Products of Stochastic Matrices: Large Deviation Rate for Markov Chain Temporal Dependencies

Dragana Bajović, João Xavier, and Bruno Sinopoli

***Abstract*— We find the large deviation rate I for convergence in probability of the product $W_k \cdots W_1 W_0$ of temporally dependent random stochastic matrices. As the model for temporal dependencies, we adopt the Markov chain whose set of states is the set of all possible graphs that support the matrices W_k. Such model includes, for example, 1) token-based protocols, where a token is passed among nodes to determine the order of processing; and 2) temporally dependent link failures, where the temporal dependence is modeled by a Markov chain. We characterize the rate I as a function of the Markov chain transition probability matrix P. Examples further reveal how the temporal correlations (dependencies) affect the rate of convergence in probability I.**

I. Introduction

We study large deviations rates for products of *temporally dependent*, random, symmetric, stochastic matrices $\{W_t\}_{t\geq 0}$. Products of stochastic matrices arise, e.g., in the analysis of consensus or gossip algorithms, e.g., [1], [2], [3], consensus+innovations algorithms, e.g., [4], [5] and diffusion algorithms, e.g., [6]. It is well-known that the product $W_k W_{k-1} \cdots W_0$ converges in probability to $J = (1/N)11^\top$, if the second largest (in modulus) eigenvalue of $\mathbb{E}\left[W_k\right]$ is strictly less than one and W_t have positive diagonals. We have recently computed [7] the *large deviation rate* for the convergence in probability (here $\|\cdot\|$ denotes the spectral norm):

$$I := \lim_{k\to\infty} -\frac{1}{k}\log \mathbb{P}\left(\|W_k W_{k-1}\cdots W_0\| \geq \epsilon\right),\ \epsilon \in (0,1), \tag{1}$$

when the sequence $\{W_t\}_{t\geq 0}$ is *independent identically distributed* (i.i.d.) The quantity I in (1) is an important metric and appears naturally in studying consensus+innovations distributed inference, e.g., distributed detection [4]. Reference [4] shows that performance of consensus+innovations distributed detector significantly depends on I and in a highly nonlinear way. (See [4] for details.)

In this paper, we calculate the rate I in (1) for a random model with temporally dependent matrices W_t. Specifically, we consider the model where the underlying graphs G_t – the graphs that support the matrices W_t – follow a Markov chain model. We associate to each realizable graph (out of M graphs) a state of the Markov chain; the distribution of G_t is then determined by a specified $M \times M$ transition probability matrix P. This model subsumes, e.g., the token-based protocols similar to [8], or temporally dependent link failure models, where the on/off state of each link follows a Markov chain.

We characterize the rate I in (1) as a function of the transition probability matrix P. We refer to Theorem 4 for details, but here we convey the general idea. Namely, we find that the probability that the Markov chain stays for t time steps in a sub-collection $\mathcal{H}$ of graphs behaves as $\left(\rho(P_{\mathcal{H}})\right)^t$, where $\rho(P_{\mathcal{H}})$ is the spectral radius of the submatrix of P associated with the sub-collection $\mathcal{H}$. Then, the rate I equals $\rho(P_{\mathcal{H}^\star})$, where $\mathcal{H}^\star$ is the sub-collection whose graphs are disconnected in union, and that gives the largest spectral radius $\rho(P_{\mathcal{H}^\star})$. Further, we illustrate the results on two examples, namely gossip with Markov dependencies and temporally correlated link failures. The example with temporally correlated link failures shows that "negative temporal correlations" of the links' states (being ON or OFF) increase (improve) the rate I when compared with the uncorrelated case, while positive correlations decrease (degrade) the rate. This result is in accordance with standard large deviations results on temporally correlated sequences, see, e.g., [[9], exercise V.12, page 59.]

Paper organization. The next paragraph introduces notation. Section II describes the problem setup and states our main result on the rate I. Section III proves the result. Section IV gives two examples, the gossip with Markov dependencies, and temporally correlated link failures. Finally, Section V 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; $\rho(A)$ the spectral radius of A; 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 ith entry equal to 1 and the rest being zeros), respectively. Further, for a vector a, the inequality $a > 0$ is understood component wise. Given a selection $S \subseteq \{1,...,N\}$ of rows and columns of a matrix

Work of Dragana Bajović, João Xavier and Bruno Sinopoli is partially supported by grants CMU-PT/SIA/0026/2009 and SFRH/BD/33517/2008 (through the Carnegie Mellon/Portugal Program managed by ICTI) from Fundação para a Ciência e Tecnologia and also by ISR/IST plurianual funding (POSC program, FEDER). Dragana Bajović holds fellowship from the Carnegie Mellon/Portugal Program.

