A Distributed Dynamical Scheme for Fastest Mixing Markov Chains

Michael M. Zavlanos, Daniel E. Koditschek and George J. Pappas

Abstract— This paper introduces the problem of determining through distributed consensus the fastest mixing Markov chain with a desired sparsity pattern. In contrast to the centralized optimization-based problem formulation, we develop a novel distributed relaxation by constructing a dynamical system over the cross product of an appropriately patterned set of stochastic matrices. In particular, we define a probability distribution over the set of such patterned stochastic matrices and associate an agent with a random matrix drawn from this distribution. Under the assumption that the network of agents is connected, we employ consensus to achieve agreement of all agents regardless of their initial states. For sufficiently many agents, the law of large numbers implies that the asymptotic consensus limit converges to the mean stochastic matrix, which for the distribution under consideration, corresponds to the chain with the fastest mixing rate, relative to a standard bound on the exact rate. Our approach relies on results that express general element-wise nonnegative stochastic matrices as convex combinations of 0-1 stochastic matrices. Its performance, as a function of the weights in these convex combinations and the number of agents, is illustrated in computer simulations. Because of its differential and distributed nature, this approach can handle large problems and seems likely to be well suited for applications in distributed control and robotics.

I. INTRODUCTION

Markov chains are stochastic processes describing discrete time trajectories of distributions over discrete state spaces whose iterates are prescribed probabilistically according to the value of their immediate predecessors. Under certain conditions, these processes converge to an equilibrium distribution, the so called stationary distribution. The rate at which a Markov chain converges to its stationary distribution is called the mixing rate of the chain and is determined by the second largest eigenvalue of its transition matrix. In this paper, we investigate the problem of determining the fastest mixing Markov chain over the set of appropriately patterned stochastic matrices whose convergence yields a representation of the desired chain.

Markov chains arise frequently in the areas of statistics, physics, biology, computer science [1]–[7] and distributed robotics [8]–[13] and, in particular, in the context of Markov chain Monte Carlo (McMC) simulation [14, 15]. McMC allows simulation of stochastic processes with high-dimensional state spaces and known probability distributions by simulating instead Markov chains that have the distribution of interest as their stationary distribution [16]. Such chains can be obtained by the Metropolis-Hastings algorithm [2, 17]. Since sampling takes place while the chain converges to stationarity, rapid mixing rates are necessary for valid inference results [16]. Therefore, determining or bounding the second largest eigenvalue of Markov chains is vital.

This paper is strongly influenced by a centralized formulation of a closely related problem, namely, the assignment of transition probabilities to the edges of a graph on which resulting random walks are required to converge as quickly as possible [18]. The authors of [18] contrasted prior effort to determine analytical bounds for the second largest eigenvalue of a Markov chain [19]–[24] with their new observation that the fastest mixing Markov chain on a given graph could be computed exactly by means of a polynomial time optimization algorithm. The proposed approach was restricted to chains with symmetric transition patterns and moderate size (at most 100 states). For larger problems subgradient methods were proposed. This convex optimization formulation and duality theory allowed derivation of improved bounds on the second largest eigenvalue of a Markov chain.

Motivated by the scalability and robustness properties of distributed control as well as the many applications of Markov chains in distributed robotics, ranging from motion planning in probabilistic environments to probabilistic target tracking, we consider a novel distributed relaxation to the problem of determining the fastest mixing Markov chain with a desired sparsity pattern. In particular, we associate every chain with a stochastic transition matrix, define a probability distribution over the set of these matrices, and define a network of agents sampling matrices from this distribution. Under the assumption that this network is connected, we construct differential flows on the set of stochastic matrices that achieve consensus of all agents on their initial samples. For sufficiently many agents, the law of large numbers guarantees that the asymptotic consensus limit converges to the mean stochastic matrix, which for the distribution under consideration, corresponds to the chain with the fastest mixing rate. The proposed distribution as well as the desired sparsity patterns, rely on results that express general element-wise nonnegative stochastic matrices as convex combinations of 0-1 stochastic matrices. The efficiency of our approach is illustrated in representative computer simulations.

The rest of this paper is organized as follows. In Section II we define the problem of determining the fastest mixing Markov chain with a desired sparsity pattern. In Section III we define the consensus algorithm on the set of stochastic matrices and discuss its convergence properties. Performance of the algorithm is discussed in Section IV, while in Section V, we illustrate our approach in nontrivial simulations.

II. Preliminaries & Problem Definition

Let $\{X_t\}_{t=0}^{\infty}$ denote a sequence of random variables, where every variable X_t takes values in a finite set $\mathcal{X} = \{1, \ldots, n\}$. We call the stochastic process $\{X_t\}_{t=0}^{\infty}$ a Markov chain with state space $\mathcal{X}$ if it satisfies the Markov property:

$$\mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) =$$
$$= \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}),$$

for all $t \geq 1$ and all $x_t, \ldots, x_0 \in \mathcal{X}$ [25, 26]. We assume time-homogenous Markov chains, for which the transition probabilities $s_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$ from state i to state j are independent of t. Then, the matrix $S = (s_{ij}) \in \mathbb{R}^{n \times n}$ is called the transition matrix of the Markov chain and is a non-negative and *stochastic* matrix, i.e., its row sums are all equal to one. Denote, further, by $\pi_t = (\pi_t^{[i]}) \in \mathbb{R}_+^n$ the distribution of X_t so that $\pi_t^{[i]} = \mathbb{P}(X_t = i)$.[1] Then, we have the following update rule

