Distributed Parameter Estimation in Networks

Kamiar Rahnama Rad and Alireza Tahbaz-Salehi

***Abstract*— In this paper, we present a model of distributed parameter estimation in networks, where agents have access to partially informative measurements over time. Each agent faces a local identification problem, in the sense that it cannot consistently estimate the parameter *in isolation*. We prove that, despite local identification problems, if agents update their estimates recursively as a function of their neighbors' beliefs, they can consistently estimate the true parameter provided that the communication network is strongly connected; that is, there exists an information path between any two agents in the network. We also show that the estimates of all agents are asymptotically normally distributed. Finally, we compute the asymptotic variance of the agents' estimates in terms of their observation models and the network topology, and provide conditions under which the distributed estimators are as efficient as any centralized estimator.**

I. Introduction

One of the central problems in the study of multi-agent systems is the information aggregation problem. In many scenarios, information is spread throughout the network in such a way that no agent has access to enough data to learn a relevant parameter in isolation, and therefore, agents face the task of recovering the truth by engaging in communication with one another. Such problems are ubiquitous in social and economic networks, as well as networks engineered for specific applications. For example, Kotler [1] and Ioannides and Loury [2] document how people base their decisions on their neighbors' information when purchasing consumer products or adopting new technologies, respectively. Similarly, the main goal of distributed sensor and robotic networks is to aggregate relevant decentralized information, so that a prespecified task can be performed properly (see e.g., Jadbabaie, Lin, and Morse [3] and Bullo, Cortés, and Martínez [4]).

The goal of this paper is to develop a recursive model for aggregation of dispersed information over networks, where the measurements of each agent are only partially informative about the unknown parameter. In order to resolve the local identification problems they face,[1] agents in our model update their estimates as a function of their neighbors' beliefs. More specifically, we assume that at discrete time intervals, each agent sets its belief as the geometric mean of the likelihood of its observation and its neighbors' beliefs, and uses the mode of the updated belief function as the estimate for the unknown parameter.

We show that despite the absence of local identifiability across the network, agents' estimates are weakly consistent (i.e., converge to the truth in probability), provided that there exists a directed information path connecting any two agents in the network. In other words, we prove that as long as the underlying network is *strongly connected*, information is properly aggregated over the network and the local identification problems are resolved. We also show that as observations accumulate, the distribution of agents' estimates converge to a normal distribution. The consistency and asymptotic normality of agents' estimates hold regardless of the distribution of their measurements and the structure of the network (beyond of course, the strong connectivity requirement). Furthermore, we characterize the asymptotic covariance matrix of the distributed estimates in terms of agents' signal structures, as well as the network topology. Using this characterization, we show that in bidirectional networks, distributed estimators are as efficient as any centralized estimator with access to the collection of signals observed across the network. This efficiency is achieved even if the communication network is highly sparse.

Our work is related to the collection of works on learning in networks in economics, as well as distributed estimation and consensus algorithms in the control literature. The consensus literature (such as DeGroot [5], Jadbabaie, Lin, and Morse [3], and Golub and Jackson [6]) studies models in which a collection of agents asymptotically agree on the same value. Golub and Jackson provide conditions under which the asymptotic consensus value coincides with the true underlying parameter in *large* networks. In the same spirit is Xiao, Boyd, and Lall [7], which uses the consensus update to compute the maximum-likelihood estimate of the underlying parameter in a distributed fashion. These papers, however, do not address the problem of local identifiability, as they assume that all agents' observations are equally informative. This is the main point of departure of this paper from the above mentioned works, as we assume agents face local identification problems due to their different signal structures. Moreover, we show that as time progresses, not only the agents agree on their estimates, but also their consensus estimate converges to the true underlying parameter.

More relevant to our paper is Jadbabaie, Sandroni, and Tahbaz-Salehi [8], which studies distributed non-Bayesian learning in social networks. However, unlike [8], we study the problem of estimating a parameter in a continuum and in presence of continuous observations. Furthermore, we

Kamiar Rahnama Rad is with Department of Statistics, Columbia University, New York, NY 10027 (e-mail: kamiar@stat.columbia.edu).

Alireza Tahbaz-Salehi is with Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139. (e-mail: alirezat@mit.edu).

[1]Throughout the paper, by local (global) identifiability, we mean the possibility of consistently estimating the parameter through an agent's private data (the data observed by all agents). The terminology should not be mistaken by the concepts of local and global indistinguishably in a neighborhood of the true parameter in the parameter space.

characterize the rate of convergence and the efficiency of the estimates. Finally, our work is also relevant to Kar, Moura, and Ramanan [9], who focus on a non-stationary update with time-decaying weight sequences associated with consensus and innovation updates. In contrast to [9], in this paper, we address general non-linear observation models and present a stationary update for the beliefs.

