Asymptotically Efficient Distributed Estimation
With Exponential Family Statistics
Soumya Kar and José M. F. Moura

Abstract

The paper studies the problem of distributed parameter estimation in multi-agent networks with exponential family observation statistics. A certainty-equivalence type distributed consensus + innovations estimator is proposed, which, under global observability of the networked sensing model and mean connectivity of the inter-agent communication network, is shown to yield consistent parameter estimates at each network agent. Further, 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. From a technical viewpoint, the proposed distributed estimator leads to non-Markovian mixed time-scale stochastic recursions and the analytical methods developed in the paper contribute to the general theory of distributed stochastic approximation.

Index Terms

Distributed estimation, exponential family, consistency, asymptotic efficiency, stochastic approximation.

1. INTRODUCTION

A. Motivation

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 with exponential family statistics\(^1\) parameterized by the (vector) parameter of interest. 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 observation statistics 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

---

\(^1\)Exponential families subsume most of the distributions encountered in practice, for example, Gaussian, gamma, beta etc.
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 + innovations type, which, among other things, accomplishes the following:

**Consistency under distributed observability:** Under global observability 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. Interestingly, the above leads to the following characterization of distributed observability: distributed observability, i.e., the minimal structural conditions on the sensing and communication models such that there exists a distributed estimation scheme leading to consistent parameter estimates at each network agent, is equivalent to global observability and mean network connectivity.

**Asymptotic efficiency:** Under the same conditions of distributed observability, 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, i.e., they all 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.

Conforming to the sensing and communication architecture, our distributed estimation approach is of the consensus + innovations type, in which at every observation sampling epoch the local agent estimate refinement step embeds a single round of local neighborhood estimate mixing, the consensus or agreement potential, with local processing of the sampled new observation, the innovation potential. Multi-agent stochastic recursive algorithms of the above type have been proposed in prior work – see, for example, early work on parallel stochastic gradient and stochastic approximation; consensus + innovation approaches for nonlinear distributed estimation, detection, adaptive control and learning; diffusion approaches for network inference and optimization; networked LMS and variants. The key distinction between the above and the current work 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

---

2Global 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.
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 [20], [21]. However, in contrast to optimal estimation in linear statistical models [20], [21], 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 (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 which 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, [22], [23] and the references therein, in the context of parameter estimation), primarily due to the distributed nature of our problem. Specifically, in our approach, each agent runs simultaneously three local time recursions: (1) an auxiliary distributed consensus + innovations estimator driven by non-adaptive innovation gains; (2) an online distributed learning procedure that uses the auxiliary distributed local estimators to generate a sequence of optimal adaptive innovations gains; and 3) the desired distributed consensus + innovations estimator whose innovations are weighted by the optimal adaptive innovations gains, thus achieving asymptotic efficiency. We note in this context that the idea of recovering asymptotically efficient estimates from consistent (but suboptimal) auxiliary estimates, although novel from a distributed estimation standpoint, has been investigated in prior work on (centralized) recursive estimation, see, for example, [24], [25].

In summary, in contrast to existing work, the current paper presents a principled development of distributed parameter estimation as applicable to the general and important class of multi-agent statistical exponential families; paralleling the classical development of centralized parameter estimation, it quantifies notions of distributed observability, performance metrics, information measures and algorithmic optimality. Due to the mixed time-scale behavior and the non-Markovianity (induced by the learning process), the stochastic procedure does not fall under the purview of standard stochastic approximation (see, for example, [26]) or distributed stochastic approximation (see, for example, [11], [6–8], [16], [19], [27–29]) procedures. In fact, some of the intermediate results on the pathwise convergence rates of mixed time-scale stochastic procedures obtained in the paper are more broadly applicable and contribute to the general theory of distributed stochastic approximation. In this context, we note the study of mixed time-scale stochastic procedures that arise in algorithms of the simulated annealing type (see, for example, [30]). Apart from being distributed, our scheme technically differs from [30] in that, whereas the additive perturbation in [30] is a martingale difference sequence, ours is a network dependent consensus potential manifesting past dependence. In fact, intuitively, a key step in the analysis is to derive pathwise strong approximation results to characterize the rate at which the consensus term/process converges to a martingale difference process. We also emphasize that our notion of mixed time-scale is different from that of stochastic algorithms with coupling (see [31], [32]), where a quickly switching parameter influences the relatively slower dynamics of another state, leading to averaged dynamics. Mixed time scale procedures of this latter type arise in multi-scale distributed
information diffusion problems, see, in particular, the paper \cite{33}, that studies interactive consensus formations in Markov-modulated switching networks.