Dragana Bajović is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal, and with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA dragana@isr.ist.utl.pt, dbajovic@andrew.cmu.edu

João Xavier is with the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), Technical University of Lisbon, Lisbon, Portugal jxavier@isr.ist.utl.pt

Bruno Sinopoli is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA brunos@ece.cmu.edu

A: $\{A_l : l \in S\}$ and $\{A^l : l \in S\}$, we denote by A_S the submatrix of A corresponding to the selection S. Similarly, if S is a selection of rows, we denote by A_{Sl} the part of A^l that corresponds to the selection S. Likewise, for the selection of columns S, we denote by A_{lS} the part of A_l that corresponds to S.

II. Problem setup and statement of main result

In this section, we state the assumptions of the paper in Subsection II-A, introduce several key concepts needed for our analysis in Subsection II-B, and state the main result on the large deviation rate I in Subsection II-C.

A. Assumptions

We study the sequence of random matrices $\{W_t\}_{t\geq 0}$, $W_t : \Omega \mapsto \mathbb{R}^{N\times N}$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (W_t is $(\mathcal{F}, \mathcal{B}(\mathbb{R}^{N\times N}))$-measurable, for all $t \geq 1$). We denote by $\omega \in \Omega$ an arbitrary element of Ω. We make the following assumption on the random matrices W_t.

Assumption 1 (Matrices W_t) 1) $W_t(\omega)$ is symmetric and stochastic (row sums equal 1 and the entries are nonnegative), $\forall \omega$.
Further, there exists $\delta > 0$ such that the following two conditions are satisfied.
2) $[W_t(\omega)]_{ii} \geq \delta$, $\forall i, \forall t, \forall \omega$.
3) The entries $[W_t(\omega)]_{ij}$ are bounded away from zero whenever positive, that is, for any i, j, t, ω $[W_t(\omega)]_{ij} \geq \delta$ whenever $[W_t(\omega)]_{ij} > 0$.

Thus, we assume that $W_t(\omega)$ has positive diagonals and, also, that the positive entries of $W_t(\omega)$ are bounded away from zero, for all ω. Throughout, for a fixed ω, we refer to $W_t(\omega)$ as a realization. To explain the temporal dependence in the matrix sequence, we need to introduce the sequence of graphs that underly the matrices W_t.

For a stochastic symmetric N by N matrix W, we define its induced graph, denoted by $G(W)$, by

$$G(W) = \left\{ V, \left\{ \{i,j\} \in \binom{V}{2} : W_{ij} > 0,\ i \neq j \right\} \right\}, \quad (2)$$

where $V = \{1, ..., N\}$. Thus, $G(W)$ is a graph on N nodes, without self-loops, with edges between those nodes i and j for which the entry W_{ij} is positive. As W is symmetric, $G(W)$ is undirected. We denote by G_t the induced graph of the random matrix W_t, for $t = 0, 1, \ldots$, i.e., $G_t = G(W_t)$. Intuitively, the graph G_t underlying W_t at some time $t \geq 0$ is the collection of all communication links that are active at time t. Also, as W_t can assume realizations with different underlying graphs, G_t is in general random. (Formally, G_t is a random map from Ω to the set of all graphs on N vertices.)

Consider now the sequence of random graphs $\{G_t\}_{t\geq 0}$. Let $\mathcal{G} = \{H_1, \ldots, H_M\}$ be the minimal set of subgraphs of $\left(V, \binom{V}{2}\right)$ (on the same set of vertices V)[1] such that $G_t(\omega) \in \mathcal{G}$, for all ω and all t. Then, the induced graph $G_t = G(W_t)$ along the sequence $\{W_t\}_{t\geq 0}$ draw their realizations from $\mathcal{G}$.

Graph temporal dependence. We consider temporally correlated matrices W_t. We encode the temporal dependence through random graphs G_t, by assuming that the process G_t follows a Markov chain model. To each H_l, $l = 1, ..., M$, we associate a state (indexed by l) of a (finite state) Markov chain; the corresponding M by M transition matrix we denote by P.

Assumption 2 (Graph temporal dependence) There exist a nonnegative matrix $P \in \mathbb{R}^{M\times M}$ and a nonnegative vector $v \in \mathbb{R}^M$ satisfying $\sum_{m=1}^{M} P_{lm} = 1$ for all $l = 1, \ldots, M$ and $\sum_{l=1}^{M} v_l = 1$, such that for all t and all $l_0, \ldots, l_t \in \{1, \ldots, M\}$

$$\mathbb{P}\left(G_0 = H_{l_0}, G_1 = H_{l_1}, \ldots, G_t = H_{l_t}\right) = v_{l_0} P_{l_0 l_1} \cdots P_{l_{t-1} l_t}.$$

We assume in the sequel that $v > 0$[2]. Examples of the model are considered in Section IV.

B. Key objects

This subsection recalls the concepts from [7] that are our main analytical tools in the computation of I.