The rest of the paper is organized as follows. In the next section, we describe the model and present the dynamics according to which agents update their estimates of the true parameter. In Section III, we prove that all agents' estimates are consistent. Asymptotic normality is proved in Section IV, where we also compute the asymptotic variance of agents' estimates. In Section V, we investigate the efficiency of the distributed estimators and compare our results with centralized maximum likelihood estimation. Section VI concludes.

II. The Model

A. Agents and Observations

Let $N = \{1, 2 \ldots, n\}$ denote a group of agents, located on a network, who are assigned the task of estimating an unknown parameter $\theta^* \in \Theta$, where $\Theta \subseteq \mathbb{R}^d$ is a convex parameter space. At discrete time steps $t \in \mathbb{N}$, each agent observes noisy and partially informative signals that can be used in estimating the parameter. More specifically, at any given time period t, agent i observes a random signal $s_t^i \in \mathbb{R}^p$, drawn from a distribution with conditional probability density $\ell_i(\cdot|\theta)$. We assume that agents' signals are i.i.d. over time and independent from the observations of all other agents.

The signals observed by a single agent, although potentially informative, do not reveal the parameter completely; i.e., each agent faces an identification problem. Two parameters are said to be observationally equivalent from the point of view of an agent if the conditional distributions of the signals coincide. We denote the set of parameters that are observationally equivalent to θ^* from the point of view of agent i by $\bar{\Theta}_i \triangleq \big\{\theta \in \Theta : \mathbb{P}[\ell_i(s^i|\theta) = \ell_i(s^i|\theta^*)] = 1\big\}$.[2]

Despite the local identification problems faced by the agents, we assume that the true parameter is identifiable if one has access to the signals observed by all agents.

Assumption (GI): The true parameter is *globally identifiable*; that is, $\bigcap_{i=1}^n \bar{\Theta}_i = \{\theta^*\}$.

The above assumption plays a key role in our main results. Clearly, in its absence, even an agent with access to all the data collected across the network over time would not be able to consistently estimate θ^*.

In addition to Assumption (GI), we impose the following regularity conditions on the observation models of the agents:

- (A1) $\ell_i(\cdot|\theta)$ is twice continuously differentiable in θ for all realizations of data.
- (A2) $\log \ell_i(\cdot|\theta)$ is concave in θ for all observations.
- (A3) $\ell_i(s^i|\theta)$ is a measurable function of s^i for all $\theta \in \Theta$.
- (A4) $\mathbb{E}[\log^2 \ell_i(s_1^i|\theta)] < \infty$ for all i.
- (A5) $\mathbb{E}\left[\sup_{\theta\in\mathcal{B}} \|\nabla_{\theta\theta} \log \ell_i(s_1^i|\theta)\|\right] < \infty$, for some neighborhood $\mathcal{B}$ of θ^*, where $\nabla_{\theta\theta}$ denotes the Hessian with respect to the parameter vector θ.

The above assumptions are quite mild and many of the usual distribution families, such as normals and exponentials, satisfy them. We have made these assumptions for simplicity, and our results hold under much weaker restrictions as well.

Finally, we define the *Fisher information matrix* corresponding to agent i's observation model as the covariance of its score function; that is,

$$\mathcal{I}_i(\theta) = \mathbb{E}\left[\nabla_\theta \psi_\theta^i(s_1^i) \nabla_\theta \psi_\theta^i(s_1^i)'\right] \tag{1}$$

where $\psi_\theta^i(s_t^i) \triangleq \log \ell_i(s_t^i|\theta)$ and ∇_θ denotes the gradient with respect to the parameter vector θ. As the definition suggests $\mathcal{I}_i$ is a $d \times d$ symmetric and positive semi-definite matrix.

B. Network Structure

In addition to signals $\{s_t^i\}_{t=1}^\infty$ observed privately over time, each agent can communicate with a subset of other agents known as its *neighbors*. We capture this neighborhood relation with a directed graph $G = (V, E)$, where each vertex in V corresponds to an agent $i \in N$, and there exists a directed edge $(j, i) \in E$ from vertex j to vertex i if agent i has access to the belief function of agent j. We denote the set of neighbors of agent i with N_i, and impose the following restriction on the network:

Assumption (C): The communication graph G is strongly connected; that is, there exists a directed path from any vertex to any other vertex in G.