We comment briefly on the organization of the rest of the paper. Section 1-B sets up notation. The multi-agent sensing model is formalized in Section 2-A whereas, preliminary facts pertaining to the model and assumptions are summarized in Section 2-B. Section 3-A describes the distributed estimation approach and the main results of the paper (concerning consistency and asymptotic efficiency of the proposed approach) are stated in Section 3-B. The major technical developments are presented in Section 4 culminating to the proofs of the main results in Section 5. Finally, Section 6 concludes the paper.

B. Notation

We denote by $\mathbb{R}$ the set of reals, $\mathbb{R}_+$ the set of non-negative reals, and by $\mathbb{R}^k$ the $k$-dimensional Euclidean. For $a, b \in \mathbb{R}$, we use $a \vee b$ and $a \wedge b$ to denote the maximum and minimum of $a$ and $b$ respectively. For deterministic $\mathbb{R}_+$-valued sequences $\{a_t\}$ and $\{b_t\}$, the notation $a_t = O(b_t)$ denotes the existence of a constant $c > 0$ such that $a_t \leq c b_t$ for all $t$ sufficiently large. Further, the notation $a_t = o(b_t)$ is used to indicate that $a_t/b_t \to 0$ as $t \to \infty$. For $\mathbb{R}_+$-valued stochastic processes $\{a_t\}$ and $\{b_t\}$, the corresponding order notations are to be interpreted to hold pathwise almost surely (a.s.).

The set of $k \times k$ real matrices is denoted by $\mathbb{R}^{k \times k}$. The corresponding subspace of symmetric matrices is denoted by $\mathbb{S}^k$. The cone of positive semidefinite matrices is denoted by $\mathbb{S}^k_+$, whereas $\mathbb{S}^k_++$ denotes the subset of positive definite matrices. The $k \times k$ identity matrix is denoted by $I_k$, while $1_k$ and $0_k$ denote respectively the column vector of ones and zeros in $\mathbb{R}^k$. Often the symbol $0$ is used to denote the $k \times p$ zero matrix, the dimensions being clear from the context. The symbol $\top$ denotes matrix transpose, whereas, for a finite set of matrices $A_n \in \mathbb{R}^{k_n \times p}$, $n = 1, \cdots, N$, the quantity $\text{Vec}(A_n)$ denotes the $(k_1 + \cdots + k_N) \times p$ matrix $[A_1^\top, \cdots, A_N^\top]^\top$ obtained as the (column-wise) stack of the matrices $A_n$. The operator $\|\cdot\|$ applied to a vector denotes the standard Euclidean $L_2$ norm, while applied to matrices it denotes the induced $L_2$ norm, which is equivalent to the matrix spectral radius for symmetric matrices. Also, for $a \in \mathbb{R}^k$ and $\varepsilon > 0$, we will use $B_\varepsilon(a)$ to denote the closed $\varepsilon$-neighborhood of $a$, i.e.,

$$B_\varepsilon(a) = \{b \in \mathbb{R}^k : \|b - a\| \leq \varepsilon\}.$$  

The notation $A \otimes B$ is used to denote the Kronecker product of two matrices $A$ and $B$.

The following notion of consensus subspace and its complement will be used:

**Definition 1.1.** Let $N$ and $M$ be positive integers and consider the Euclidean space $\mathbb{R}^{NM}$. The consensus or agreement subspace $C$ of $\mathbb{R}^{NM}$ is then defined as

$$C = \{z \in \mathbb{R}^{NM} : z = 1_N \otimes a \text{ for some } a \in \mathbb{R}^M\}.$$
The orthogonal complement of $C$ in $\mathbb{R}^{NM}$ is denoted by $C^\perp$. Finally, for a given vector $z \in \mathbb{R}^{NM}$, its projection on the consensus subspace $C$ is to be denoted by $z_C$, whereas, $z_{C^\perp} = z - z_C$ denotes the projection on the orthogonal complement $C^\perp$.

Also, for $z \in C$, we will denote by $z^a$ the vector $a \in \mathbb{R}^M$ such that $z = 1_N \otimes a$.