$$\pi_{t+1} = S^T \pi_t, \tag{1}$$

which implies that $\pi_t = (S^T)^t \pi_0$, where π_0 denotes the initial distribution of the chain. If the Markov chain is *irreducible* and *ergodic*, then there exists a distribution $\pi^\star = (\pi_i^\star)$ over $\mathcal{X}$ such that

$$\lim_{t \to \infty} \pi_t = \pi^\star.$$

The distribution $\pi^\star$ is called the *stationary distribution* of the chain and is the unique vector that satisfies $\pi^\star = S^T \pi^\star$ and $\mathbf{1}^T \pi^\star = 1$, where $\mathbf{1}$ is a column vector of all entries equal to one. A chain is irreducible if every state in the chain can be reached from every other state, i.e., if for every pair of states $i, j \in \mathcal{X}$, there exists a $t \in \mathbb{N}$ such that $\mathbb{P}(X_t = j \mid X_0 = i) > 0$. On the other hand, necessary and sufficient conditions for ergodicity are that the chain is

(a) *persistent*, i.e., $f_{ii} \triangleq \sum_{t=1}^{\infty} f_{ii}^t = 1$ for all $i \in \mathcal{X}$, where f_{ij}^t is the probability that starting from i, the first entry to j occurs at the t-th step,

(b) *non-null*, i.e., the mean recurrence time $\mu_i \triangleq \sum_{t=1}^{\infty} t f_{ii}^t$ for every state $i \in \mathcal{X}$ is finite,

(c) *aperiodic*, i.e., $\gcd\{t \mid \mathbb{P}(X_t = i \mid X_0 = i) > 0\} = 1$ for all $i \in \mathcal{X}$, where gcd indicates the greatest common divisor.

Hence, a time-homogeneous Markov chain converges to the unique stationary distribution $\pi^\star$, independent of π_0, if it is irreducible and ergodic. In this case, $\pi_i^\star = \frac{1}{\mu_i}$ [26]. If the chain is also *symmetric*, then $\pi_i^\star = \frac{1}{n}$ [18].

The time required for a Markov chain to converge to its stationary distribution is called the *mixing time* of the chain and is determined by the second largest eigenvalue modulus of the transition probability matrix of the chain. Let $0 \leq |\lambda_1| \leq \cdots \leq |\lambda_{n-1}| \leq |\lambda_n| = 1$ be the ordered modula of the eigenvalues of S, which by the Perron-Frobenious theorem can not be greater that one [27]. Define, further, the sparsity pattern $P = (p_{ij}) \in \{0, 1\}^{n \times n}$ of the chain with

$p_{ij} = 1$ if $s_{ij} > 0$,[2] and let $\mathcal{S}_n^P$ denote the set of all $n \times n$ transition matrices S that respect the pattern P. Then, the problem addressed in this paper can be stated as follows.

Problem 1: Given a sparsity pattern P, design a distributed dynamical estimation scheme whose execution, from a set of arbitrarily initialized local estimates $\{S_i\}_{i=1}^m \in \mathcal{S}_n^P$ of the underlying Markov chain, collectively converges to a common, pattern-preserving transition matrix, $S \in \mathcal{S}_n^P$, with the smallest mixing rate $|\lambda_{n-1}|$.

Our approach to Problem 1 relies on defining a consensus flow on the transition matrices $\{S_i(t)\}_{i=1}^m$ that lies in the set $\mathcal{S}_n^P$ for all time t. The following result characterizes the set $\mathcal{S}_n^P$ in terms of the 0-1 stochastic matrices $\Sigma = (\sigma_{ij}) \in \{0, 1\}^{n \times n}$, and is similar in nature to a well known result that expresses the doubly stochastic matrices as convex combinations of permutation matrices [27].

Lemma 2.1: The set $\mathcal{S}_n^P$ of all $n \times n$ stochastic matrices is a convex polyhedron whose vertices are the 0-1 stochastic matrices $\Sigma \in \{0, 1\}^{n \times n}$.

Proof: Let $S \in \mathcal{S}_n^P$. We use induction on the number of positive entries in S. If S has exactly n positive entries, then S is a 0-1 stochastic matrix, so the result holds. If S is not a 0-1 stochastic matrix, let $\theta = \min_{i,j}\{s_{ij}\} < 1$ and define the 0-1 stochastic matrix $\Sigma \in \mathcal{S}_n^P$, such that if $\sigma_{ij} = 1$, then $s_{ij} > 0$. Clearly, such a matrix Σ exists for any $S \in \mathcal{S}_n^P$. Then, $\tilde{S} = \frac{1}{1-\theta}(S - \theta \Sigma)$ is also a stochastic matrix in $\mathcal{S}_n^P$ and has fewer positive entries than S. Hence, by the induction hypothesis, $\tilde{S}$ is a convex combination of 0-1 stochastic matrices, and therefore, so is $S = (1-\theta)\tilde{S} + \theta \Sigma$. This shows that $\mathcal{S}_n^P$ is the convex hull of 0-1 stochastic matrices, which are the vertices of $\mathcal{S}_n^P$, since no 0-1 stochastic matrix can be written as a convex combination of other 0-1 stochastic matrices. ∎

III. Consensus on the Set of Stochastic Matrices

A. Sampling Stochastic Matrices

Before addressing Problem 1, we characterize the transition matrix $S \in \mathcal{S}_n^P$ that corresponds to the Markov chain with the fastest mixing rate.

