matrix $\mathbf{W}_t$ that collects the weights $W_{ij,t}$ in (24) adhere to the model in Assumption 2. In other words, sensors utilize a randomized communication protocol as described in Assumptions 1 and 2 and the preceding paragraphs. We additionally assume that sensors' observations are independent from the activation protocol, i.e., from matrices $\mathbf{W}_t$. To assess communication-wise benefits of the sparsified communication protocol, we benchmark detector (24) against the detector with the same algorithmic form as in (24), except that the weight matrix is replaced by a constant doubly stochastic matrix W. Intuitively, the benchmark utilizes communications across all links at all times, and it is hence natural to expect that it has a better performance with respect to time, i.e., with respect to the number of measurements processed. However, as shown ahead, the detector with sparsified communications practically matches the benchmark's performance time-wise while achieving a better performance communication-wise.

For detector (24), rate of consensus $\mathcal{R}$ plays a major role in its asymptotic performance, as measured by the worst-sensor error exponent of the Bayes error probability: $\min_{i=1,\ldots,N} \left\{-\frac{1}{t} \log\left(P^{\mathrm{e}}_{i,t}\right)\right\}$, where $P^{\mathrm{e}}_{i,t}$ is the Bayes error probability for sensor i at time t. While prior work [3] characterized asymptotic performance of detectors of form (24) when the weight matrices are deterministic or randomly varying in an i.i.d. fashion, Theorem 3 gives us the opportunity to characterize here the detection performance under the assumed sparsified time-varying protocol. Namely, it can be shown that, if the rate of consensus $\mathcal{R}$ in (1) satisfies: $\mathcal{R} \geq (N-1) C_{\mathrm{tot}}$, where $C_{\mathrm{tot}} = C_{\mathrm{tot}}(N, f_1, f_0)$ is the exponential decay rate of the error probability of the best centralized detector (Chernoff information), then distributed detector (24) is asymptotically optimal, i.e., it achieves the best possible error exponent. (See [3] for details.)

We now comment on the achieved result. It is known that a detector of the form (24) is asymptotically optimal under either deterministic or under an i.i.d. weight matrix model, provided that $\mathcal{R}$ exceeds $(N-1)C_{\mathrm{tot}}$. Here, we show that asymptotic optimality is still achievable under the time-varying weight model satisfying Assumption 2 with slowly increasing node activation probabilities. This result has, for example, the following implication on improving communication efficiency in distributed detection: a detector of form (24) with the assumed time-varying randomized communication protocol – wherein the activation probabilities are slowly increasing to unity – is asymptotically equivalent to the detector with a constant weight matrix W. Hence, as the detector with the randomized protocol has a lower communication cost while essentially equivalent performance time-wise, one can expect that it becomes more communication efficient. We next present a simulation example that confirms such improvements in communication efficiency.

We consider a geometric network with $N = 20$ sensors. We place the sensors uniformly over a unit square, and connect those sensors whose distance d_{ij} is less than a radius. The total number of (undirected) links is 63. For the sensors' measurements, we use the Gaussian distribution $f_1 \sim \mathcal{N}(m, \sigma^2)$, $f_0 \sim \mathcal{N}(0, \sigma^2)$, with $m = 0.01$, and $\sigma^2 = 0.2$. We consider two different detectors of form (24). The first one is the benchmark for which each link is online at all times. The second detector utilizes at each node i activation probability $p_t = 1 - 1/\log(t+2)$, $t = 1, 2, \ldots$. For the averaging weights of the benchmark, we use for each link $\{i, j\}$ a constant weight $W_{ij} = 0.1$. To compensate for random activations, the second detector assigns weight $0.1/p_t$ whenever a link is online. Figure 1 (top) plots the simulated Bayes error probability versus per node communication cost, averaged across nodes and across $20,000$ Monte Carlo algorithm runs. We can see that the detector with time-varying sparsified communications (solid line) achieves significant communication savings with respect to the benchmark, while at the same time practically matches the benchmark's performance with respect to time (see Figure 1, bottom).

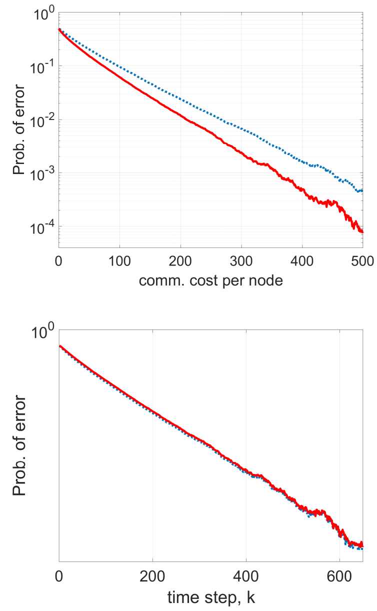

Fig. 1: Estimated probability of error (in the $\log$ scale) versus per node communication cost (top) and versus time (bottom) for the benchmark detector (blue dotted line) and the detector with sparsifying communications (red solid line).

V. Conclusion

We have derived the exact large deviations rate for products of a class of non-i.i.d. random stochastic and symmetric matrices that arise with distributed inference under randomized communication protocols. We applied the results to consensus+innovations distributed detection to derive universally asymptotically optimal detectors with significantly reduced communication cost.

References

[1] A. K. Sahu, D. Jakovetic, and S. Kar, "Communication optimality trade-offs in distributed estimation," 2018, arxiv preprint. [Online]. Available: http://users.ece.cmu.edu/~anits/comm_ci_jmlr_prefinal.pdf

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

[3] 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.

[4] 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.

[5] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, pp. 118–121, 1974.

[6] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on Automatic Control*, vol. 49, no. 9, p. 15201533, Sept. 2004.

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

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

[9] P. Braca, S. Marano, V. Matta, and P. Willet, "Asymptotic optimality of running consensus in testing binary hypothesis," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 814–825, Feb. 2010.

[10] F. Cattivelli and A. Sayed, "Distributed detection over adaptive networks using diffusion adaptation," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 1917–1932, May 2011.

[11] D. Bajović, J. Xavier, and B. Sinopoli, "Products of stochastic matrices: large deviation rate for markov chain temporal dependencies," in *Allerton'12, 50th Allerton Conference on Communication, Control, and Computing*, Monticello, Il, October 2012.

[12] 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.

[13] A. K. Sahu and S. Kar, "Recursive distributed detection for composite hypothesis testing: Nonlinear observation models in additive gaussian noise," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 4797–4828, 2017.