Distributed Parameter Estimation with Exponential Family Statistics: Asymptotic Efficiency

Soummya Kar* and José M. F. Moura*

*Department of ECE, Carnegie Mellon University, Pittsburgh, PA 15213, {soummyak, moura}@ece.cmu.edu

***Abstract*—This paper studies the problem of distributed parameter estimation in multi-agent networks with exponential family observation statistics. Conforming to a given inter-agent communication topology, a distributed recursive estimator of the *consensus-plus-innovations* type is presented in which at every observation sampling epoch the network agents exchange a single round of messages with their communication neighbors and recursively update their local parameter estimates by simultaneously processing the received neighborhood data and the new information (innovation) embedded in the observation sample. Under *global observability* of the networked sensing model and mean connectivity of the inter-agent communication network, the proposed estimator is shown to yield consistent parameter estimates at each network agent. Furthermore, it is shown that the distributed estimator is asymptotically efficient, in that, the asymptotic covariances of the agent estimates coincide with that of the optimal centralized estimator, i.e., the inverse of the centralized Fisher information rate.**

***Index Terms*—Multi-agent networks, distributed estimation, exponential family, collaborative network processing, consensus, stochastic aproximation.**

1. Introduction

Motivated by applications in multi-agent networked information processing, we revisit the problem of distributed sequential parameter estimation. The setup considered is a highly non-classical distributed information setting, in which each network agent samples over time an independent and identically distributed (i.i.d.) time-series which constitute noisy (nonlinear) functions (transformations) of the (vector) parameter of interest with exponential family statistics. Further, in the spirit of typical agent-networking and wireless sensing applications with limited agent communication and computation capabilities, we restrict ourselves to scenarios in which each agent is only aware of its local sensing model and, assuming slotted-discrete time, may only communicate (collaborate) with its agent-neighborhood (possibly dynamic and random) once per epoch of new observation acquisition, i.e., we consider scenarios in which the inter-agent communication rate is at most as high as the observation sampling rate. Broadly speaking, the goal of distributed parameter estimation in such multi-agent scenarios is to update over time the local agent estimates by effectively processing local observation samples and exchanging information with neighboring agents. To this end, the paper presents a distributed estimation approach of the consensus-plus-innovations type, which accomplishes the following:

Consistency under distributed observability: Under *global observability*[1] of the multi-agent sensing model and *mean connectivity* of the inter-agent communication-collaboration network, our distributed estimation approach is shown to yield strongly consistent parameter estimates at each agent. Conversely, it may be readily seen that the conditions of global observability and mean network connectivity are in fact necessary for obtaining consistent parameter estimates in our distributed information-collaboration setup. Indeed, global observability is the minimal requirement for consistency even in centralized estimation, whereas, in the absence of network connectivity, there may be locally unobservable agent-network components which, under no circumstance, will be able to generate consistent parameter estimates.

Asymptotic efficiency: Under the same conditions of global observability of the multi-agent sensing model and mean connectivity of the inter-agent communication-collaboration network, the proposed distributed estimation approach is shown to be asymptotically efficient. In other words, in terms of asymptotic convergence rate, the local agent estimates are as good as the optimal centralized[2], i.e., the local estimates achieve asymptotic covariance equal to the inverse of the centralized Fisher information rate. The key point to note here is that the above optimality holds as long as the mean communication network is connected irrespective of how sparse the link realizations are.

In the context of parallel computing and optimization in multi-agent environments, interacting stochastic gradient and stochastic approximation algorithms have been proposed–see, for example, early work [1], [2]. In contrast, to cope with scenarios where local observations are sensed sequentially over time and inter-agent communication is restricted to arbitrary preassigned topologies and occurs at the same rate as sensing, we have proposed *consensus-plus-innovations* type architectures, see [3]. *Consensus-plus-innovations* algorithms embed a single round of neighborhood consensus or agreement like in [4]–[6], with in addition local processing of the sampled new observation, the local innovation; see for example consensus-plus-innovation approaches for nonlinear distributed estimation [3], detection [7], [8], adaptive control [9] and learning [10]. Other approaches for distributed optimization and inference in multi-agent networks have been considered, see for example diffusion for network inference and optimization [11], [12] and networked LMS and vari-

The work was partially supported by NSF grant 1306128, NSF grant 1011903, and by AFOSR grant FA95501010291.

[1] Global observability means that for every pair of different parameter values, the corresponding probability measures induced on the aggregate or collective agent observation set are *distinguishable*. For setups involving exponential families distinguishability is aptly captured by strict positivity of the Kullback-Liebler (KL) divergence between the corresponding measures, see Assumption 2.2 for details.