Intuitively, Assumption (C) guarantees the possibility of information flow between any two agents (either directly or indirectly) in the network. The next sections will highlight the role played by this assumption in guaranteeing consistency and asymptotic normality of agents' estimates.

C. Belief Dynamics and Estimates

In order to aggregate the information provided to them over time – either through observations or communication with neighbors – agents hold and update beliefs over the parameter space Θ. More specifically, we denote the belief of agent i at time t with $\mu_{i,t} : \Theta \longrightarrow \mathbb{R}^+$, a probability measure over Θ. As for the dynamics, we assume that each agent updates its belief function as a geometric mean of its neighbors' beliefs and its own observation likelihood function; or equivalently, the log-posterior beliefs of each agent is a linear combination of its neighbors' log-beliefs and its log-likelihood function:

$$\nu_{i,t+1}(\theta) = \lambda_i \log \ell_i(s_{t+1}^i|\theta) + \sum_{j\in N_i\cup\{i\}} w_{ij}\nu_{j,t}(\theta) + c_{i,t} \tag{2}$$

where $\nu_{i,t}(\theta) \triangleq \log \mu_{i,t}(\theta)$ is the logarithm of the belief function, $\lambda_i > 0$ is the weight that agent i assigns to its private observations, $w_{ij} > 0$ is the weight assigned to the beliefs of agent j in its neighborhood, and $c_{i,t}$ is

[2] Throughout the paper, $\mathbb{P}$ refers to the probability distribution induced by the true parameter θ^*, and $\mathbb{E}$ denotes expectation with respect to $\mathbb{P}$.

a normalization constant which ensures that $\mu_{i,t+1}(\theta)$ is a well-defined probability density over Θ. Note that constants $c_{i,t}$ do not depend on the parameter θ. Throughout the paper, we assume that $\lambda_i = \lambda$, and $\sum_{j\in N_i\cup\{i\}} w_{ij} = 1$, for all $i \in N$.

Given its beliefs at any given time period, agent i's estimate of the true parameter is defined as a maximizer of its belief function; that is,[3]

$$\hat{\theta}_{i,t} \in \arg\max_{\theta\in\Theta} \nu_{i,t}(\theta). \tag{3}$$

Note that $\hat{\theta}_{i,t}$ is a random variable that depends on the data observed by agents up to time t. In the next section we show that this point estimator always exists and is a measurable function of the data. Moreover, note that due to the identification problem faced by each agent, the maximizer is not necessarily unique at all times. In that case, $\hat{\theta}_{i,t}$ can correspond to any solution of (3).

In order to simplify notation, we write update (2) in matrix form as

$$\nu_{t+1}(\theta) = W\nu_t(\theta) + \lambda\psi_\theta(s_{t+1}) + c_t \qquad \forall\theta\in\Theta$$

where $W = [w_{ij}]$ is a stochastic matrix with $w_{ij} = 0$ if $j \notin N_i \cup \{i\}$, and c_t is a vector of constants independent of θ. Thus, at any time t, we have

$$\nu_t(\theta) = W^t\nu_0(\theta) + \lambda\sum_{\tau=1}^{t} W^{t-\tau}\psi_\theta(s_\tau) + c'_t,$$

where c'_t is a vector that depends on past observations of all agents, but not θ. Finally, we define

$$\Phi_{i,t}(\theta) \triangleq \frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}[W^{t-\tau}]_{ij}\psi^j_\theta(s^j_\tau)$$

which is a function of agents' observations as well as the parameter. Therefore,

$$\nu_{i,t}(\theta) = \lambda t\Phi_{i,t}(\theta) + \sum_{j=1}^{n} W^t_{ij}\nu_{j,0}(\theta) + c'_{i,t} \tag{4}$$

where the second term only depends on the priors and the last term is a constant not depending on θ. This immediately implies that for large enough t, the point estimator $\hat{\theta}_{i,t}$ coincides with the maximizer of $\Phi_{i,t}(\theta)$ over Θ.

III. Consistency

In this section, we prove that under relatively mild assumptions, all agents' estimates of the true parameter are asymptotically consistent in probability; that is, $\hat{\theta}_{i,t} \xrightarrow{p} \theta^*$ for all i as $t \to \infty$. Before presenting our results on consistency, we state a few lemmas. The proofs can be found in the Appendix.

Our first lemma establishes that the point estimator of each agent is well-defined.

[3]Given the fact that log is a monotone function, defining the estimate as the mode of the log-belief function is equivalent to defining it as the maximizer of the belief function itself.

Lemma 1: Suppose that $\theta^* \in \operatorname{int}\Theta$. Then, there exists a measurable function of the data $\hat{\theta}_{i,t}$ that solves (3).