Lemma 3.1: Let $\psi(S) \triangleq \|S - \frac{1}{n}\mathbf{1}\mathbf{1}^T\|_F^2$. Then, $|\lambda_{n-1}(S)| \leq \psi(S)$ for all $S \in \mathcal{S}_n^P$.

Proof: Similarly to [18], note that $|\lambda_{n-1}(S)|$ corresponds to the spectral radius of S restricted to the subspace $\mathbf{1}^\perp$, i.e., $|\lambda_{n-1}(S)| = \rho(B^T S B)$, where $B = [b_1 \ldots b_n] \in \mathbb{R}^{n \times n}$ denotes an orthogonal projection on $\mathbf{1}^\perp$, such that $b_i^T b_j = 0$ and $b_i^T \mathbf{1} = 0$ for all $i, j = 1, \ldots, n$. Moreover, $\rho(B^T S B) \leq \|B^T S B\|_2$, and letting $B = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, we get

$$|\lambda_{n-1}(S)| \leq \|(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^T S (I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\|_2$$
$$= \|S - \frac{1}{n}\mathbf{1}\mathbf{1}^T\|_2 \leq \|S - \frac{1}{n}\mathbf{1}\mathbf{1}^T\|_F,$$

where $\|X\|_F^2 = \mathbf{tr}(XX^T)$ for any $X \in \mathbb{R}^{n \times n}$ denotes the Frobenious norm and the last inequality results from the

---

[1] We denote by $\mathbb{R}_+$ the set $[0, \infty)$.

[2] Note that $s_{ij} = 0$ does not imply that $p_{ij} = 0$.

equivalence of the norms $\|\cdot\|_2$ and $\|\cdot\|_F$, i.e., $\|X\|_2 \leq \|X\|_F \leq \sqrt{n}\|X\|_2$. ∎

Minimizing $|\lambda_{n-1}(S)|$ directly is, in general, hard. Instead, using Lemma 3.1 we can minimize the relaxation $\psi(S)$, which is a convex function of the stochastic matrix S. Let

$$S^\star \triangleq \operatorname{argmin}_{S \in \mathcal{S}_n^P} \psi(S). \tag{2}$$

The following result makes use of Lemma 2.1 and characterizes a class of distributions of stochastic matrices $S \in \mathcal{S}_n^P$ with mean $S^\star$. Essentially, it provides a way of sampling from the set $\mathcal{S}_n^P$ and motivates the design of the distributed algorithm for Problem 1.

Lemma 3.2: For any collection of k uniformly distributed 0-1 stochastic matrices $\{\Sigma_j\}_{j=1}^k \in \mathcal{S}_n^P$ and any random vector of nonnegative weights $\alpha = [\alpha_1 \dots \alpha_k]^T \in \mathbb{R}_+^k$, let $S = \frac{1}{\alpha^T \mathbf{1}} \sum_{j=1}^k \alpha_j \Sigma_j \in \mathcal{S}_n^P$. Then, assuming the usual independence, we have $\mathbb{E}(S) = S^\star$.

Proof: Since $\mathbb{E}(S)$ is a stochastic matrix, we have that $\mathbb{E}(S)\mathbf{1}\mathbf{1}^T = \mathbf{1}\mathbf{1}^T$. Hence, using the property that the trace of a matrix is equal to the trace of its transpose, we get

$$\begin{aligned}
\psi(\mathbb{E}(S)) &= \mathbf{tr}\big(\mathbb{E}(S) - \frac{1}{n}\mathbf{1}\mathbf{1}^T)(\mathbb{E}(S)^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T\big) \\
&= \mathbf{tr}\big(\mathbb{E}(S)\mathbb{E}(S)^T\big) - 1,
\end{aligned}$$

where

$$\mathbb{E}(S) = \sum_{j=1}^k \mathbb{E}\Big(\frac{\alpha_j}{\alpha^T\mathbf{1}}\Big)\mathbb{E}(\Sigma_j) = \mathbb{E}(\Sigma)\mathbb{E}\Big(\sum_{j=1}^k \frac{\alpha_j}{\alpha^T\mathbf{1}}\Big) = \mathbb{E}(\Sigma),$$

for any uniformly distributed 0-1 stochastic matrix $\Sigma \in \mathcal{S}_n^P$. Note, further, that $\mathbb{E}(\Sigma) = \sum_{j=1}^{N_p} \frac{1}{N_p}\Sigma_j$, where N_p is the total number of 0-1 stochastic matrices that satisfy the pattern $\mathcal{S}_n^P$, and for any $0 < \epsilon < \frac{1}{N_p}$, define the ϵ-perturbed mean of the 0-1 stochastic matrix Σ by