Random supergraph $\Gamma(s,t)$ and the error matrix $\widetilde{\Phi}(s,t)$. For a collection of graphs $\mathcal{H} \subseteq \mathcal{G}$, let $\Gamma(\mathcal{H})$ denote the graph that contains all edges from all graphs in $\mathcal{H}$. That is, $\Gamma(\mathcal{H})$ is the minimal graph (i.e., the graph with the minimal number of edges) that is a supergraph of every graph in $\mathcal{H}$:

$$\Gamma(\mathcal{H}) := (V, \bigcup_{G\in\mathcal{H}} E(G)), \quad (3)$$

where $E(G)$ denotes the set of edges of graph G.

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

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

Denote $\Phi(s,t) := W_s W_{s-1} \cdots W_t$, and $\widetilde{\Phi}(s,t) := \Phi(s,t) - J$, for $s \geq t \geq 0$. The norm of $\widetilde{\Phi}(s,t)$ characterizes the distance of the product $W_s \cdots W_t$ from the perfect consensus matrix J and we call $\widetilde{\Phi}(s,t)$ the error matrix. We have the following lemma that relates the supergraph $\Gamma(s,t)$ and the error matrix $\widetilde{\Phi}(s,t)$. The proof of Lemma 1 can be found in [7].

Lemma 1 For all $s \geq t \geq 0$, $\left\|\widetilde{\Phi}(s,t)\right\| < 1$ if and only if $\Gamma(s,t)$ is connected. Furthermore, if $\Gamma(s,t)$ is connected, then

$$\left\|\widetilde{\Phi}(s,t)\right\| \leq \left(1 - c\delta^{2(s-t+1)}\right)^{\frac{1}{2}}, \quad (4)$$

where $c = 2(1 - \cos\frac{\pi}{N})$ is the Fiedler value (the second smallest eigenvalue λ_F) of the path graph on N vertices, i.e., the minimum of $\lambda_{\mathrm{F}}(L(G)) > 0$ over all Laplacians $L(G)$ of connected graphs G on N vertices [10].

[1] As $\binom{V}{2}$ is a finite set, such a minimal set of subgraphs $\mathcal{G}$ exists.

[2] The large deviation result from Section III holds also under a less restrictive condition that the Markov chain can start in every initial (source) communication class with positive probability. The result for this case is omitted due to lack of space and will be pursued elsewhere.

Lemma 1 says that the norm of the error matrix drops exactly at the times when the supergraph $\Gamma(s,t)$ becomes connected. This is an intuitive result, as it simply says that there cannot be any improvement in the error until a network-wide information exchange has occurred.

The event whose probability determines the rate I, as we show in the proof of Theorem 4, is the event in which the supergraph $\Gamma(s,t)$ stays disconnected over a long time interval. We compute this probability by relating it with the concept of disconnected collection, which we introduce next.

Definition 2 A collection $\mathcal{H} \subseteq \mathcal{G}$ is a disconnected collection on $\mathcal{G}$ if its supergraph $\Gamma(\mathcal{H})$ is disconnected.

Thus, a disconnected collection is any collection of realizable graphs such that the union of all the edges of its graphs yields a disconnected graph. We also define the set of all possible disconnected collections on $\mathcal{G}$:

$$\Pi(\mathcal{G}) = \{\mathcal{H} \subseteq \mathcal{G} : \mathcal{H} \text{ is a disconnected collection on } \mathcal{G}\}.$$

Observation 3 If for some realization of the graph sequence the supergraph $\Gamma(s,t)$ is disconnected for some $0 \le t \le s$, then there must exist $\mathcal{H} \in \Pi(\mathcal{G})$ such that $G_r \in \mathcal{H}$, for all $t \le r \le s$.

C. Statement of the main result

Theorem 4 Consider the sequence $\{W_t\}_{t\geq 0}$ of stochastic symmetric matrices satisfying Assumptions 1 and 2. Then, for any $\epsilon \in (0,1]$

$$\lim_{k\to+\infty} \frac{1}{k} \log \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) = -I, \tag{5}$$

where

$$I = \begin{cases} |\log \rho_{\max}|, & \text{if } \Pi(\mathcal{G}) \neq \emptyset \\ +\infty, & \text{otherwise} \end{cases}, \tag{6}$$

and $\rho_{\max} = \max_{\mathcal{H}\in\Pi(\mathcal{G})} \rho\left(P_{\mathcal{H}}\right)$.

III. Proof of Theorem 4

This section proves Theorem 4: Subsection III-B proves the large deviation lower bound, while Subsection III-C proves the upper bound. Before the large deviation bounds, Subsection III-A derives certain intermediate results on the Markov chain temporal correlation model.

A. Intermediate results

The following lemma explains how to estimate the probability of the event that $\Gamma(s,t)$ is disconnected using the probability that the graph realizations remain in a certain disconnected collection $\mathcal{H} \in \Pi(\mathcal{G})$.