The next lemma shows that the beliefs of all agents converge asymptotically to a limit independent of their priors.

Lemma 2: Suppose that Assumption (C) holds. Then,

$$\Phi_{i,t}(\theta) \xrightarrow{p} \Phi_\infty(\theta) \triangleq \sum_{j=1}^{n} z_j\mathbb{E}[\log\ell_j(s^j_1|\theta)] \tag{5}$$

for all $\theta\in\Theta$, where $z = [z_i]$ is the stationary distribution of a Markov chain with W as its probability transition matrix.

Note that under Assumption (C), matrix W corresponds to an aperiodic and irreducible Markov chain, and therefore, has a unique stationary distribution z, with all elements strictly positive. Moreover, the limiting normalized log-posterior belief function $\Phi_\infty(\theta)$ is independent of i for all values of θ, and as a result, for large enough t, the beliefs of all agents get arbitrarily close. This implies that, as observations accumulate, the agents' estimates get closer to one another.

The next lemma establishes that the limiting log-posterior belief function $\Phi_\infty(\theta)$ is uniquely maximized at the true parameter θ^*, if the truth is globally identifiable and the network of agents is strongly connected.

Lemma 3: Suppose that Assumptions (C) and (GI) hold. Then,

$$\arg\max_{\theta\in\Theta}\ \Phi_\infty(\theta) = \{\theta^*\},$$

where $\Phi_\infty(\theta)$ is defined in (5).

Both Assumptions (C) and (GI) are required for the above lemma to hold. Clearly, in the presence of a global identification problem in the network, there exists a $\theta \neq \theta^*$ for which $\Phi_\infty(\theta) = \Phi_\infty(\theta^*)$ on almost all sample paths, and therefore, the limiting log-posterior belief function is not uniquely maximized. On the other hand, a network which is not strongly connected corresponds to a random walk with some transient states which implies that vector z will have at least one element, say z_k, equal to zero. As a result, the identification problem of agent k persists and leads to a non-unique solution to the maximization problem.

We now present the main result of this section.

Theorem 1: Suppose that $\theta^* \in \operatorname{int}\Theta$ and that Assumptions (C) and (GI) hold. Then, the point estimators of all agents are weakly consistent; that is

$$\hat{\theta}_{i,t} \xrightarrow{p} \theta^* \qquad \forall i.$$

Proof: First, note that for large enough t, the estimate $\hat{\theta}_{i,t}$ coincides with the maximizer of $\Phi_{i,t}(\theta)$ over Θ. On the other hand, by Lemma 2, the convex function $\Phi_{i,t}(\theta)$ converges to $\Phi_\infty(\theta)$ in probability for all θ. As established by Lemma 3, $\Phi_\infty(\theta)$ is uniquely maximized at θ^*, and therefore, by Theorem 2.7 of Newey and McFadden [10], the maximizer of $\Phi_{i,t}(\theta)$ converges in probability to θ^* for all $i \in N$. Thus, the estimator of ever agent is weakly consistent. ■

Theorem 1 establishes that as the number of observations grows, the estimate of each agent converges to the parameter corresponding to the true data generating process. The

importance of this result lies in the fact that asymptotic consistency is achieved despite the fact that all agents face some identification problem – in the sense that no agent can consistently estimate the true parameter in isolation. However, if agents have access to the information held by their neighbors and the communication graph is strongly connected, then information is properly aggregated over the network, and the estimate of every agent converges to the true parameter.

The other notable fact about Theorem 1 is that consistency is achieved regardless of the network's structure. More specifically, as long as the network is strongly connected, its topology and the weights w_{ij} assigned by the agents to their neighbors do not affect convergence of the estimates to the truth. However, in the next sections, we show that the network structure determines the efficiency of the distributed estimators.

IV. Asymptotic Normality

In this section, we prove that the agents' estimates are asymptotically normally distributed and characterize their asymptotic covariance matrices.

We start by stating two auxiliary lemmas, which are proved in the Appendix. Lemma 4 is simply a weak law of large numbers for the Hessian of the log-likelihood of the observations, whereas Lemma 5 is a central limit theorem for the gradients.

Lemma 4: Suppose that $\{\bar{\theta}_{i,t}\}_{i\in N}$ are consistent estimators of θ^*, and suppose Assumption (C) holds. Then,

$$-\nabla_{\theta\theta}\Phi_{i,t}(\bar{\theta}_{i,t}) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*) \qquad \forall i.$$

Lemma 5: Suppose that Assumption (C) holds. Then, for all $i \in N$

$$\sqrt{t}\nabla_\theta \Phi_{i,t}(\theta^*) \xrightarrow{d} \mathcal{N}\big(0, \sum_{j=1}^{n} z_j^2 \mathcal{I}_j(\theta^*)\big).$$

We are now ready to state and prove the main result of this section.