$$\begin{aligned}
\mathbb{E}_\epsilon(\Sigma) &= \Big(\frac{1}{N_p} + \epsilon\Big)\Sigma_1 + \Big(\frac{1}{N_p} - \epsilon\Big)\Sigma_2 + \frac{1}{N_p}\sum_{j=3}^{N_p}\Sigma_j \\
&= \mathbb{E}(\Sigma) + \epsilon(\Sigma_1 - \Sigma_2).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\psi(\mathbb{E}(\Sigma)) &- \psi(\mathbb{E}_\epsilon(\Sigma)) = \\
&= \mathbf{tr}\big(\mathbb{E}(\Sigma)\mathbb{E}(\Sigma)^T - \mathbb{E}_\epsilon(\Sigma)\mathbb{E}_\epsilon(\Sigma)^T\big) \\
&= -2\epsilon\mathbf{tr}\big(\mathbb{E}(\Sigma)(\Sigma_1 - \Sigma_2)^T\big) - \epsilon^2\|\Sigma_1 - \Sigma_2\|_F^2 \\
&= -\frac{2\epsilon}{N_p}\mathbf{tr}\Big(\sum_{j=1}^{N_p}(\Sigma_1 - \Sigma_2)\Sigma_j^T\Big) - \epsilon^2\|\Sigma_1 - \Sigma_2\|_F^2 \\
&= -\epsilon^2\|\Sigma_1 - \Sigma_2\|_F^2 < 0,
\end{aligned}$$

since $\mathbf{tr}\big(\sum_{j=1}^{N_p}(\Sigma_1 - \Sigma_2)\Sigma_j^T\big) = 0$. To see this, suppose that Σ_1 and Σ_2 differ in the i-th row. Then, there exist indices $s \neq t$ such that $[\Sigma_1 - \Sigma_2]_{is} = 1$ and $[\Sigma_1 - \Sigma_2]_{it} = -1$, where $[\cdot]_{ij}$ denotes the ij-th entry of a matrix. Observe, further, that the number of 0-1 stochastic matrices Σ_j with $[\Sigma_j]_{is} = 1$ is equal to the number of 0-1 stochastic matrices Σ_j with $[\Sigma_j]_{it} = 1$, and corresponds to the number of all 0-1

stochastic matrices with all remaining rows $k \neq i$ satisfying the sparsity pattern $\mathcal{S}_n^P$. Hence, $[\sum_{j=1}^{N_p}(\Sigma_1 - \Sigma_2)\Sigma_j^T]_{ii} = 0$ for all $i = 1, \dots, n$, which implies that $\mathbf{tr}\big(\sum_{j=1}^{N_p}(\Sigma_1 - \Sigma_2)\Sigma_j^T\big) = 0$, as desired.[3]

We conclude that $\psi(\mathbb{E}(\Sigma)) < \psi(\mathbb{E}_\epsilon(\Sigma))$ for any perturbation $0 < \epsilon < \frac{1}{N_p}$, which implies that $\psi(\mathbb{E}(\Sigma)) = \min_{S \in \mathcal{S}_n^P} \psi(S)$ and, hence, $\mathbb{E}(S) = S^\star$, by convexity of the function $\psi(S)$. ∎

In other words, if we sample stochastic matrices $S \in \mathcal{S}_n^P$ according to the assumptions of Lemma 3.2, then their mean value is equal to $S^\star$. Combined with Lemma 3.1 and the law of large numbers [25, 26]

$$\lim_{m \to \infty} \frac{1}{m}\sum_{i=1}^m S_i = \mathbb{E}(S)$$

we conclude that we can obtain a stochastic matrix $S \in \mathcal{S}_n^P$ with the lowest second largest eigenvalue $|\lambda_{n-1}(S)|$ by sampling the set $\mathcal{S}_n^P$ of stochastic matrices and averaging these samples. This observation leads to a distributed control scheme for Problem 1 using distributed averaging or *consensus*.

B. The Distributed Consensus Algorithm

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote a network of m agents, where $\mathcal{V} = \{1, \dots, m\}$ is the set of vertices indexed by the set of agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of communication links. We assume bidirectional communication links and so $(i, j) \in \mathcal{E}$ if and only if $(j, i) \in \mathcal{E}$. Such graphs are called *undirected* and are the main focus of this paper. Furthermore, assume $\mathcal{G}$ is such that there exists a path (i.e., a sequence of distinct vertices such that consecutive vertices are adjacent) between any two of its vertices. Then, we say that $\mathcal{G}$ is *connected*. Any vertices i and j of an undirected graph $\mathcal{G}$ that are joined by a link $(i, j) \in \mathcal{E}$, are called *adjacent* or *neighbors*. Hence, we can define the set of neighbors of agent i by $\mathcal{N}_i \triangleq \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$.

In what follows, we make use of an equivalent algebraic representation of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ using a *laplacian matrix*

$$L = \Delta - A, \tag{3}$$

where $A = (a_{ij}) \in \mathbb{R}^{m \times m}$ corresponds to the *adjacency matrix* of the graph $\mathcal{G}$, which is such that $a_{ij} = \frac{1}{|\mathcal{N}_i|}$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise and $\Delta = \operatorname{diag}\big(\sum_{j=1}^n a_{ij}\big)$ denotes the *valency matrix*.[4] The spectral properties of the laplacian matrix are closely related to graph connectivity. In particular, if $\nu_1 \leq \nu_2 \leq \dots \leq \nu_m$ are the ordered eigenvalues of the laplacian matrix L, then $\nu_1 = 0$, with corresponding eigenvector $\mathbf{1}$, and $\nu_2 > 0$ if and only if $\mathcal{G}$ is connected [28]. Hence, we have the following result.