[2] The term centralized estimator refers to a hypothetical fusion center based estimator that has access to all agent observations at all times.

ants [11], [13]–[16]. The key distinction between this prior art and the current paper is that, in the former the focus has been mainly on consistency (or minimizing the asymptotic error residual between the estimated and the true parameter), but not on asymptotic efficiency. The requirement of asymptotic efficiency complicates the construction of such distributed algorithms non-trivially and necessitates the use of time-varying consensus and innovation gains in the update process; further these time-varying gains driving the persistent consensus and innovation potentials need to decay at strictly different rates in order for the distributed scheme to achieve the asymptotic covariance of the optimal centralized estimator. Such mixed time-scale construction for asymptotically efficient distributed parameter estimation in linear statistical models was obtained in [17], [18]. However, in contrast to optimal estimation in linear statistical models [17], [18], in the nonlinear non-Gaussian setting, the local innovation gains that achieve asymptotic efficiency are necessarily dependent on the true value of the parameter to be estimated and on the statistics of the global sensing model. Since the value of the parameter (and hence the optimal estimator gains) are not available in advance, our proposed distributed estimation approach involves a distributed online gain learning procedure that proceeds in conjunction with the sequential estimation task. As a result, a closed-loop interaction occurs between the gain learning and parameter estimation that is reminiscent of the certainty-equivalence approach for adaptive estimation and control–although the analysis methodology is significantly different from classical techniques used in adaptive processing (see, for example, [19], [20] and also [21]–[23] in the context of parameter estimation), primarily due to the distributed nature of our problem.

The rest of the paper is organized as follows. Spectral graph theory notation is reviewed next. The multi-agent sensing model is described in Section 2, where we also review some classical concepts from (centralized) estimation theory. Section 3 presents the proposed distributed parameter estimation algorithm and derives its convergence, the main result of the paper. Finally, Section 4 concludes the paper and discusses research avenues for future research.

Detailed proofs of the technical results presented in this paper can be found in the longer manuscript [24].

A. Notation

We denote by $\mathbb{R}$ the set of reals, $\mathbb{R}_+$ the set of non-negative reals, by $\mathbb{R}^k$ the k-dimensional Euclidean space and by $\mathbb{R}^{k\times k}$ the set of $k \times k$ matrices with real entries. Time t is assumed to be discrete or slotted throughout the paper.

Spectral graph theory: The inter-agent communication topology at a given time instant may be described by an *undirected* graph $G = (V, E)$, with $V = [1 \cdots N]$ and E denoting the set of agents (nodes) and inter-agent communication links (edges) respectively. The unordered pair $(n, l) \in E$ if there exists an edge between nodes n and l. We consider simple graphs, i.e., graphs devoid of self-loops and multiple edges. A graph is connected if there exists a path between any pair of nodes. The neighborhood of node n is $\Omega_n = \{l \in V \ : \ (n, l) \in E\}$. The structure of the graph can be described by the symmetric $N \times N$ adjacency matrix, $A = [A_{nl}]$, $A_{nl} = 1$, if $(n, l) \in E$, $A_{nl} = 0$, otherwise. Let the degree matrix be the diagonal matrix $D = \text{diag}\,(d_1 \cdots d_N)$. The positive semidefinite matrix $L = D - A$ is called the graph Laplacian matrix. The eigenvalues of L can be ordered as $0 = \lambda_1(L) \leq \lambda_2(L) \leq \cdots \leq \lambda_N(L)$. The multiplicity of the zero eigenvalue equals the number of connected components of the network; for a connected graph, $\lambda_2(L) > 0$ (see [25]).

2. Multi-agent sensing model

Let $\boldsymbol{\theta}^* \in \mathbb{R}^M$ be an M-dimensional (vector) parameter that is to be estimated by a network of N agents. Throughout, we assume that all the random objects are defined on a common measurable space $(\Omega, \mathcal{F})$ equipped with a filtration $\{\mathcal{F}_t\}$. Probability and expectation, when the true (but unknown) parameter value $\boldsymbol{\theta}^*$ is in force, are denoted by $\mathbb{P}_{\boldsymbol{\theta}^*}(\cdot)$ and $\mathbb{E}_{\boldsymbol{\theta}^*}[\cdot]$ respectively. All inequalities involving random variables are to be interpreted a.s.

Since the *sources of randomness* in our formulation are the observations $\mathbf{y}_n(t)$'s sensed by the network agents at each time $t = 0, 1, \cdots$, and the Laplacian matrices L_t's modeling the stochastic inter-agent communication graphs over time (to be made precise soon), the filtration $\{\mathcal{F}_t\}$ may be taken to be the natural filtration induced by these random quantities, i.e., $\mathcal{F}_t = \sigma\left(\{L_s, \{\mathbf{y}_n(s)\}_{n=1}^{N}\}_{s=0}^{t-1}\right)$ is the σ-algebra induced by the observation and communication processes. Finally, a stochastic process $\{\mathbf{z}_t\}$ is said to be $\{\mathcal{F}_t\}$-adapted if the σ-algebra $\sigma(\mathbf{z}_t)$ is a subset of $\mathcal{F}_t$ at each t; in particular, if $\{\mathcal{F}_t\}$ is the natural filtration induced by the observations and Laplacians, then a process $\{\mathbf{z}_t\}$ is $\{\mathcal{F}_t\}$-adapted if for each t there exists a measurable function $\mathcal{Z}_t(\cdot)$ such that $\mathbf{z}_t = \mathcal{Z}_t\left(\{L_s, \{\mathbf{y}_n(s)\}_{n=1}^{N}\}_{s=0}^{t-1}\right)$.