Theorem 2: Suppose that Assumptions (C) and (GI) hold. Then,

$$\sqrt{t}(\hat{\theta}_{i,t} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathrm{Avar}) \tag{6}$$

where the asymptotic covariance matrix is given by

$$\mathrm{Avar} = \left[\sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*)\right]^{-1} \sum_{j=1}^{n} z_j^2 \mathcal{I}_j(\theta^*) \left[\sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*)\right]^{-1}. \tag{7}$$

Proof: By definition, $\hat{\theta}_{i,t}$ is a maximizer of $\Phi_{i,t}(\theta)$, and therefore, it must be the case that $\nabla_\theta \Phi_{i,t}(\hat{\theta}_{i,t}) = 0$. On the other hand, by the mean value theorem, we have

$$\nabla_\theta \Phi_{i,t}(\hat{\theta}_{i,t}) = \nabla_\theta \Phi_{i,t}(\theta^*) + \nabla_{\theta\theta} \Phi_{i,t}(\bar{\theta}_{i,t})(\hat{\theta}_{i,t} - \theta^*),$$

where $\bar{\theta}_{i,t}$ is a mean value between θ^* and $\hat{\theta}_{i,t}$. Thus, we can solve for $(\hat{\theta}_{i,t} - \theta^*)$ and get

$$\sqrt{t}(\hat{\theta}_{i,t} - \theta^*) = -\sqrt{t}\left[\nabla_{\theta\theta} \Phi_{i,t}(\bar{\theta}_{i,t})\right]^{-1} \nabla_\theta \Phi_{i,t}(\theta^*).$$

Since $\bar{\theta}_{i,t}$ lies between θ^* and $\hat{\theta}_{i,t}$, it is a consistent estimator for θ^*,[4] and therefore, Lemma 4 implies that $\nabla_{\theta\theta} \Phi_{i,t}(\bar{\theta}_{i,t}) \xrightarrow{p} -\sum_{j=1}^{n} z_j^2 \mathcal{I}_j(\theta^*)$. Note that the global identifiability assumption guarantees that $\sum_j z_j^2 \mathcal{I}_j(\theta^*)$ is non-singular. On the other hand, Lemma 5 guarantees that $\sqrt{t}\nabla_\theta \Phi_{i,t}(\theta^*) \xrightarrow{d} \mathcal{N}(0, \sum_{j=1}^{n} z_j \mathcal{I}_j(\theta^*))$. At this point, the theorem trivially follows by Slutsky's theorem.[5] ∎

Theorem 2 states that the agents' estimates are normally distributed as the sample size grows. As the proof suggests, the key idea behind asymptotic normality is that in large samples, estimators are approximately equal to linear combinations of sample averages (a consequence of applying the mean value theorem), so that the central limit theorem can be applied [10]. The theorem also states that distributed estimators, like the centralized maximum likelihood estimator, are $\sqrt{t}$-consistent. Finally, expression (7) provides the asymptotic covariance matrix of the estimates in terms of the network structure and information matrices corresponding to agents' observation models.

V. Estimator Efficiency and Network Topology

In the previous section, we derived asymptotic variance of the distributed estimators. In this section, we investigate their efficiency in terms of the network structure, as well as the observation model of each agent. Our next theorem compares the distributed estimator with a centralized estimator, and provides a bound for its performance.

Theorem 3: Suppose that Assumptions (GI) and (C) hold. Then, asymptotic variance of the distributed estimator satisfies

$$\mathrm{Avar} \succeq [\mathcal{I}_c(\theta^*)]^{-1} \tag{8}$$

where $\mathcal{I}_c(\theta)$ denotes the Fisher information matrix of a centralized estimator with access to the observations of all agents. Moreover, the above bound is tight if W is doubly stochastic.

Before presenting the proof, a few remarks are in order. First note that $[\mathcal{I}_c(\theta^*)]^{-1}$ is equal to asymptotic variance of the maximum-likelihood estimator of a centralized entity with access to the measurements of all agents. In other words, equation (8) simply means that the distributed estimators are never more efficient (in the Cramér-Rao sense) than a centralized maximum likelihood estimator. This is not surprising, as one expects that decentralization can never lead to a more efficient estimation.