---

[3]Note that if Σ_1 and Σ_2 do not differ in the i-th row, then the i-th row of $\Sigma_1 - \Sigma_2$ is zero, which again results in $[\sum_{j=1}^{N_p}(\Sigma_1 - \Sigma_2)\Sigma_j^T]_{ii} = 0$.

[4]Since we do not allow self-loops, we define $a_{ii} = 0$ for all i. Also, $|\mathcal{N}_i|$ denotes the cardinality of the set $\mathcal{N}_i$.

Lemma 3.3 (Consensus on $\mathcal{S}_n^P$): Let $\mathcal{G}$ denote a network of m agents and assume that every agent i is associated with a stochastic matrix $S_i \in \mathcal{S}_n^P$. Then, the closed loop system

$$\dot{S}_i = -\frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} (S_i - S_j), \;\; \forall \, i = 1, \ldots, m, \quad (4)$$

defines a consensus algorithm on the set of stochastic matrices $\mathcal{S}_n^P$, and if $\mathcal{G}$ is connected, it guarantees that $S_i - S_j \to \mathbf{0}$ for all i, j as $t \to \infty$.

Proof: For all agents i, observe that $\sum_{j \in \mathcal{N}_i}(S_i - S_j)\mathbf{1} = \mathbf{0}$, which implies that $\dot{S}_i\mathbf{1} = \mathbf{0}$. Hence, $S_i(t)\mathbf{1} = \mathbf{c}_i$, for any constant vector $\mathbf{c}_i \in \mathbb{R}^n$ and all time $t \geq 0$. Since, $S_i(0)\mathbf{1} = \mathbf{1}$ for all agents i, we have that $\mathbf{c}_i = \mathbf{1}$ for all i, and so $S_i(t)\mathbf{1} = \mathbf{1}$ for all time $t \geq 0$ and all agents i. The fact that $S_i(t) \geq 0$ for all i and all time $t \geq 0$ follows from the fact that $S_i(0) \in \mathcal{S}_n^P$ for all agents i and the distributed averaging law (4), which ensures that $S_i(t) \in \text{conv}\{S_j(0) \mid j = 1, \ldots, m\}$ for all time $t \geq 0$. Hence, $S_i(t) \in \mathcal{S}_n^P$ for all time $t \geq 0$ and all agents i.

Consider now the Lyapunov function candidate

$$V(S) \triangleq \frac{1}{2} \sum_{i=1}^{m} \sum_{j \in \mathcal{N}_i} \|S_i - S_j\|_F^2 = \frac{1}{2}\mathbf{tr} S^T (L \otimes I_n) S,$$

where $\otimes$ denotes the Kronecker product of matrices and $S = [S_1^T \; \ldots \; S_m^T]^T \in \mathbb{R}^{mn \times m}$. Taking the time derivative of $V(S)$ we get

$$
\begin{aligned}
\dot{V}(S) &= \frac{1}{2}\mathbf{tr}\dot{S}^T(L \otimes I_n)S + \frac{1}{2}\mathbf{tr} S^T (L \otimes I_n)\dot{S} \\
&= -\mathbf{tr} S^T (L \otimes I_n)^T (L \otimes I_n) S \\
&= -\|(L \otimes I_n)S\|_F^2 \leq 0.
\end{aligned}
$$

The set of critical points satisfies $(L \otimes I_n)S = \mathbf{0}$ and if the network is connected, then $S_i - S_j \to \mathbf{0}$ as $t \to \infty$. ∎

If we initialize consensus (4) according to the assumptions of Lemma 3.2, then the law of large numbers implies that for sufficiently large number m of agents, the asymptotic limit of the consensus approximates the optimal solution $S^\star$. In other words, Lemmas 3.2 and 3.3 provide a distributed algorithm for Problem 1, as desired.

IV. CONSENSUS PERFORMANCE

In this section we characterize the performance of the consensus update (4) for a given number of agents m. In particular, given a set of 0-1 stochastic matrices $\{\Sigma_{ij}\}_{j=1,\ldots,k_i}^{i=1,\ldots,m}$, with $k_i > 0$ the number of such matrices associated with agent i, and recalling that (Lemma 2.1)

$$S_i = \frac{1}{\alpha_i^T \mathbf{1}} \sum_{j=1}^{k_i} \alpha_{ij} \Sigma_{ij},$$

where $\alpha_i = [\alpha_{i1} \ldots \alpha_{ik_i}]^T$ is a vector of positive weights, we are interested in the solution of the following optimization problem

$$
\begin{aligned}
\min_{\alpha_{ij} \geq 0} \quad & \mathbb{P}\left(\|\text{vec}(S) - \text{vec}(S^\star)\|_2 \geq \epsilon\right) \\
\text{s.t.} \quad & S \triangleq \frac{1}{m}\sum_{i=1}^{m} S_i, \quad S_i = \frac{1}{\alpha_i^T \mathbf{1}} \sum_{j=1}^{k} \alpha_{ij}\Sigma_{ij}
\end{aligned} \quad (5)
$$