Lemma 5 For every $\mathcal{H} \in \Pi(\mathcal{G})$ and every $0 \le t \le s$

$$\mathbb{P}(G_r \in \mathcal{H},\ t \le r \le s) \le \mathbb{P}(\Gamma(s,t) \text{ is disconnected}) \le \sum_{\mathcal{H}\in\Pi(\mathcal{G})} \mathbb{P}(G_r \in \mathcal{H},\ t \le r \le s). \tag{7}$$

Proof: By Observation 3 we have

$$\{\Gamma(s,t) \text{ is disconnec.}\} = \left\{\bigcup_{\mathcal{H}\in\Pi(\mathcal{G})} \{G_r \in \mathcal{H},\ t \le r \le s\}\right\}. \tag{8}$$

Applying now the union bound to the probability of the union of events in (8), the right hand side inequality in (7) follows. The left hand side inequality follows by bounding the probability of the union of events by the probability of a single event from the union. ■

Lemma 6 For $s > t \geq 0$, $\mathcal{S} \subseteq \mathcal{G}$ and $1 \le l, m \le M$, if $\mathbb{P}\left(G_t = H_l\right) > 0$,

$$\mathbb{P}\left(G_r \in \mathcal{S}, t+1 \le r \le s, G_{s+1} = H_m | G_t = H_l\right) = P_{l\mathcal{S}} P_{\mathcal{S}}^{s-t-1} P_{\mathcal{S}m}. \tag{9}$$

We omit the proof of Lemma 6 for brevity. The lemma is a simple result for Markov chains: if we start from the state H_l at time t, end up in the state H_m at time $s+1$, and we restrict the trajectory ($G_r,\ t+1 \le r \le s$) to belong to a subset of states $\mathcal{S}$, then the corresponding probability is determined by the submatrix $P_{\mathcal{S}}$.

Lemma 7 Let $A \in \mathbb{R}^{N\times N}$ be a nonnegative matrix. For every $\varsigma > 0$, there exists C_ς such that for all $t \geq 1$

$$\rho(A)^t \le 1^\top A^t\, 1 \le C_\varsigma \left(\rho(A) + \varsigma\right)^t. \tag{10}$$

Proof of Lemma 7 is omitted for brevity; the lemma essentially follows from the simple relation $\|A^t\|_1 \le 1^\top A^t\, 1 \le N\|A^t\|_1$ and Gelfand's formula [11].

Corollary 8 follows by combining the results of Lemmata 5-7.

Corollary 8 For any $\varsigma > 0$, there exists $\overline{C}_\varsigma$ such that, for all $1 \le t \le s$ and H_m such that $\mathbb{P}(G_{t-1} = H_m) > 0$,

$$\mathbb{P}\left(\Gamma(s,t) \text{ is disconnected} | G_{t-1} = H_m\right) \le |\Pi(\mathcal{G})| \overline{C}_\varsigma (\rho_{\max} + \varsigma)^{s-t}.$$

Proof: Combining the results of Lemma 5 and Lemma 6 we get

$$\mathbb{P}\left(\Gamma(s,t) \text{ is disconnected} | G_{t-1} = H_m\right) \le \sum_{\mathcal{H}\in\Pi(\mathcal{G})} 1^\top P_{\mathcal{H}}^{s-t} 1, \tag{11}$$

where we used the simple bound $P_{m\mathcal{H}} P_{\mathcal{H}}^{s-t} 1 \le 1^\top P_{\mathcal{H}}^{s-t} 1$. Now, by Lemma 7, for any $\varsigma > 0$ and $\mathcal{H} \in \Pi(\mathcal{G})$ there exists $C_{\mathcal{H},\varsigma}$ such that $1^\top P_{\mathcal{H}}^{s-t} 1 \le C_{\mathcal{H},\varsigma} (\rho(P_{\mathcal{H}} + \varsigma))^{s-t}$. Bounding each term in the sum in (11) by $\overline{C}_\varsigma (\rho_{\max} + \varsigma)^{s-t}$, where $\overline{C}_\varsigma = \max_{\mathcal{H}\in\Pi(\mathcal{G})} C_{\mathcal{H},\varsigma}$, yields the claim. ■

We next prove Theorem 4 for the case when $\Pi(\mathcal{G}) \neq \emptyset$ by showing the upper and the lower large deviation bound:

$$\liminf_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) \geq \log \rho_{\max} \tag{12}$$

$$\limsup_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) \le \log \rho_{\max}. \tag{13}$$

We prove the lower bound (12) in Subsection III-B and the upper bound (13) we prove in Subsection III-C. The proof for the case when $\Pi(\mathcal{G}) = \emptyset$ can be derived by using similar arguments to the ones given in [7] (when $\Pi(\mathcal{G}) = \emptyset$) and is omitted here.