The second part of the theorem, however, is more striking. It basically states if the weight matrix W is doubly stochastic, then the distributed estimator is as efficient as any centralized estimator. For example, if all communication links are bidirectional and the weights that each pair of agents assign to one another are equal (i.e., $w_{ij} = w_{ji}$), then decentralization does not sacrifice efficiency, regardless of how sparse the network is.

[4] Note that in Theorem 1 we established that $\hat{\theta}_{i,t}$ is consistent.

[5] Slutsky's theorem states that if $x_t \xrightarrow{d} x$ and $y_t \xrightarrow{p} c$ where c is a constant, then, $x_t y_t \xrightarrow{d} cY$.

Proof of Theorem 3: We first compute $\mathcal{I}_c(\theta)$ in terms of the Fisher information matrices corresponding to agents' observation models. By independence of observations across agents, we have

$$\ell(s_t|\theta) = \ell_1(s_t^1|\theta)\ell_2(s_t^2|\theta)\cdots\ell_n(s_t^n|\theta),$$

which implies

$$\begin{aligned}\mathcal{I}_c(\theta^*) &= \mathbb{E}\left[\sum_{j=1}^n \nabla_\theta \psi^j_{\theta^*}(s_1^j)\sum_{i=1}^n \nabla'_\theta \psi^i_{\theta^*}(s_1^i)\right]\\ &= \sum_{j=1}^n \mathbb{E}\left[\nabla_\theta \psi^j_{\theta^*}(s_1^j)\nabla_\theta \psi^j_{\theta^*}(s_1^j)'\right]\\ &= \sum_{j=1}^n \mathcal{I}_j(\theta^*),\end{aligned}$$

where we have used the fact that $\mathbb{E}\left[\nabla_\theta \log \ell_i(s_1^i|\theta^*)\right] = 0$ (see proof of Lemma 5). Therefore, in order to prove (8), we need to show that

$$Q = \sum_{j=1}^n \mathcal{I}_j(\theta^*) - \sum_{j=1}^n z_j\mathcal{I}_j(\theta^*)\left[\sum_{j=1}^n z_j^2\mathcal{I}_j(\theta^*)\right]^{-1}\sum_{j=1}^n \mathcal{I}_j(\theta^*)$$

is positive semi-definite. Note that Q is the Schur complement of

$$X = \begin{bmatrix}\sum_j z_j^2\mathcal{I}_j(\theta^*) & \sum_j z_j\mathcal{I}_j(\theta^*)\\ \sum_j z_j\mathcal{I}_j(\theta^*) & \sum_j \mathcal{I}_j(\theta^*)\end{bmatrix}$$

which can be easily verified to be positive semi-definite.[6] Thus, Q is also positive semi-definite, which proves the first part of the theorem.[7]

To prove the second part, we use the fact that if W is doubly stochastic, then its corresponding Markov chain has a uniform stationary distribution, that is, $z_i = \frac{1}{n}$. Therefore, expression (7) reduces to

$$\text{Avar} = \left[\sum_{j=1}^n \mathcal{I}_j(\theta^*)\right]^{-1} = [\mathcal{I}_c(\theta^*)]^{-1}$$

which is the asymptotic covariance matrix of the centralized maximum likelihood estimator. This proves that the bound is tight. ∎

As a final remark, we emphasize that although sufficient, double stochasticity of W is not necessary for efficiency of the distributed estimator. For example, it is possible to achieve efficiency by assigning a zero weight on an agent whose signals are non-informative, and have the rest of the weights equally shared among the rest of the agents. A complete characterization of efficiency conditions is part of our ongoing research.

[6]Note that $u'Xu = \sum_j (z_j u'_1 + u'_2)\mathcal{I}_j(\theta^*)(z_j u_1 + u_2) \geq 0$ for all $u' = [u'_1\ u'_2]$.

[7]For more on Schur complement and its properties, see for example, Boyd and Vandenberghe [11], page 650.

VI. Conclusions

In this paper, we studied a model of distributed estimation over a network, where each agent faces a local identification problem – in the sense that it cannot consistently estimate a parameter of interest in isolation. The agents engage in communication with their neighbors in order to resolve their identification problems. We showed that as long as the true parameter is globally identifiable (i.e., there is enough information across the network for it to be uniquely identified) and the communication network is strongly connected (i.e., there exists a direct or indirect information path connecting any two agents), then all agents can consistently estimate the true parameter as observations accumulate. Moreover, we proved that under some regularity assumptions on the observation models, the agents' estimates are asymptotically normally distributed. Finally, we computed the asymptotic variance of the distributed estimators, and showed that in bidirectional networks, the agents' estimators are as efficient as any centralized estimator, regardless of the sparsity of the network.