for any $\epsilon > 0$, where S and $S^\star$ are the limit of the consensus update (4) and the sought expectation, respectively, and $\text{vec} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n^2}$ denotes the *vectorization* of an $n \times n$ matrix. Optimization problem (5) is, in general, hard to solve. Instead, we solve the simpler relaxation

$$
\begin{aligned}
\min_{\alpha_{ij} \geq 0} \quad & \text{Var}(\text{vec}(S)) \\
\text{s.t.} \quad & S \triangleq \frac{1}{m}\sum_{i=1}^{m} S_i, \quad S_i = \frac{1}{\alpha_i^T\mathbf{1}} \sum_{j=1}^{k} \alpha_{ij}\Sigma_{ij},
\end{aligned} \quad (6)
$$

which results from an application of Chebyshev's inequality

$$\mathbb{P}\left(\|\text{vec}(S) - \text{vec}(S^\star)\|_2 \geq \epsilon\right) \leq \frac{\text{Var}(\text{vec}(S))}{\epsilon^2},$$

where the variance $\text{Var}(\text{vec}(S))$ is defined as [29, pp. 446–451]

$$\text{Var}(\text{vec}(S)) \triangleq \mathbb{E}\left(\|\text{vec}(S) - \mathbb{E}(\text{vec}(S))\|_2^2\right).$$

In particular, we have the following result.

Lemma 4.1: Problem (6) has a unique minimum obtained when $\alpha_i = c_i\mathbf{1}$, for any set of scalars $c_i > 0$ and all agents $i = 1, \ldots, m$.

Proof: Let $\beta_{ij} \triangleq \alpha_{ij}/\alpha_i^T\mathbf{1}$. Then, $\sum_{j=1}^{k_i}\beta_{ij} = 1$ for all $i = 1, \ldots, m$ and using Lemma 1.1 in the Appendix, the the objective function becomes

$$
\begin{aligned}
\text{Var}(\text{vec}(S)) &= \frac{1}{m^2}\sum_{i=1}^{m}\text{Var}(\text{vec}(S_i)) \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{k_i}\beta_{ij}^2\text{Var}(\text{vec}(\Sigma_{ij})),
\end{aligned}
$$

by independence of the 0-1 stochastic matrices Σ_{ij}. Hence, the optimization problem (6) is equivalent to

$$
\begin{aligned}
\min_{\beta_{ij} \geq 0} \quad & \sum_{i=1}^{m}\sum_{j=1}^{k_i}\beta_{ij}^2 \\
\text{s.t.} \quad & \sum_{j=1}^{k_i}\beta_{ij} = 1 \;\; \forall \, i = 1, \ldots, m,
\end{aligned} \quad (7)
$$

which is separable, and so corresponds to the solution of m copies of the problem

$$
\begin{aligned}
\min_{\beta_{ij} \geq 0} \quad & \sum_{j=1}^{k_i}\beta_{ij}^2 \\
\text{s.t.} \quad & \sum_{j=1}^{k_i}\beta_{ij} = 1,
\end{aligned} \quad (8)
$$

It can be shown that the solution to problem (8) is $\beta_{ij} = \frac{1}{k_i}$ for all $j = 1, \ldots, k_i$, which along with its convex nature, completes the proof. ∎

In other words, Lemma 4.1 implies that in order to increase the performance of the consensus update (4), every agent i should initialize its stochastic matrix $S_i = \frac{1}{\alpha_i^T\mathbf{1}}\sum_{j=1}^{k_i}\alpha_{ij}\Sigma_{ij}$ with equal weights α_{ij}.

V. SIMULATIONS

In this section we simulate the distributed consensus algorithm discussed in Sections III and IV. In particular, for the pattern illustrated in Fig. 1 and $k_i = 5$ for all agents i, we construct the initial samples $S_i = \frac{1}{\alpha_i^T\mathbf{1}}\sum_{j=1}^{k_i}\alpha_{ij}\Sigma_{ij}$ choosing the 0-1 stochastic matrices Σ_{ij} uniformly in the set $\mathcal{S}_n^P$ and the positive weights α_{ij} either randomly or
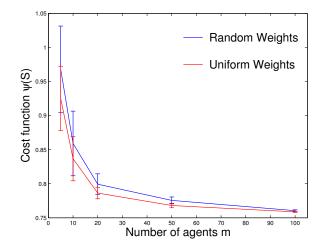
Fig. 1. Markov chain of size $n = 10$.