B. Lower bound

The proof of the lower bound is based on the intuitive fact that the norm of the error matrix will stay equal to 1 and never drop (and thus remain above ϵ), if the sequence of graphs G_t continues to be drawn in time from some disconnected collection $\mathcal{H} \in \Pi(\mathcal{G})$. By Lemma 6, the probability of the event that the Markov chain reduces to a subset of states $\mathcal{S} \subseteq \mathcal{G}$ is determined by the submatrix of the transition matrix corresponding to this subset of states $\mathcal{S}$. Next lemma combines these two facts to derive a family of lower bounds, indexed by $\mathcal{H} \in \Pi(\mathcal{G})$, on the probability of the event of interest.

Lemma 9 For any $\mathcal{H} \in \Pi(\mathcal{G})$

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) \geq v_{\mathcal{H}}^{\top} P_{\mathcal{H}}^{k} 1. \tag{14}$$

Proof: Using the result of Lemma 1, we have that for any fixed realization of the sequence $\{W_t\}_{t\geq 0}$ and for any $k \geq 1$ $\|\widetilde{\Phi}(k,0)\| \geq 1$ if and only if $\Gamma(k,0)$ is disconnected, and so

$$\mathbb{P}\left(\|\widetilde{\Phi}(k,0)\| \geq \epsilon\right) \geq \mathbb{P}\left(\|\widetilde{\Phi}(k,0)\| \geq 1\right) = \mathbb{P}\left(\Gamma(k,0) \text{ is disconnected}\right).$$

Applying Lemma 5 for fixed $\mathcal{H} \in \Pi(\mathcal{G})$ together with Lemma 6 yields

$$\mathbb{P}\left(\Gamma(k,0) \text{ is disconnected}\right) \geq \mathbb{P}\left(G_0 \in \mathcal{H}, G_1 \in \mathcal{H}, \ldots, G_k \in \mathcal{H}\right) = v_{\mathcal{H}}^{\top} P_{\mathcal{H}}^{k} 1.$$

■

Combining the result of Lemma 9 with Lemma 7, we get

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) \geq v_{\min} 1^{\top} P_{\mathcal{H}}^{k} 1 \geq v_{\min} \rho^{k}\left(P_{\mathcal{H}}\right),$$

where $v_{\min} = \min_{1\leq l\leq M} v_l > 0$. Taking the log, dividing by k, and taking the $\liminf$ over $k \to +\infty$ yields

$$\liminf_{k\to\infty} \frac{1}{k} \log \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) \geq \log \rho\left(P_{\mathcal{H}}\right).$$

Finally, since the previous bound holds for arbitrary $\mathcal{H} \in \Pi(\mathcal{G})$, we obtain the best bound by finding the maximal spectral radius $\rho\left(P_{\mathcal{H}}\right)$ over all $\mathcal{H} \in \Pi(\mathcal{G})$. This completes the proof of the lower bound.

C. Upper bound

As we have seen in the previous subsection, to derive the lower bound (12), we only had to consider the event that $\Gamma(k,0)$ remains disconnected when $k \to \infty$. This argument is no longer applicable for the upper bound (13) and one has to consider all possible realizations of the matrix sequence that satisfy $\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon$. We handle the combinatorial nature of the upper bound through the sequence of stopping times T_i which we introduced in [7].

Sequence of stopping times. Define the sequence of stopping times T_i, $i = 1, 2, \ldots$ by:

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

and $T_0 = 0$. The sequence $\{T_i\}_{i\geq 1}$ registers the times along the graph sequence when the union graph becomes connected, and, equivalently, when the norm of the error matrix drops.

We also borrow from [7] the increasing sequence M_k which counts the number of improvements along the graph sequence. For fixed time $k \geq 1$, we define M_k to be the number of improvements until time k, which can be expressed through stopping times T_i by:

$$M_k = \max\{i \geq 0 : T_i \leq k\}. \tag{15}$$

The idea that we use to solve the upper bound is to partition the probability space according to the number of improvements M_k. Namely, for fixed $\alpha \in (0,1)$ we consider separately two disjoint events, $\{M_k > \lceil \alpha k \rceil\}$ and $\{M_k \leq \lceil \alpha k \rceil\}$, and apply the law of total probability:

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) = \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon, M_k > \lceil \alpha k \rceil\right) + \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon, M_k \leq \lceil \alpha k \rceil\right). \tag{16}$$

We will show that, conditioned on the first event $\{M_k > \lceil \alpha k \rceil\}$, our event of interest $\left\{\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right\}$ has probability zero for sufficiently large k. In other words, we will show that, if the number of improvements until time k is on the order of k, then it is not possible that the norm of the error matrix stays above any fixed number $\epsilon > 0$ as k increases, and thus the first term in (16) becomes firm zero for k sufficiently large. This is formally stated in Lemma 10, the proof of which can be found in [7][3]. We only remark here that the key property from which this result follows is Lemma 1.