Each network agent n sequentially observes an independent and identically distributed (i.i.d.) time-series $\{\mathbf{y}_n(t)\}$ of noisy measurements of $\boldsymbol{\theta}^*$, where the distribution $\boldsymbol{\mu}_n^{\boldsymbol{\theta}^*}$ of $\mathbf{y}_n(t)$ belongs to a $\boldsymbol{\theta}$-parameterized exponential family, formalized as follows:

Assumption 2.1. *For each n, let $\boldsymbol{\nu}_n$ be a σ-finite measure on $\mathbb{R}^{M_n}$. Let $g_n : \mathbb{R}^{M_n} \mapsto \mathbb{R}^M$ be a Borel function such that for all $\boldsymbol{\theta} \in \mathbb{R}^M$ the following expectation exists:*

$$\lambda_n(\boldsymbol{\theta}) = \int_{\mathbb{R}^{M_n}} e^{\boldsymbol{\theta}^\top g_n(\mathbf{y}_n)} d\boldsymbol{\nu}_n(\mathbf{y}_n) < \infty. \tag{1}$$

Finally, let $\{\boldsymbol{\mu}_n^{\boldsymbol{\theta}}\}$, for $\boldsymbol{\theta} \in \mathbb{R}^M$, be the corresponding $\boldsymbol{\theta}$-parameterized exponential family of distributions on $\mathbb{R}^{M_n}$ (see [26]), i.e., for each $\boldsymbol{\theta} \in \mathbb{R}^M$ the probability measure $\boldsymbol{\mu}_n^{\boldsymbol{\theta}}$ on $\mathbb{R}^{M_n}$ is given by the Radon-Nikodym derivative

$$\frac{d\boldsymbol{\mu}_n^{\boldsymbol{\theta}}}{d\boldsymbol{\nu}_n}(\mathbf{y}_n) = e^{\left(\boldsymbol{\theta}^\top g_n(\mathbf{y}_n) - \psi_n(\boldsymbol{\theta})\right)} \tag{2}$$

for all $\mathbf{y}_n \in \mathbb{R}^{M_n}$, where $\psi_n(\cdot)$ denotes the function $\psi_n(\boldsymbol{\theta}) = \log \lambda_n(\boldsymbol{\theta})$.

We assume that each network agent n obtains an $\{\mathcal{F}_{t+1}\}$-adapted independent and identically distributed (i.i.d.) sequence $\{\mathbf{y}_n(t)\}$ of observations of the (true) parameter $\boldsymbol{\theta}^*$ with distribution $\boldsymbol{\mu}_n(\boldsymbol{\theta}^*)$, and, for each t, $\mathbf{y}_n(t)$ is independent of $\mathcal{F}_t$. Further, we assume that the observation sequences $\{\mathbf{y}_n(t)\}$ and $\{\mathbf{y}_l(t)\}$ at any two agents n and l are mutually independent.

As an example, consider the familiar linear Gaussian setup in which agent n observes

$$\mathbf{y}_n(t) = H_n\boldsymbol{\theta} + \mathbf{v}_n(t), \tag{3}$$

where $H_n \in \mathbb{R}^{M_n \times M}$ is the local sensing matrix and $\{\mathbf{v}_n(t)\}$ denotes i.i.d. zero-mean Gaussian noise sequence with positive definite covariance matrix R_n. Denoting by $|R_n|$ the determinant of R_n, the probability density function (p.d.f.) of $\mathbf{y}_n(t)$ is given by

$$f_n(\mathbf{y}_n) = \frac{1}{\sqrt{(2\pi)^{M_n}|R_n|}} e^{-\frac{1}{2}(\mathbf{y}_n - H_n\boldsymbol{\theta})^\top R_n^{-1}(\mathbf{y}_n - H_n\boldsymbol{\theta})}, \quad (4)$$

which is the Radon-Nikodym derivative of the multi-variate Gaussian measure $\mathcal{N}(H_n\boldsymbol{\theta}, R_n)$ w.r.t. the Lebesgue measure on $\mathbb{R}^{M_n}$. Rewriting the above as