Time is assumed to be discrete or slotted throughout the paper. The symbols $t$ and $s$ denote time, and $\mathbb{T}_+$ is the discrete index set $\{0, 1, 2, \cdots \}$. The parameter to be estimated belongs to a subset $\Theta$ (generally open) of the Euclidean space $\mathbb{R}^M$. We reserve the symbol $\theta$ to denote a canonical element of the parameter space $\Theta$, whereas, the true (but unknown) value of the parameter (to be estimated) is denoted by $\theta^*$. The symbol $x_n(t)$ is used to denote the $\mathbb{R}^M$-valued estimate of $\theta^*$ at time $t$ at agent $n$. Without loss of generality, the initial estimate, $x_n(0)$, at time 0 at agent $n$ is assumed to be a non-random quantity.

**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\(^3\) between each pair of nodes. The neighborhood of node $n$ is

$$\Omega_n = \{l \in V \mid (n, l) \in E\}.$$  

Node $n$ has degree $d_n = |\Omega_n|$ (the number of edges with $n$ as one end point.) 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 $D = \text{diag}(d_1, \cdots d_N)$. By definition, 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 eigenvector corresponding to $\lambda_1(L)$ being $(1/\sqrt{N})1_N$. The multiplicity of the zero eigenvalue equals the number of connected components of the network; for a connected graph, $\lambda_2(L) > 0$. This second eigenvalue is the algebraic connectivity or the Fiedler value of the network; see [34] for detailed treatment of graphs and their spectral theory.

### 2. Multi-agent sensing model

Let $\theta^* \in \mathbb{R}^M$ 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 $\theta^*$ is in force, are denoted by $\mathbb{P}_{\theta^*}(\cdot)$ and $\mathbb{E}_{\theta^*}[\cdot]$ respectively. All inequalities involving random variables are to be interpreted a.s.

#### A. Sensing Model

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

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

$$
\lambda_n(\theta) = \int_{\mathbb{R}^{M_n}} e^{\theta^\top g_n(y_n)} d\nu_n(y_n) < \infty.
$$

(1)

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

$$
d\mu^\theta_n(y_n) = e^{\theta^\top g_n(y_n) - \psi_n(\theta)}
$$

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

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

We will also denote by $y_t$ the totality of agent observations at a given time $t$, i.e., $y_t = \text{Vec}(y_n(t)) = [y_1^\top(t), \ldots, y_N^\top(t)]^\top$. For $\theta \in \mathbb{R}^M$ let $\mu^\theta$ denote the product measure $\mu_1^\theta \otimes \cdots \otimes \mu_N^\theta$ on the product space $\mathbb{R}^{M_1} \otimes \cdots \otimes \mathbb{R}^{M_N}$; it is readily seen that $\{\mu^\theta\}$ is a $\theta$-parameterized exponential family with respect to (w.r.t.) the product measure $\nu = \nu_1 \otimes \cdots \otimes \nu_N$ and given by the Radon-Nikodym derivatives

$$
\frac{d\mu^\theta}{d\nu}(y) = e^{\theta^\top g(y) - \psi(\theta)},
$$

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

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

(2)

respectively.

It is readily seen that under Assumption 2.1 the global observation sequence $\{y_t\}$ is $\{\mathcal{F}_{t+1}\}$-adapted, with $y_t$ being independent of $\mathcal{F}_t$ and distributed as $\mu^{\theta^*}$ 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 $\theta^*$. To formalize, while we do not require local observability for $\theta^*$, 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 $\tilde{\mathcal{R}}(\theta, \theta') > 0$ and $\tilde{\mathcal{R}}(\theta', \theta) > 0$ for each pair $(\theta, \theta')$ of parameter values, where $\tilde{\mathcal{R}}(\theta, \theta')$ denotes the Kullback-Leibler divergence between the
distributions $\mu^\theta$ and $\mu^{\theta'}$, i.e.,
\[ \mathcal{R}(\theta, \theta') = \int_y \log \left( \frac{d\mu^\theta(y)}{d\mu^{\theta'}(y)} \right) \, d\mu^\theta(y). \]

B. Some preliminaries

We state some useful analytical properties associated with the multi-agent sensing model, in particular, the implications of the global observability condition (see Assumption 2.2). Most of the listed properties are direct consequences of standard analytical arguments involving statistical exponential families, see, for example, [35].