Lemma 10 For every $\alpha \in (0,1)$, there exists $k_0 = k_0(\alpha, \epsilon)$ such that for every $k \geq k_0$

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon, \; M_k > \lceil \alpha k \rceil\right) = 0.$$

Thus, the first term in (16) vanishes for $k \geq k_0$, which further yields for $k \geq k_0$

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) = \mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon, M_k \leq \lceil \alpha k \rceil\right) \leq \mathbb{P}\left(M_k \leq \lceil \alpha k \rceil\right).$$

We now focus on computing the probability of the event $\{M_k \leq \lceil \alpha k \rceil\}$. By definition of M_k, this event is the same as the event $\left\{T_{\lceil \alpha k \rceil + 1} > k\right\}$, implying that, for $k \geq k_0$,

$$\mathbb{P}\left(\left\|\widetilde{\Phi}(k,0)\right\| \geq \epsilon\right) \leq \mathbb{P}\left(T_{\lceil \alpha k \rceil + 1} > k\right). \tag{17}$$

[3] Although reference [7] considers the case of independent and identically distributed random matrices W_t, the result of Lemma 10 holds for arbitrary random model.

We give without proof, which is left for a companion journal paper, the following lemma on the asymptotic behavior of the probability in the right hand side of (17). We remark here that the proof can be derived from the exponential Markov inequality and from the property that the increments of the stopping times $T_{i+1} - T_i$ are independent given realizations of random graphs G_{T_i} (at the stopping times T_i), $i = 1, ..., \lceil \alpha k \rceil$.

Lemma 11 For every $\varsigma > 0$ and $\lambda < |\log(\rho_{\max} + \varsigma)|$

$$\frac{1}{k}\log \mathbb{P}\left(T_{\lceil \alpha k\rceil + 1} > k\right) \leq \frac{\lceil \alpha k \rceil + 1}{k}\left(\log(|\mathcal{G}||\Pi(\mathcal{G})|\overline{C}_\varsigma) + |\log\left(\rho_{\max}+\varsigma\right)| + \log\frac{e^{-(|\log(\rho_{\max}+\varsigma)|-\lambda)}}{1-e^{-(|\log(\rho_{\max}+\varsigma)|-\lambda)}}\right) - \lambda \quad (18)$$

where $\overline{C}_\varsigma$ is given in Corollary 8.

Consider now eq. (18) for fixed $\varsigma > 0$. Taking the limit $k \to +\infty$, and then inf over $\alpha > 0$, yields

$$\inf_{\alpha>0}\lim_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(T_{\lceil\alpha k\rceil+1} > k\right) \leq -\lambda.$$

Since the last inequality holds for all $\lambda < |\log(\rho_{\max}+\varsigma)|$ and every $\varsigma > 0$, we have

$$\inf_{\alpha>0}\lim_{k\to+\infty}\frac{1}{k}\log\mathbb{P}\left(T_{\lceil\alpha k\rceil+1} > k\right) \leq \inf_{\varsigma>0}\inf_{\lambda<|\log(\rho_{\max}+\varsigma)|} -\lambda = -\left|\log\rho_{\max}\right|.$$

This completes the proof of the upper bound (14) and the proof of Theorem 4.

IV. Examples

In this section, we give two instances for the Markov chain model described in Section II and we compute the rate I for each of the given examples. The first example is a gossip-type averaging protocol with Markov dependencies, similar to the protocol in [8] (except that protocol in [8] corresponds to directed graphs). One particular instance of this protocol is a random walk of a token along the edges of a given graph, according to a given transition probability matrix. In the second example, we consider a network with temporal correlations of the link failures, where we model the correlations by a Markov chain. We detail these two instances of the Markov chain model in the next two subsections.

A. Gossip with Markov dependencies

Let $G = (V, E)$ be a connected graph on N vertices. We assume that at each time $t \geq 0$ only one link of G can be active; if $e = \{u, v\} \in E$ is active at time t, then $W_t = I - \frac{1}{2}(e_u - e_v)(e_u - e_v)^\top$. The sequence of active one link graphs G_t, $t \geq 0$ is generated according to a Markov chain:

$$\mathbb{P}(G_0 = (V, e)) = v_e, \text{ for } e \in E$$
$$\mathbb{P}(G_{t+1} = (V, f) | G_t = (V, e)) = P_{ef}, \text{ for } e, f \in E,$$

where $v_e > 0, P_{ef} \geq 0$, $\sum_{f\in E} P_{ef} = 1$, for each $e \in E$, and $\sum_{e\in E} v_e = 1$. The set of states of the Markov chain is therefore

$$\mathcal{G}^{\text{Gossip}} = \{(V, e) : e \in E\}$$

and there are $M = |E|$ states. A disconnected collection on $\mathcal{G}^{\text{Gossip}}$ is of the form $\{(V, e) : e \in E \setminus F\}$, for some set of edges F that disconnects G. Thus, the set of all disconnected collections on $\mathcal{G}^{\text{Gossip}}$ is

$$\Pi(\mathcal{G}^{\text{Gossip}}) = \{\mathcal{H}_F : F \text{ disconnects } G\}.$$

where $\mathcal{H}_F =:= \{(V, e) : e \in E \setminus F\}$, for $F \subseteq E$. By Theorem 4, we get the formula for $\rho_{\max}$:

$$\rho_{\max} = \max_{F\subseteq E:\ F \text{ disconnects } G} \rho(P_{\mathcal{H}_F}).$$