$$f_n(\mathbf{y}_n) = e^{\boldsymbol{\theta}^\top H_n^\top R_n^{-1}\mathbf{y}_n - (1/2)H_n^\top \boldsymbol{\theta}^\top R_n^{-1} H_n \boldsymbol{\theta}} \quad (5)$$

$$\times \frac{1}{\sqrt{(2\pi)^{M_n}|R_n|}} e^{-\frac{1}{2}\mathbf{y}_n^\top R_n^{-1}\mathbf{y}_n}, \quad (6)$$

we note that the observation statistics satisfy Assumption 2.1, with $\boldsymbol{\nu}_n$ in (2) corresponding to the multi-variate zero-mean Gaussian measure on $\mathbb{R}^{M_n}$ with covariance R_n and $\boldsymbol{\mu}_n^{\boldsymbol{\theta}}$ is absolutely continuous w.r.t. $\boldsymbol{\nu}_n$ with Radon-Nikodym derivative

$$\frac{d\boldsymbol{\mu}_n^{\boldsymbol{\theta}}}{d\boldsymbol{\nu}_n}(\mathbf{y}_n) = e^{\left(\boldsymbol{\theta}^\top g_n(\mathbf{y}_n) - \psi_n(\boldsymbol{\theta})\right)} \quad (7)$$

$$= e^{\boldsymbol{\theta}^\top H_n^\top R_n^{-1}\mathbf{y}_n - (1/2)H_n^\top \boldsymbol{\theta}^\top R_n^{-1} H_n \boldsymbol{\theta}} \quad (8)$$

in the formalism of (2).

We will also denote by $\mathbf{y}_t$ the totality of agent observations at a given time t, i.e., $\mathbf{y}_t = \mathbf{Vec}(\mathbf{y}_n(t)) = \left[\mathbf{y}_1^\top(t), \cdots, \mathbf{y}_N^\top(t)\right]^\top$. For $\boldsymbol{\theta} \in \mathbb{R}^M$ let $\boldsymbol{\mu}^{\boldsymbol{\theta}}$ denote the product measure $\boldsymbol{\mu}_1^{\boldsymbol{\theta}} \otimes \cdots \otimes \boldsymbol{\mu}_N^{\boldsymbol{\theta}}$ on the product space $\mathbb{R}^{M_1} \otimes \cdots \otimes \mathbb{R}^{M_N}$, which means the measures $\boldsymbol{\mu}_n$, $n = 1 \cdots N$, are independent; it is readily seen that $\{\boldsymbol{\mu}^{\boldsymbol{\theta}}\}$ is a $\boldsymbol{\theta}$-parameterized exponential family with respect to (w.r.t.) the product measure $\boldsymbol{\nu} = \boldsymbol{\nu}_1 \otimes \cdots \otimes \boldsymbol{\nu}_N$ and given by the Radon-Nikodym derivatives

$$\frac{d\boldsymbol{\mu}^{\boldsymbol{\theta}}}{d\boldsymbol{\nu}}(\mathbf{y}) = e^{\left(\boldsymbol{\theta}^\top g(\mathbf{y}) - \psi(\boldsymbol{\theta})\right)}, \quad (9)$$

where $\mathbf{y} = \mathbf{Vec}(\mathbf{y}_n)$ denotes a generic element of the product space and the functions $g(\cdot)$ and $\psi(\cdot)$ are given by

$$g(\mathbf{y}) = \sum_{n=1}^{N} g_n(\mathbf{y}_n) \quad \text{and} \quad \psi(\boldsymbol{\theta}) = \sum_{n=1}^{N} \psi_n(\boldsymbol{\theta}) \quad (10)$$

respectively.

It is readily seen that under Assumption 2.1 the global observation sequence $\{\mathbf{y}_t\}$ is $\{\mathcal{F}_{t+1}\}$-adapted, with $\mathbf{y}_t$ being independent of $\mathcal{F}_t$ and distributed as $\boldsymbol{\mu}^{\boldsymbol{\theta}^*}$ (due to mutual independence of the local agent observations) for all t.

For most practical agent network applications, each agent observes only a subset of M_n of the components of the parameter vector, with $M_n \ll M$. It is then necessary for the agents to collaborate by means of occasional local inter-agent message exchanges to achieve a reasonable estimate of the parameter $\boldsymbol{\theta}^*$. To formalize, while we do not require *local observability* for $\boldsymbol{\theta}^*$, we assume that the network sensing model is *globally observable* as follows:

Assumption 2.2. *The network sensing model is globally observable, i.e., we assume $D(\boldsymbol{\theta}, \boldsymbol{\theta}') > 0$ and $D(\boldsymbol{\theta}', \boldsymbol{\theta}) > 0$ for each pair $(\boldsymbol{\theta}, \boldsymbol{\theta}')$ of parameter values, where $D(\boldsymbol{\theta}, \boldsymbol{\theta}')$ denotes the Kullback-Leibler divergence between the distributions $\boldsymbol{\mu}^{\boldsymbol{\theta}}$ and $\boldsymbol{\mu}^{\boldsymbol{\theta}'}$, i.e.,*

$$D(\boldsymbol{\theta}, \boldsymbol{\theta}') = \int_{\mathbf{y}} \log\left(\frac{d\boldsymbol{\mu}^{\boldsymbol{\theta}}}{d\boldsymbol{\mu}^{\boldsymbol{\theta}'}}(\mathbf{y})\right) d\boldsymbol{\mu}^{\boldsymbol{\theta}}(\mathbf{y}). \quad (11)$$

Returning to the linear Gaussian sensing model example (see (3)), the global observability condition in Assumption 2.2 reduces to the invertibility of the aggregate Grammian matrix $\sum_{n=1}^{N} H_n^\top H_n$, where H_n denotes the sensing matrix associated with the n-th agent. Standard results from linear estimation theory confirms that such invertibility (in the linear case) is necessary and sufficient for obtaining consistent estimates of $\boldsymbol{\theta}^*$ in centralized settings.

As a direct consequence of the assumptions stated above, we obtain the following properties on Fisher information matrices associated with the multi-agent sensing model (see [24], [26]).

Proposition 2.1. *Let Assumption 2.1 hold. Then,*

(1) *For each n and $\boldsymbol{\theta} \in \mathbb{R}^M$, let $I_n(\boldsymbol{\theta})$ denote the Fisher information matrix associated with the exponential family $\{\boldsymbol{\mu}_n^{\boldsymbol{\theta}}\}$, i.e.,*

$$I_n(\boldsymbol{\theta}) = -\int_{\mathbf{y}_n} \left(\nabla_{\boldsymbol{\theta}}^2 \frac{d\boldsymbol{\mu}_n^{\boldsymbol{\theta}}}{d\boldsymbol{\nu}_n}(\mathbf{y}_n)\right) d\boldsymbol{\mu}_n^{\boldsymbol{\theta}}(\mathbf{y}_n), \quad (12)$$

where the expectation integral is to be interpreted entry-wise. Then, $I_n(\boldsymbol{\theta})$ is positive semidefinite and satisfies $I_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\left(h_n(\boldsymbol{\theta})\right)$ for all $\boldsymbol{\theta}$, with $h_n(\cdot)$ denoting the function

$$h_n(\boldsymbol{\theta}) = \int_{\mathbf{y}_n \in \mathbb{R}^{M_n}} g_n(\mathbf{y}_n) d\boldsymbol{\mu}_n^{\boldsymbol{\theta}}(\mathbf{y}_n) \quad \forall \boldsymbol{\theta} \in \mathbb{R}^M. \quad (13)$$

(2) *If, in addition, Assumption 2.2 holds, the global Fisher information matrix $I(\boldsymbol{\theta})$, given by,*

$$I(\boldsymbol{\theta}) = -\int_{\mathbf{y}} \left(\nabla_{\boldsymbol{\theta}}^2 \frac{d\boldsymbol{\mu}^{\boldsymbol{\theta}}}{d\boldsymbol{\nu}}(\mathbf{y})\right) d\boldsymbol{\mu}^{\boldsymbol{\theta}}(\mathbf{y}), \quad (14)$$

is positive definite and satisfies

$$I(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}^2 h(\boldsymbol{\theta}) = \sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}}^2 h_n(\boldsymbol{\theta}) = \sum_{n=1}^{N} I_n(\boldsymbol{\theta}) \quad (15)$$

for all $\boldsymbol{\theta} \in \mathbb{R}^M$.

For the multi-agent statistical exponential families under consideration, the well-known Cramér-Rao characterization holds, and it may be shown that the mean-squared estimation error of any (centralized) estimator based on t sets of observation samples from all the agents is lower bounded by the quantity $t^{-1}I^{-1}(\boldsymbol{\theta}^*)$, where $\boldsymbol{\theta}^*$ denotes the true value of the parameter. Making t tend to ∞, the class of asymptotically efficient (optimal) estimators is defined as follows:

Definition 2.1. *An asymptotically efficient estimator of $\boldsymbol{\theta}^*$ is an $\{\mathcal{F}_t\}$-adapted sequence $\{\widehat{\boldsymbol{\theta}}_t\}$, such that $\{\widehat{\boldsymbol{\theta}}_t\}$ is asymptotically normal with asymptotic covariance $I^{-1}(\boldsymbol{\theta}^*)$, i.e.,*

$$\sqrt{t+1}\left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}, I^{-1}(\boldsymbol{\theta}^*)\right), \quad (16)$$

where $\Longrightarrow$ and $\mathcal{N}(\cdot, \cdot)$ denote convergence in distribution and the normal distribution respectively.