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

1. For each $n$, the function $\psi_n(\cdot)$ is infinitely differentiable on $\mathbb{R}^M$.

2. For each $n$, let $h_n : \mathbb{R}^M \mapsto \mathbb{R}^M$ be the gradient of $\psi_n(\cdot)$, i.e., $h_n(\theta) = \nabla_\theta \psi_n(\theta)$ for all $\theta \in \mathbb{R}^M$. Then

\[ h_n(\theta) = \int_{y_n \in \mathbb{R}^M_n} g_n(y_n) d\mu^\theta_n(y_n) \quad \forall \theta \in \mathbb{R}^M \tag{3} \]

and the following inequality (monotonicity) holds for each pair $(\theta, \theta')$ in $\mathbb{R}^M$:
\[ (\theta - \theta')^\top (h_n(\theta) - h_n(\theta')) \geq 0. \tag{4} \]

3. If, in addition, Assumption 2.2 holds, denoting by $h(\cdot)$ the gradient of $\psi(\cdot)$, see [2], we have the following strict monotonicity

\[ (\theta - \theta')^\top (h(\theta) - h(\theta')) = \sum_{n=1}^N (\theta - \theta')^\top (h_n(\theta) - h_n(\theta')) > 0 \]

for each pair $(\theta, \theta')$ in $\mathbb{R}^M$ such that $\theta \neq \theta'$.

**Proof:** The first assertion is an immediate consequence of the fact that the function $\psi_n(\theta)$ associated with the exponential family $\{\mu^\theta_n\}$ is infinitely differentiable on the interior of the natural parameter space (the set on which the expectation in (1) exists), see Theorem 2.2 in [35]. The second assertion constitutes a well-known property of statistical exponential families (see Corollary 2.5 in [35]). The same corollary in [35] asserts that the inequality in (4) is strict whenever the measures $\mu^\theta$ and $\mu^{\theta'}$ are different for $\theta \neq \theta'$, the latter being ensured by the positivity of the Kullback-Leibler divergences as in Assumption 2.2.

The next proposition characterizes the information matrices (or Fisher matrices) associated with the sensing model and may be stated as follows (see [35] for a proof):

**Proposition 2.2.** Let Assumption 2.1 hold. Then,

For a function $f(\theta)$, $\nabla_\theta f(\theta) \in \mathbb{R}^M$ denotes the vector of partial derivatives, i.e., the $i$-th component of $\nabla_\theta f(\theta)$ is given by $\frac{\partial f(\theta)}{\partial \theta_i}$. The Hessian $\nabla^2_\theta f(\theta) \in \mathbb{R}^{M \times M}$ denotes the matrix of second order partial derivatives, whose $i, j$-th entry corresponds to $\frac{\partial^2 f(\theta)}{\partial \theta_i \partial \theta_j}$. 

\[ \frac{\partial}{\partial \theta_i} \nabla_\theta f(\theta) \in \mathbb{R}^M \]

\[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(\theta) \in \mathbb{R}^{M \times M} \]
(1) For each \( n \) and \( \theta \in \mathbb{R}^M \), let \( I_n(\theta) \) denote the Fisher information matrix associated with the exponential family \( \{\mu_{\theta}^n\} \), i.e.,
\[
I_n(\theta) = -\int_{y_n} \left( \nabla_{\theta}^2 \frac{d\mu_{\theta}^n}{d\nu_n}(y_n) \right) d\mu_{\theta}^n(y_n),
\]
where the expectation integral is to be interpreted entry-wise. Then, \( I_n(\theta) \) is positive semidefinite and satisfies \( I_n(\theta) = \nabla_{\theta} (h_n(\theta)) \) for all \( \theta \), with \( h_n(\cdot) \) denoting the function in (3).

(2) If, in addition, Assumption 2.2 holds, the global Fisher information matrix \( I(\theta) \), given by,
\[
I(\theta) = -\int_{y} \left( \nabla_{\theta}^2 \frac{d\mu_{\theta}}{d\nu}(y) \right) d\mu_{\theta}(y),
\]
is positive definite and satisfies
\[
I(\theta) = \nabla_{\theta} h(\theta) = \sum_{n=1}^{N} \nabla_{\theta}^2 h_n(\theta) = \sum_{n=1}^{N} I_n(\theta)
\]
for all \( \theta \in \mathbb{R}^M \).