Computing $\rho_{\max}$ for this model is difficult in general, as it involves computing the spectral radius for all submatrices $P_{\mathcal{H}_F}$ of the transition matrix P associated with disconnected collections $\mathcal{H}_F$. A simple approximation for $\rho_{\max}$ can be obtained using the row-sum based lower bound for the spectral radius. We explain this next. For any fixed disconnected collection $\mathcal{H}_F$, we denote by $\underline{c}(P_{\mathcal{H}_F})$ the minimal row sum of its associated submatrix $P_{\mathcal{H}_F}$: $\underline{c}(P_{\mathcal{H}_F}) = \min_{i=1,...,|\mathcal{H}_F|}\sum_{j=1}^{|\mathcal{H}_F|}[P_{\mathcal{H}_F}]_{ij}$ We then have, for any $\mathcal{H}_F$ [11]: $\underline{c}(P_{\mathcal{H}_F}) \leq \rho(P_{\mathcal{H}_F})$, implying

$$\max_{F\subseteq E:\ F \text{ disconnects } G} \underline{c}(P_{\mathcal{H}_F}) \leq \rho_{\max}. \quad (19)$$

In particular, for gossip on a tree, we get a very simple lower bound on $\rho_{\max}$ that involves no computations (it involves only $O(M^2)$ comparisons of certain entries of the matrix P.) When $G = (V, E)$ is a tree, removal of any edge $f \in E$ disconnects G. Also, for any $F' \subseteq F \subseteq E$, the matrix $P_{\mathcal{H}_F}$ is a submatrix of $P_{\mathcal{H}_{F'}}$, and so $\underline{c}(P_{\mathcal{H}_F}) \leq \underline{c}(P_{\mathcal{H}_{F'}})$, i.e., the minimal row sum can only grow as the edges are removed from F. This implies that we can decrease the space of search in (19) to the set of edges of G:

$$\max_{F\subseteq E:\ F \text{ disconnects } G} \underline{c}(P_{\mathcal{H}_F}) = \max_{f\in E}\underline{c}(P_{\mathcal{H}_f}) \leq \rho_{\max}. \quad (20)$$

Now, for any fixed $f \in E$, since P is stochastic, it holds that $\underline{c}(P_{\mathcal{H}_f}) = 1 - \max_{e\in E\setminus f} P_{ef}$; that is, to compute the minimal row sum of $P_{\mathcal{H}_f}$, we only have to find the maximal entry of the column P^f, with entry P_{ff} excluded. This finally implies:

$$\rho_{\max} \geq \max_{f\in E} 1 - \max_{e\in E\setminus f} P_{ef} = 1 - \min_{f\in E}\max_{e\in E\setminus f} P_{ef}. \quad (21)$$

We can see an interesting phenomenon in the lower bound on $\rho_{\max}$ in eq. (21): when $\max_{e\in E\setminus f} P_{ef}$ is high for every link e, that is, when the gossip token is more likely to jump to a different link $f \neq e$, rather than to stay on the same link e ($P_{ef} >> P_{ee}$, for some $f \neq e$), the bound in eq. (21) has a small value. Assuming that $\rho_{\max}$ follows the tendency of its lower bound, we obtain a high rate I for this case of "negative correlations". This is in accordance with the intuition: if every link has a low probability P_{ee} to be repeated (repeating a link is a wasteful transmission in gossip), the convergence of gossip is faster and thus the rate I is higher.

B. Link failures with temporal correlations

Let $G=(V,E)$ be a connected graph on N vertices. For each $e \in E$ and $t \geq 0$, let $Y_{e,t} \in \{0,1\}$ be a random variable that models the occurrence of the link e at time t: if $Y_{e,t}=1$ then e is online at time t, and e is offline otherwise. For each link e, we assume that the failures of e occur in time according to a Markov chain. Also, the failures of different links are independent. More precisely, we assume that $Y_{e,t}$ and $Y_{f,s}$ are independent for all $t,s \geq 0$ if $e \neq f$, and, for $e \in E$ and $t \geq 1$:

$$\mathbb{P}(Y_{e,t+1}=1|Y_{e,t}=1)=p_e,$$
$$\mathbb{P}(Y_{e,t+1}=0|Y_{e,t}=0)=q_e,$$

$\mathbb{P}(Y_{e,0}=1)=v_e$, for some $p_e, q_e, v_e \in (0,1)$. In other words, the joint state of all the links in the network evolves according to the $|E|$ independent Markov chains, where each Markov chain determines the state of one link. Given the network realization G_t, the averaging matrix W_t can be chosen, e.g., as the Metropolis or an equal weight matrix [1].