Centralized estimators that are asymptotically efficient for the proposed multi-agent setting may be obtained using now-standard results in point estimation theory. For instance, the (centralized) maximum likelihood estimator is known to achieve asymptotic efficiency; however, apart from being centralized, the maximum likelihood estimator is realized in batch form, i.e., requires access to the entire past observation history at all times. To cope with this, extensive research has focused on the development of time-sequential (but centralized) estimators based on recursively processing the agents' observation data $\mathbf{y}_t$; asymptotically efficient recursive centralized estimators of the stochastic approximation type have been developed by several authors, see, for example, [27]–[31], that are asymptotically efficient. In contrast, in this paper, we take a further leap and provide distributed recursive estimators which ensure that each agent obtains an asymptotically efficient estimator of $\boldsymbol{\theta}^*$.

3. Asymptotically Efficient Distributed Estimator

In this section, we provide distributed sequential estimators for $\boldsymbol{\theta}^*$ that are not only consistent but asymptotically optimal, in that, the local asymptotic covariances at each agent coincide with the inverse of the centralized Fisher information rate $I^{-1}(\boldsymbol{\theta}^*)$ associated with the exponential observation statistics in consideration. Specifically, the main idea in the proposed distributed estimation methodology is to generate simultaneously two distributed estimators $\{\breve{\mathbf{x}}_n(t)\}$ and $\{\mathbf{x}_n(t)\}$ at each agent n; the former, the auxiliary estimate sequences $\{\breve{\mathbf{x}}_n(t)\}$, are driven by constant (non-adaptive) innovation gains, and, while supposed to be consistent for $\boldsymbol{\theta}^*$, are suboptimal in the sense of asymptotic covariance. The consistent auxiliary estimates are used to generate the sequence of optimal adaptive innovation gains through another online distributed learning procedure; the resulting adaptive gain process is in turn used to drive the evolution of the desired estimate sequences $\{\mathbf{x}_n(t)\}$ at each agent n, which will be shown to be asymptotically efficient from the asymptotic covariance viewpoint. As will be seen below, we emphasize here that the construction of the auxiliary estimate sequences, the adaptive gain refining, and the generation of the optimal estimators are all executed simultaneously.

A. Algorithms and Assumptions

The proposed optimal distributed estimation methodology consists of the following three simultaneous update processes at each agent n: (i) auxiliary estimate sequence $\{\breve{\mathbf{x}}_n(t)\}$ generation; (ii) adaptive gain refinement; and (iii) optimal estimate sequence $\{\mathbf{x}_n(t)\}$ generation. Formally:

Auxiliary Estimate Generation: Each agent n maintains an $\{\mathcal{F}_t\}$-adapted $\mathbb{R}^M$-valued estimate sequence $\{\breve{\mathbf{x}}_n(t)\}$ for $\boldsymbol{\theta}^*$, recursively updated in a distributed fashion as follows:

$$\breve{\mathbf{x}}_n(t+1) = \breve{\mathbf{x}}_n(t) - \beta_t \sum_{l\in\Omega_n(t)} \left(\breve{\mathbf{x}}_n(t) - \breve{\mathbf{x}}_l(t)\right) \quad (17)$$

$$+\alpha_t \left(g_n(\mathbf{y}_n(t)) - h_n(\breve{\mathbf{x}}_n(t))\right), \quad (18)$$

where $\{\beta_t\}$ and $\{\alpha_t\}$ correspond to appropriate time-varying weighting factors for the agreement (consensus) and innovation (new observation) potentials, respectively, whereas, $\Omega_n(t)$ denotes the $\{\mathcal{F}_{t+1}\}$-adapted time-varying random neighborhood of agent n at time t.

Optimal Estimate Generation: In addition, each agent n generates an optimal (or refined) estimate sequence $\{\mathbf{x}_n(t)\}$, which is also $\{\mathcal{F}_t\}$-adapted and evolves as

$$\mathbf{x}_n(t+1) = \mathbf{x}_n(t) - \beta_t \sum_{l\in\Omega_n(t)} \left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right) \quad (19)$$

$$+\alpha_t K_n(t) \left(g_n(\mathbf{y}_n(t)) - h_n(\mathbf{x}_n(t))\right). \quad (20)$$

Note that the key difference between the estimate updates in (17) and (19) is in the use of adaptive (time-varying) gains $K_n(t)$ in the innovation part in the latter, as opposed to static gains in the former. Specifically, the adaptive gain sequence $\{K_n(t)\}$ at an agent n is an $\{\mathcal{F}_t\}$-adapted $\mathbb{R}^{M\times M}$-valued process which is generated according to a distributed learning process as follows.