Fig. 2. Plot of the cost function $\psi(S)$ obtained by applying consensus (4) for the pattern illustrated in Fig. 1, corresponding to the data presented in Table I. Note that the larger the number of agents, the better the approximation of the sought expectation $S^\star$. Moreover, uniform weights $\alpha_{ij} \geq 0$ result in better performance, as discussed in Section IV. The communication networks $\mathcal{G}$ underlying the consensus algorithm are taken to be random and connected.

uniformly, as discussed in Section IV. Following is the initial stochastic matrix S_i of a sample agent i:

$$
S_i = \begin{bmatrix}
.2 & 0 & .4 & 0 & 0 & 0 & 0 & 0 & 0 & .4 \\
0 & .4 & .6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & .6 & 0 & .2 & .2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .6 & 0 & .2 & .2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & .2 & .2 & 0 & .6 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & .4 & .4 & .2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & .4 & 0 & .4 & .2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .2 & .2 & .6 & 0 \\
.2 & 0 & 0 & 0 & 0 & 0 & 0 & .4 & .2 & .2 \\
.6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .4
\end{bmatrix}
$$

We run the consensus update (4) developed Section III for different numbers of agents m and the results are shown in Table I and the associated Fig. 2. Note that the larger the number of agents, the better the approximation of the sought expectation $S^\star$, as predicted by the law of large numbers. Furthermore, choosing the weights α_{ij} uniformly results in a better performance of the algorithm, as discussed in Section IV. Following is the final stochastic S for the case of $m = 100$ agents and uniform weights:

$$
\begin{bmatrix}
.2501 & .2813 & .2502 & 0 & 0 & 0 & 0 & 0 & 0 & .2183 \\
.2432 & .2674 & .2638 & .2257 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & .2079 & .2484 & .2916 & .2521 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .2377 & .2514 & .2867 & .2242 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & .2313 & .2311 & .2797 & .2580 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & .2302 & .2473 & .2871 & .2354 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & .2487 & .2415 & .2654 & .2443 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .2134 & .2422 & .2746 & .2698 \\
.2619 & 0 & 0 & 0 & 0 & 0 & 0 & .2570 & .2313 & .2498 \\
.2448 & .2434 & 0 & 0 & 0 & 0 & 0 & 0 & .2498 & .2620
\end{bmatrix}
$$

Note that S approximates well the transition matrix $S^\star$ with the the fastest mixing rate, which for the particular example has all positive entries equal to 0.25.

VI. Conclusions

In this paper, we considered the problem of determining the fastest mixing Markov chain with a desired sparsity pattern, captured by a stochastic transition matrix. We developed

a novel distributed relaxation to the problem by constructing differential flows on the set of stochastic matrices. In particular, we defined a probability distribution over the set of stochastic matrices and associated an agent with any random matrix drawn from this distribution. Under the assumption that the network of agents is connected, we employed consensus to achieve agreement of all agents independent of their initial states. For sufficiently many agents, we showed that the asymptotic consensus limit converged to the mean stochastic matrix, which for the distribution under consideration, corresponded to the chain with the fastest mixing rate. The proposed distribution as well as the desired sparsity patterns, relied on results that express general stochastic matrices as convex combinations of 0-1 stochastic matrices. Due to its differential and distributed nature, our approach can handle large problems and seems likely to be well suited for applications in distributed control and robotics. Future work involves determining more quantitative bounds on the number of agents required to obtain almost optimal solutions, as well as applications in distributed robotics, in the context of probabilistic mapping of environments and target tracking.

References

[1] L. Billera, and P. Diaconis. *A Geometric Interpretation of the Metropolis-Hastings Algorithm*, Statistical Science, vol. 16, pp. 335–339, Nov. 2001.

[2] W. Hastings. *Monte Carlo Sampling Methods using Markov Chains and their Applications*, Biometrika, vol. 57, pp. 97–109, 1970.

[3] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*, Chapman and Hall, 1996.

[4] G. Casella, and E. I. George. *Explaining the Gibbs Sampler*, The American Statistician, vol. 46(3), pp. 167–174, 1992.

[5] S. Chib, and E. Greenberg. *Markov Chain Monte Carlo Simulation Methods in Econometrics*, Econometric Theory, vol. 12(3), pp. 409–431, 1996.

TABLE I

THE COST FUNCTION $\psi(S)$ AFTER APPLYING CONSENSUS (4) FOR THE MARKOV CHAIN SHOWN IN FIG. 1.

$k_i = 5, \forall i$	Number of agents m				
	5	10	20	50	100
Random α_i	.9678 ±.0636	.8591 ±.0474	.7993 ±.0152	.7755 ±.0051	.7605 ±.0012
Uniform α_i	.9252 ±.0472	.8367 ±.0324	.7863 ±.0085	.7680 ±.0033	.7587 ±.0010

[6] Z. Yang, and B. Rannala. *Bayesian Phylogenetic Inference using DNA Sequences: A Markov Chain Monte Carlo Method*, Molecular Biology and Evolution, vol. 14, pp. 717–724, 1997.

[7] C. J. Geyer, and E. A. Thompson. *Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference*, Journal of American Statistical Association, vol. 90(431), pp. 909–920, 1995.

[8] S. Thrun, D. Fox, W. Burgard, and F. Dellaert. *Robust Monte Carlo Localization for Mobile Robots*, Artificial Intelligence, vol. 128(1-2), pp. 99-141, May 2001.

[9] C. Hue, J.-P. Le Cadre, and P. Perez. *Sequential Monte Carlo Methods for Multiple Target Tracking and Data Fusion*, IEEE Transactions on Signal Processing, vol. 50(2), pp. 309–325, Feb. 2002.

[10] M. Kaess, and F. Dellaert. *A Markov Chain Monte Carlo Approach to Closing the Loop in SLAM*, Proc. of the IEEE Conference on Robotics and Automation, Barcelona, Spain, April 2005, pp. 643–648.