We compute the rate I for this model, following the reasoning in the proof of Theorem 4, and exploiting the decoupled single-link Markov chains. We first find the set of all network realizations at time t. Due to the independence in space of the link failures, and the fact that each link is on/off at time t with positive probability, the set of all network realizations at time t is the set of all subgraphs of G:

$$\mathcal{G}^{\text{Link fail.}}=\{(V,E'): E' \subseteq E\}.$$

Consider now a fixed disconnected collection $\mathcal{H}$ on $\mathcal{G}^{\text{Link fail.}}$ and let F be $\Gamma(\mathcal{H})=E \setminus F$; note that F disconnects G. Then $\mathcal{H}$ is necessarily a subset of the (bigger) collection $\mathcal{H}_F=\{(V,E'): E' \subseteq E \setminus F\}$ and thus $\mathbb{P}(G_t \in \mathcal{H}, 0 \leq t \leq k) \leq \mathbb{P}(G_t \in \mathcal{H}_F, 0 \leq t \leq k)$. The latter implies that, in order to find the most likely $\mathcal{H}$ that determines the rate I, we can search over the smaller set $\{\mathcal{H}_F : F \text{ disconnects } G\}$. Thus, we focus on the right hand side of the latter inequality:

$$\mathbb{P}(G_t \in \mathcal{H}_F,\ 0 \leq t \leq k)=\mathbb{P}(Y_{e,t}=0,\ \text{for } e \in F,\ 0 \leq t \leq k)$$
$$=\prod_{e \in F} \mathbb{P}(Y_{e,t}=0,\ 0 \leq t \leq k)=\prod_{e \in F}(1-v_e)q_e^k; \tag{22}$$

the second equality in (22) follows by the independence of failures of different links. The rate at which the probability in (22) decays is equal to $\sum_{e \in F}|\log q_e|$, and thus the rate I equals

$$I=I(\{q_e\})=\min_{F \subseteq E:\ F \text{ disconnects } G} \sum_{e \in F}|\log q_e|. \tag{23}$$

Optimization problem in (23) is the minimum cut problem [12], with the cost of edge $e \in E$ equal to $|\log q_e|$. (Recall that q_e is the probability that the link e stays offline, given that in the previous time it was also offline.) Problem (23) is a convex problem, and there are efficient numerical algorithms to solve it, e.g., [12].

To get some intuition on the effect of temporal correlations, we let $q=q_e=p_e$, for all $e \in E$, i.e., all the links have the same symmetric 2×2 transition matrix. Note that $q=1/2$ corresponds to the temporally uncorrelated link failures. When $q<1/2$, a link *is more likely to change its state* (on/off) with respect to its state in the previous time ("negative correlation") than to maintain it. From (23), the rate $I(q)>I(1/2)$ for $q<1/2$. We conclude that a "negative correlation" increases (improves) the rate. Likewise, a "positive correlation" ($q>1/2$) decreases (degrades) the rate.

V. Conclusion

We studied the products $W_k W_{k-1} \cdots W_0$ of temporally dependent random stochastic matrices. We modeled the temporal dependence through a Markov chain with a transition matrix P, whose set od states is the set of all possible graphs that support the matrices W_k. For this model, we calculated the large deviations rate I for convergence in probability of the product $W_k W_{k-1} \cdots W_0$. We found that the rate I is determined by the spectral radii of the submatrices of P associated with the sub-collections of graphs that are disconnected in union. We supported our analysis with two example models, namely token-based averaging protocol and temporally dependent link failures. The examples showed that the "negative temporal correlations" of the links binary states (being on or off) increase (improve) the rate I with respect to the temporally independent links.

References

[1] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Syst. Contr. Lett.*, vol. 53, no. 1, Sep. 2004.

[2] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, Jan. 2007.

[3] B. Touri and A. Nedić, "Product of random stochastic matrices," *Submitted to the Annals of Probability*, 2011, available at: http://arxiv.org/pdf/1110.1751v1.pdf.

[4] 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.

[5] 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," 2011, available at: http://arxiv.org/abs/1111.4555.

[6] 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.

[7] D. Bajović, J. Xavier, J. M. F. Moura, and B. Sinopoli, "Consensus and products of random stochastic matrices: Exact rate for convergence in probability," *submitted to IEEE Transactions on Signal Processing*, Feb. 2012, available at: http://arxiv.org/abs/1202.6389.

[8] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1157–1170, 2009.

[9] F. d. Hollander, *Large deviations*. Fields Institute Monographs, American Mathematical Society, 2000.

[10] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, no. 98, pp. 298–305, 1973.

[11] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambrige, United Kingdom: Cambrige Univesity Press, 1990.

[12] R. D. Carr, G. Konjevod, G. Little, V. Natarajan, and O. Parekh, "Compacting cuts: a new linear formulation for minimum cut," *ACM Transactions on Algorithms*, vol. 5, no. 3, July 2009, DOI: 10.1145/1541885.1541888.