Adaptive Gain Refinement: The $\{\mathcal{F}_t\}$-adapted gain sequence $\{K_n(t)\}$ at an agent n is generated according to a distributed learning process, driven by the auxiliary estimates $\{\breve{\mathbf{x}}_n(t)\}$ obtained in (17), as follows:

$$K_n(t) = \left(G_n(t) + \varphi_t I_M\right)^{-1} \quad \forall n, \quad (21)$$

where, $\{\varphi_t\}$ is a deterministic sequence of positive numbers such that $\varphi_t \to 0$ as $t \to \infty$, I_M denotes the $M \times M$ identity matrix, and each agent n maintains another $\{\mathcal{F}_t\}$-adapted matrix-valued process $\{G_n(t)\}$ evolving in a distributed fashion as

$$G_n(t+1) = G_n(t) - \beta_t \left(G_n(t) - G_l(t)\right) \quad (22)$$

$$+\alpha_t \left(I_n(\breve{\mathbf{x}}_n(t)) - G_n(t)\right) \quad (23)$$

for all t, with some positive semidefinite initial condition $G_n(0)$ and $I_n(\cdot)$ denoting the local Fisher information matrix, see (12).

Assumption 3.1. *The $\{\mathcal{F}_{t+1}\}$-adapted sequence $\{L_t\}$ of communication network Laplacians (modeling the agent communication neighborhoods $\Omega_n(t)$-s at each time t) is temporally i.i.d. with L_t being independent of $\mathcal{F}_t$ for each t. Further, the sequence $\{L_t\}$ is connected on the average, i.e., $\lambda_2(\overline{L}) > 0$, where $\overline{L} = \mathbb{E}_{\boldsymbol{\theta}^*}[L_t]$ denotes the mean Laplacian.*

Assumption 3.2. *The weight sequences $\{\beta_t\}$ and $\{\alpha_t\}$ satisfy*

$$\alpha_t = \frac{1}{(t+1)} \quad \textit{and} \quad \beta_t = \frac{b}{(t+1)^{\tau_2}}, \quad (24)$$

where $b > 0$ and $0 < \tau_2 < 1/2$.

Further, the sequence $\{\varphi_t\}$ in (21) *satisfies*

$$\lim_{t\to\infty} (t+1)^{\mu_2}\varphi_t = 0 \quad (25)$$

for some positive constant μ_2.

The following weak linear growth condition on the functions $h_n(\cdot)$ driving the (nonlinear) innovations in (17)-(19) will be assumed:

Assumption 3.3. *For each $\boldsymbol{\theta} \in \mathbb{R}^M$, there exist positive constants $c_1^{\boldsymbol{\theta}}$ and $c_2^{\boldsymbol{\theta}}$, such that, for each n, function $h_n(\cdot)$ in* (13) *satisfies the local linear growth condition,*

$$\left\|h_n(\boldsymbol{\theta}') - h_n(\boldsymbol{\theta})\right\| \leq c_1^{\boldsymbol{\theta}} \left\|\boldsymbol{\theta}' - \boldsymbol{\theta}\right\| + c_2^{\boldsymbol{\theta}}, \quad (26)$$

for all $\boldsymbol{\theta}' \in \mathbb{R}^M$.

B. Main Results

We formally state the main results of the paper concerning the performance of the proposed distributed estimation scheme.

Theorem 3.1. *Let Assumptions 2.2,3.1,3.3 and 3.2 hold. Then, for each n the estimate sequence $\{\mathbf{x}_n(t)\}$ is strongly consistent. In particular, we have*

$$\mathbb{P}_{\boldsymbol{\theta}^*}\left(\lim_{t\to\infty}(t+1)^{\tau}\left\|\mathbf{x}_n(t)-\theta^*\right\|=0\right)=1 \tag{27}$$

for each n and $\tau\in[0,1/2)$.

The consistency in Theorem 3.1 is order optimal in that (27) fails to hold with an exponent $\tau\geq 1/2$ for any (including centralized) estimation procedure.

The next result concerns the asymptotic efficiency of the estimates generated by the proposed distributed scheme.

Theorem 3.2. *Let Assumptions 2.2,3.1,3.3 and 3.2 hold. Then, for each n we have*

$$\sqrt{(t+1)}\left(\mathbf{x}_n(t)-\theta^*\right)\Longrightarrow\mathcal{N}\left(\mathbf{0},I^{-1}(\boldsymbol{\theta}^*)\right), \tag{28}$$

where $\mathcal{N}(\cdot,\cdot)$ and $\Longrightarrow$ denote the Gaussian distribution and weak convergence, respectively.

4. Conclusions

In this paper, we have addressed the problem of distributed parameter estimation in multi-agent networks with generic exponential family observation statistics. Under very weak conditions on the global observability of the agent sensing model and mean connectivity of the inter-agent communication network, we have provided distributed estimators which guarantee consistent and asymptotically efficient (hence, as good as the centralized) estimates at all agents. Natural extensions involve extending the proposed consensus-plus-innovations type architecture to related statistical inference problems such as multi-hypothesis testing, generalized likelihood ratio detection in distributed multi-agent networks with exponential family observation statistics, which we intend to pursue in the future.