Acknowledgments

The authors would like to thank Ali Jadbabaie for helpful comments and discussions.

Appendix: Omitted Proofs

Proof of Lemma 1: The proof is along the lines of the proof of Lemma 7.1 in Hayashi [12], and therefore, is omitted. ∎

Proof of Lemma 2: We first show that variance of $\Phi_{i,t}(\theta)$ converges to zero, for all i and θ:

$$\begin{aligned}\text{var}[\Phi_{i,t}(\theta)] &= \frac{1}{t^2}\sum_{\tau=1}^t\sum_{j=1}^n [W_{ij}^{t-\tau}]^2\,\text{var}[\psi_\theta^j(s_1^j)]\\ &\leq \frac{1}{t}\sum_{j=1}^n \text{var}[\psi_\theta^j(s_1^j)] \longrightarrow 0,\end{aligned}$$

and therefore, $\Phi_{i,t}(\theta) - \mathbb{E}[\Phi_{i,t}(\theta)] \xrightarrow{p} 0$. On the other hand, we have

$$\begin{aligned}\mathbb{E}[\Phi_{i,t}(\theta)] &= \sum_{j=1}^n \left[\frac{1}{t}\sum_{\tau=1}^t W^{t-\tau}\right]_{ij}\mathbb{E}[\psi_\theta^j(s_1^j)]\\ &\longrightarrow \sum_{j=1}^n [\mathbf{1}z']_{ij}\mathbb{E}[\psi_\theta^j(s_1^j)]\\ &= \sum_{j=1}^n z_j\mathbb{E}[\psi_\theta^j(s_1^j)],\end{aligned}$$

where we used the fact that W corresponds to an aperiodic and irreducible Markov chain with the unique stationary distribution z (guaranteed by Assumption (C)), and that

Cesàro means preserve convergent sequences and their limits. Thus, we have

$$\Phi_{i,t}(\theta) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathbb{E}[\psi_\theta^j(s_1^j)]$$

for all $i \in N$ and all $\theta \in \Theta$, which completes the proof. ■

Proof of Lemma 3: By Jensen's inequality,

$$\mathbb{E}\left[\log \frac{\ell_j(s_1^j|\theta)}{\ell_j(s_1^j|\theta^*)}\right] \leq \log \mathbb{E}\left[\frac{\ell_j(s_1^j|\theta)}{\ell_j(s_1^j|\theta^*)}\right] = 0,$$

implying

$$\mathbb{E}[\log \ell_j(s_1^j|\theta)] \leq \mathbb{E}[\log \ell_j(s_1^j|\theta^*)]$$

with equality holding if and only if $\theta \in \bar{\Theta}_j$. Therefore, the set of maximizers of $\mathbb{E}[\log \ell_j(s_1^j|\theta)]$ coincides with the set of parameters that are observationally equivalent to θ^*. Thus, by Assumption (GI), θ^* is the unique maximizer of their weighted sum. Notice that once again we are using the fact that all elements of vector z are strictly positive. ■

Proof of Lemma 4: First, notice that by a simple weak law of large numbers argument, $\nabla_{\theta\theta}\Phi_{i,t}(\theta) - \mathbb{E}\nabla_{\theta\theta}\Phi_{i,t}(\theta)$ converges to zero in probability, pointwise for all $\theta \in \Theta$. Moreover, we have

$$\mathbb{E}\nabla_{\theta\theta}\Phi_{i,t}(\theta) \longrightarrow \sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_\theta^j(s_1^j)]$$

for all θ, where once again we have used Assumption (C) and the convergence of Cesàro means. Therefore,

$$\nabla_{\theta\theta}\Phi_{i,t}(\theta) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_\theta^j(s_1^j)] \qquad \forall \theta \in \Theta.$$

Now Corollary 2.2 of Newey [13] implies that under Assumptions (A1)–(A5), $\nabla_{\theta\theta}\Phi_{i,t}(\theta)$ converges uniformly in probability to $\sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_\theta^j(s_1^j)]$, and therefore, by Theorem 4.1.5 of Amemiya [14], for any consistent estimator $\bar{\theta}_{i,t} \xrightarrow{p} \theta^*$, we have

$$\nabla_{\theta\theta}\Phi_{i,t}(\bar{\theta}_{i,t}) \xrightarrow{p} \sum_{j=1}^{n} z_j \mathbb{E}[\nabla_{\theta\theta}\psi_{\theta^*}^j(s_1^j)].$$