[11] D. Marinakis, G. Dudek, D. J. Fleet. *Learning Sensor Network Topology through Monte Carlo Expectation Maximization*, Proc. of the IEEE Conference on Robotics and Automation, Barcelona, Spain, April 2005, pp. 4581–4587.

[12] A. Ranganathan, and F. Dellaert. *Inference in the Space of Topological Maps: An MCMC-Based Approach*, Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, Sept. 2004, pp. 1518–1523.

[13] C.-X. Shi, B.-R. Hong, and Y.-Q. Wang. *Cooperative Exploration by Multi-Robots without Global Localization*, International Journal of Advanced Robotic Systems, vol. 5(2), pp. 129–138, 2008.

[14] J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[15] P. Bremaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues*. Texts in Applied Mathematics. Springer-Verlag, Berlin-Heidelberg, 1999.

[16] C. J. Geyer. *Practical Markov Chain Monte Carlo*. Statistical Science, vol. 7(4), pp. 473–483, Nov. 1992.

[17] S. Chib, and E. Greenberg. *Understanding the Metropolis-Hastings Algorithm*, The American Statistician, vol. 49(4), pp. 327–335, Nov. 1995.

[18] S. Boyd, P. Diaconis, and L. Xiao. *Fastest Mixing Markov Chain on a Graph*. SIAM Review, vol. 46(4), pp. 667–689, Dec. 2004.

[19] D. Aldous. *Random Walk on Finite Groups and Rapidly Mixing Markov Chains*. In Seminaire de Probabilities XVII, volume 986 of Lecture Notes in Mathematics, pp. 243–297. Springer, New York, 1983.

[20] P. Diaconis. *Group Representations in Probability and Statistics*. IMS, Hayward, CA, 1988.

[21] M. Jerrum, and A. Sinclair. *Approximating the Permanent*. SIAM Journal on Computing, vol. 18, pp. 1149–1178, 1989.

[22] P. Diaconis, and D. Stroock. *Geometric Bounds for Eigenvalues of Markov Chains*. The Annals of Applied Probability, vol. 1(1), pp. 36–61, 1991.

[23] A. Sinclair. *Improved Bounds for Mixing Rates of Markov Chains and Multicommodity Flow*. Combinatorics, Probability and Computing, vol. 1, pp. 351–370, 1992.

[24] N. Kahale. *A Semidefinite Bound for Mixing Rates of Markov Chains*. Proc. of the 5th Integer Programming and Combinatorial Optimization Conference, Springer Lecture Notes in Computer Science, vol. 1084, pp. 190–203, 1996.

[25] G. Grimmett, and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, 2003 (third edition).

[26] W. Feller. *An Introduction to Probability Theory and its Applications, Vol. 1*. John Wiley & Sons, 1968 (third edition).

[27] A. Berman, and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM Classics in Applied Mathematics, Philadelphia, 1994.

[28] C. Godsil, and G. Royle. *Algebraic Graph Theory*, Springer-Verlag, New York, 2001.

[29] R. G. Laha, and V. K. Rohatgi. *Probability Theory*, pp. 446-451, John Wiley and Sons, 1979.

APPENDIX

In this section we show a generalization of the well known identity $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$ [25] for $X, Y \in \mathbb{R}$ independent identically distributed random variables, to the case of random vectors. In particular, we have the following result.

Lemma 1.1: Let $X, Y \in \mathbb{R}^n$ be independent identically distributed random vectors with mean $\mathbb{E}(X) = \mathbb{E}(Y) = \mu \in \mathbb{R}^n$ and let $\mathrm{Var}(X) \triangleq \mathbb{E}\|X - \mathbb{E}(X)\|_2^2$ (similarly for Y) [29, pp. 446–451]. Then, $\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$, for any scalars $a, b \in \mathbb{R}$.

Proof: Observe first that

$$\mathrm{Var}(X) = \mathbb{E}(X - \mu)^T(X - \mu)$$
$$= \mathbb{E}(X^T X) - 2\mathbb{E}(X^T)\mu + \mu^T\mu = \mathbb{E}\|X\|_2^2 - \|\mu\|_2^2,$$

and similarly $\mathrm{Var}(Y) = \mathbb{E}\|Y\|_2^2 - \|\mu\|_2^2$. Then,

$$\mathrm{Var}(aX + bY) = \mathbb{E}\|aX + bY - \mathbb{E}(aX + bY)\|_2^2$$
$$= \mathbb{E}\|aX + bY - (a + b)\mu\|_2^2$$
$$= \mathbb{E}\|aX + bY\|_2^2 - (a + b)^2\|\mu\|_2^2$$
$$= a^2\mathbb{E}\|X\|_2^2 + 2ab\mathbb{E}(X^T)\mathbb{E}(Y) + b^2\mathbb{E}\|Y\|_2^2 - (a + b)^2\mu^2$$
$$= a^2\left(\mathbb{E}\|X\|_2^2 - \|\mu\|_2^2\right) + b^2\left(\mathbb{E}\|Y\|_2^2 - \|\mu\|_2^2\right)$$
$$= a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y),$$

which completes the proof. ∎