References

[1] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, September 1986.

[2] H. Kushner and G. Yin, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *SIAM J. Control Optim.*, vol. 25, no. 5, pp. 1266–1290, Sept. 1987.

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

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

[5] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[6] 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, Jun. 2003.

[7] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: large deviations analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4306–4320, 2012.

[8] S. Kar, R. Tandon, H. Poor, and S. Cui, "Distributed detection in noisy sensor networks," in *IEEE International Symposium on Information Theory*, Saint Petersburg, Russia, July 31 Aug. 5 2011, pp. 2856–2860.

[9] S. Kar, H. Poor, and S. Cui, "Bandit problems in networks: asymptotically efficient distributed allocation rules," in *50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)*, Orlando, FL, Dec. 12-15 2011, pp. 1771–1778.

[10] S. Kar, J. M. F. Moura, and H. Poor, "QD-learning: a collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations," *IEEE Transactions on Signal Processing*, vol. 61, no. 7, pp. 1848–1862, April 2013.

[11] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[12] J. Chen and A. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012.

[13] S. Stankovic, M. Stankovic, and D. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," in *46th IEEE Conference on Decision and Control*, New Orleans, LA, USA, 12-14 Dec. 2007, pp. 1535–1540.

[14] I. Schizas, G. Mateos, and G. Giannakis, "Stability analysis of the consensus-based distributed LMS algorithm," in *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, April 1-4 2008, pp. 3289–3292.

[15] S. Ram, V. Veeravalli, and A. Nedic, "Distributed and recursive parameter estimation in parametrized linear state-space models," *to appear in IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 488– 492, February 2010.

[16] S. Ram, A. Nedić, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.

[17] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE Journal of Selected Topics in Signal Processing: Signal Processing in Gossiping Algorithms Design and Applications*, vol. 5, no. 4, pp. 674–690, August 2011.

[18] S. Kar, J. M. F. Moura, and H. V. Poor, "Distributed linear parameter estimation: asymptotically efficient adaptive strategies," *SIAM J. on Control Optim.*, vol. 51, no. 3, pp. 2200 – 2229, May 2013.

[19] T. L. Lai and C. Z. Wei, "Asymptotically efficient self-tuning regulators," *SIAM J. Control and Optimization*, vol. 25, no. 2, pp. 466–481, March 1987.

[20] T. L. Lai, "Asymptotic properties of nonlinear least squares estimates in stochastic regression models," *The Annals of Statistics*, vol. 2, no. 4, pp. 1917–1930, 1994.

[21] S. Dasgupta and Y.-F. Huang, "Asymptotically convergent modified recursive least-squares with data-dependent updating and forgetting factor for systems with bounded noise," *Information Theory, IEEE Transactions on*, vol. 33, no. 3, pp. 383–392, 1987.

[22] S. Gollamudi, S. Nagaraj, S. Kapoor, and Y.-F. Huang, "Set-membership filtering and a set-membership normalized lms algorithm with an adaptive step size," *Signal Processing Letters, IEEE*, vol. 5, no. 5, pp. 111–114, 1998.

[23] S. Werner, Y.-F. Huang, M. L. De Campos, and V. Koivunen, "Distributed parameter estimation with selective cooperation," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 2849–2852.

[24] S. Kar and J. M. F. Moura, "Asymptotically efficient distributed estimation with exponential family statistics," *IEEE Transactions on Information Theory*, vol. PP, no. 99, Jun. 2014, DOI: 10.1109/TIT.2014.2331272.

[25] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI : American Mathematical Society, 1997.

[26] L. Brown, *Fundamentals of statistical exponential families with applications in statistical decision theory*. Hayward, CA: Institute of Mathematical Statistics, 1986.

[27] D. Sakrison, "Efficient recursive estimation; application to estimating the parameters of a covariance function," *International Journal of Engineering Science*, vol. 3, no. 4, pp. 461–483, 1965.

[28] R. Has'minskij, "Sequential estimation and recursive asymptotically optimal procedures of estimation and observation control," in *Proc. Prague Symp. Asymptotic Statist.*, vol. 1, Charles Univ., Prague, 1974, pp. 157–178.

[29] J. Pfanzagl, "Asymptotic optimum estimation and test procedures," in *Proceedings of the Prague Symposium on Asymptotic Statistics*, vol. 1, Sept. 3 - 6 1973.

[30] C. Stone, "Adaptive maximum likelihood estimators of a location parameter," *The Annals of Statistics*, vol. 3, no. 2, pp. 267–284, Mar. 1975.

[31] V. Fabian, "On asymptotically efficient recursive estimation," *The Annals of Statistics*, vol. 6, no. 4, pp. 854–866, Jul. 1978.