Finally, the information matrix equality implies that

$$\mathbb{E}[\nabla_{\theta\theta}\psi_{\theta^*}^j(s_1^j)] = -\mathbb{E}\left[\nabla_\theta \psi_{\theta^*}^j(s_1^j) \nabla_\theta \psi_{\theta^*}^j(s_1^j)'\right]$$

which is equal to $-\mathcal{I}_j(\theta^*)$, by definition. This completes the proof. ■

Proof of Lemma 5: The proof of this lemma relies on the multivariate extension of the Lindeberg-Feller central limit theorem, which can be found in van der Vaart [15], Proposition 2.27. But first, notice that by Lemma 3.6 of Newey and McFadden [10], we have

$$\mathbb{E}\left[\nabla_\theta \log \ell_i(s_1^i|\theta^*)\right] = 0,$$

implying that $\mathbb{E}\nabla_\theta \Phi_{i,t}(\theta^*) = 0$.

In order to apply the Lindeberg-Feller CLT, we need to show that the Lindeberg condition is satisfied; that is

$$\frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}(W^{t-\tau})_{ij}^2 \mathbb{E}\left[\|\nabla_\theta \psi_{\theta^*}^j\|^2 \mathbb{I}_{\left\{W_{ij}^{t-\tau}\|\nabla_\theta \psi_{\theta^*}^j\| > \epsilon\sqrt{t}\right\}}\right] \to 0$$

for all $\epsilon > 0$, as $t \to \infty$, where $\mathbb{I}$ denotes the indicator function, and for notational simplicity, we have dropped the dependence of $\nabla_\theta \psi_{\theta^*}^j$ on the observations s^j. Verifying that the Lindeberg condition is straightforward: the left hand-side is bounded above by expression

$$\max_{1\leq j\leq n} \mathbb{E}\left[\|\nabla_\theta \psi_{\theta^*}^j\|^2 \mathbb{I}_{\left\{\|\nabla_\theta \psi_{\theta^*}^j\| > \epsilon\sqrt{t}\right\}}\right]$$

which converges to zero for all $\epsilon > 0$ as $t \to \infty$. Thus, by the Lindeberg-Feller CLT, $\sqrt{t}\Phi_{i,t}(\theta^*) \xrightarrow{d} \mathcal{N}(0, S)$, where S is given by

$$\begin{aligned} S &= \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t}\sum_{j=1}^{n}(W^{t-\tau})_{ij}^2 \mathbb{E}\left[\nabla_\theta \psi_{\theta^*}^j(s_1^j) \nabla_\theta \psi_{\theta^*}^j(s_1^j)'\right] \\ &= \sum_{j=1}^{n} z_j^2 \mathcal{I}_j(\theta^*) \end{aligned}$$

where we have used the fact that $W^t \longrightarrow \mathbf{1}z'$, and the definition of the Fisher information matrix in (1). ■

REFERENCES

[1] P. Kotler, *The Principles of Marketing*, 3rd ed., 1986.

[2] M. Ioannides and L. Loury, "Job information networks, neighborhood effects, and inequality," *The Journal of Economic Literature*, vol. 42, no. 2, pp. 1056–1093, 2004.

[3] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of Groups of Mobile Autonomous Agents Using Nearest Neighbor Rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[4] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks*, ser. Applied Mathematics Series. Princeton University Press, 2009, Electronically available at http://coordinationbook.info.

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

[6] B. Golub and M. O. Jackson, "Naïve learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, vol. 2, pp. 112–149, Feb.

[7] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proceedings of International Conference on Information Processing in Sensor Networks*, Los Angeles, CA, Apr. 2005, pp. 63–70.

[8] A. Jadbabaie, A. Sandroni, and A. Tahbaz-Salehi, "Non-Bayesian social learning," Feb. 2010, PIER Working Paper ♯10-005. [Online]. Available: http://ssrn.com/paper=1550809

[9] S. Kar, J. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," Aug. 2008, Unpublished manuscript.

[10] W. K. Newey and D. McFadden, "Large sample estimation and hypothesis testing," ser. Handbook of Econometrics, R. F. Engle and D. L. McFadden, Eds. Elsevier, 1994, vol. 4, pp. 2111–2245.

[11] S. Boyd and L. Vanderberghe, *Convex Optimization*. New York, NY: Cambridge University Press, 2004.

[12] F. Hayashi, *Econometrics*. Princeton University Press.

[13] W. K. Newey, "Uniform convergence in probability and stochastic equicontinuity," *Econometrica*, vol. 59, no. 4, pp. 1161–1167, Jul. 1991.

[14] T. Amemiya, *Advanced Econometrics*. Cambridge, MA: Harvard University Press, 1985.

[15] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, MA: Cambridge University Press, 2000.