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 \( t^{-1}I^{-1}(\theta^*) \), where \( \theta^* \) denotes the true value of the parameter. Making \( t \) tend to \( \infty \), the class of asymptotically efficient (optimal) estimators is defined as follows:

**Definition 2.1.** An asymptotically efficient estimator of \( \theta^* \) is an \( \{\mathcal{F}_t\} \)-adapted sequence \( \{\hat{\theta}_t\} \), such that \( \{\hat{\theta}_t\} \) is asymptotically normal with asymptotic covariance \( I^{-1}(\theta^*) \), i.e.,
\[
\sqrt{t+1} \left( \hat{\theta}_t - \theta^* \right) \Rightarrow \mathcal{N} \left( 0, I^{-1}(\theta^*) \right),
\]
where \( \Rightarrow \) and \( \mathcal{N} \left( \cdot, \cdot \right) \) denote convergence in distribution and the normal distribution respectively.

**Remark 2.1.** 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; specifically, there exists an \( \{\mathcal{F}_t\} \)-adapted sequence \( \{\hat{\theta}_t\} \), such that
\[
\hat{\theta}_t \in \arg\max_{\theta \in \mathbb{R}^M} \left( \sum_{s=0}^{t-1} \log \frac{d\mu_{\theta}}{d\nu}(y_s) \right)
\]
a.s. for all \( t \), and \( \{\hat{\theta}_t\} \) is asymptotically normal with asymptotic covariance \( I^{-1}(\theta^*) \). Note that, apart from being centralized, the maximum likelihood estimator, as implemented above, consists of a batch-form realization. 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 \( y_t \); asymptotically efficient recursive centralized estimators of

---

5The term centralized estimator refers to a hypothetical fusion center based estimator that has access to all agent observations at all times.
the stochastic approximation type have been developed by several authors, see, for example, [24], [25], [30]–[38],
that are asymptotically efficient.

We emphasize that the centralized (recursive or batch) estimators, as discussed above, are based on the availability
of the entire set of agent observations at a centralized resource at all times, which further require the global model
information (the statistics of the agent exponential families \( \{ \mu^\theta_n \} \)) for all \( n \) such that the nonlinear innovation gains
driving the recursive estimators may be designed appropriately to achieve asymptotic efficiency. In contrast, the goal
of this paper is to develop collaborative distributed asymptotically efficient estimators of \( \theta^* \) at each agent \( n \) of the
network, in which, (i) the information is distributed, i.e., at a given instant of time \( t \) each agent \( n \) has access to its
local sensed data \( y_n(t) \) only; (ii) to start with, each agent \( n \) is only aware of its local sensing model \( \{ \mu^\theta_n \} \) only; and,
(iii) the agents may only collaborate by exchanging information over a (sparse) pre-defined communication network,
where inter-agent communication and observation sampling occurs at the same rate, i.e., each agent \( n \) may only
exchange one round of messages with its designated communication neighbors per sampling epoch. To this end, the
proposed estimators consist of simultaneous distributed local estimate update and distributed local gain refinement
(learning) at each network agent \( n \), with closed-loop interaction between the estimation and learning processes.

From a technical point of view, in contrast to centralized stochastic approximation based estimators, the estimators
developed in the paper are of the distributed nonlinear stochastic approximation type with necessarily mixed time-
scale dynamics; the mixed time-scale dynamics arise as a result of suitably crafting the relative intensities of the
potentials for local collaboration and local innovation, necessary for achieving asymptotic efficiency. Distributed
estimators of mixed time-scale dynamics have been introduced and studied in [8], [20]; we refer to them as consensus
+ innovations estimators.

3. ASYMPTOTICALLY EFFICIENT DISTRIBUTED ESTIMATOR

In this section, we provide distributed sequential estimators for \( \theta^* \) 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 \( I^{-1}(\theta^*) \) associated with the exponential observation statistics in consideration. Other than
challenges encountered in the distributed implementation, a major difficulty in obtaining such asymptotically efficient
distributed estimators concerns the design of the local estimator or innovation gains (to be made precise later); in
particular, in contrast to optimal estimation in linear statistical models [20], [21], in the nonlinear non-Gaussian
setting, the innovation gains that achieve asymptotic efficiency are necessarily dependent on the true value \( \theta^* \) of the
parameter to be estimated. Since the value of \( \theta^* \) (and hence the optimal estimator gains) are not available in advance.
We propose a distributed estimation approach that involves a distributed online gain learning procedure that proceeds
in conjunction with the sequential estimation task. As a result, a somewhat closed-loop interaction occurs between
the gain learning and parameter estimation that is reminiscent of the certainty equivalence approach to adaptive
estimation and control – although the analysis methodology is significantly different from classical techniques used
in adaptive processing, primarily due to the distributed nature of our problem and its mixed time-scale dynamics.
Specifically, the main idea in the proposed distributed estimation methodology is to generate simultaneously two distributed estimators \{\hat{x}_n(t)\} and \{x_n(t)\} at each agent \(n\); the former, the auxiliary estimate sequences \{\hat{x}_n(t)\}, are driven by constant (non-adaptive) innovation gains, and, while supposed to be consistent for \(\theta^*\), 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 \{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 \{\hat{x}_n(t)\} generation; (ii) adaptive gain refinement; and (iii) optimal estimate sequence \{x_n(t)\} generation. Formally:

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

\[
\hat{x}_n(t+1) = \hat{x}_n(t) - \beta_t \sum_{l \in \Omega_n(t)} (\hat{x}_n(t) - \hat{x}_l(t)) + \alpha_t \left(g_n(y_n(t)) - h_n(\hat{x}_n(t))\right),
\]

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 \{x_n(t)\}, which is also \(\{\mathcal{F}_t\}\)-adapted and evolves as

\[
x_n(t+1) = x_n(t) - \beta_t \sum_{l \in \Omega_n(t)} (x_n(t) - x_l(t)) + \alpha_t K_n(t) \left(g_n(y_n(t)) - h_n(x_n(t))\right).
\]

Note that the key difference between the estimate updates in (7) and (8) 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 \{\hat{x}_n(t)\} obtained in (7), as follows:

\[
K_n(t) = (G_n(t) + \varphi_t I_M)^{-1} \quad \forall n,
\]

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

\[
G_n(t+1) = G_n(t) - \beta_t (G_n(t) - G_l(t)) + \alpha_t (I_n(\hat{x}_n(t)) - G_n(t))
\]
for all $t$, with some positive semidefinite initial condition $G_n(0)$ and $I_n(\cdot)$ denoting the local Fisher information matrix, see (5).

Before discussing further, we formalize the assumptions on the inter-agent stochastic communication and the algorithm weight sequences $\{\alpha_t\}$ and $\{\beta_t\}$ in the following:

**Assumption 3.1.** The $\{F_{t+1}\}$-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 $F_t$ for each $t$. Further, the sequence $\{L_t\}$ is connected on the average, i.e., $\lambda_2(L) > 0$, where $L = \mathbb{E}_{\theta^*}[L_t]$ denotes the mean Laplacian.

**Assumption 3.2.** The weight sequences $\{\beta_t\}$ and $\{\alpha_t\}$ satisfy
\[
\alpha_t = \frac{1}{(t+1)} \quad \text{and} \quad \beta_t = \frac{b}{(t+1)^{\tau_2}},
\]
where $b > 0$ and $0 < \tau_2 < 1/2$.

Further, the sequence $\{\varphi_t\}$ in (9) satisfies
\[
\lim_{t \to \infty} (t+1)^{\mu_2} \varphi_t = 0
\]
for some positive constant $\mu_2$.

The following weak linear growth condition on the functions $h_n(\cdot)$ driving the (nonlinear) innovations in (7)-(8) will be assumed:

**Assumption 3.3.** For each $\theta \in \mathbb{R}^M$, there exist positive constants $c_{\theta}^1$ and $c_{\theta}^2$, such that, for each $n$, function $h_n(\cdot)$ in (3) satisfies the local linear growth condition,
\[
\|h_n(\theta') - h_n(\theta)\| \leq c_{\theta}^1 \|\theta' - \theta\| + c_{\theta}^2,
\]
for all $\theta' \in \mathbb{R}^M$.

**B. Main Results**

We formally state the main results of the paper, the proofs appearing in Section 5.

**Theorem 3.1.** Let Assumptions 2.2, 3.1, 3.3 and 3.2 hold. Then, for each $n$ the estimate sequence $\{x_n(t)\}$ is strongly consistent. In particular, we have
\[
\mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} (t+1)^{\tau} \|x_n(t) - \theta^*\| = 0 \right) = 1
\]
for each $n$ and $\tau \in [0, 1/2)$.

The consistency in Theorem 3.1 is order optimal in that (13) 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, 3.2 and 3.2 hold. Then, for each \( n \) we have

\[
\sqrt{(t + 1)} (\mathbf{x}_n(t) - \theta^*) \implies \mathcal{N} \left( 0, \Gamma^{-1}(\theta^*) \right),
\]

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

## 4. A Generic Consistent Distributed Estimator

With a view to understanding the asymptotic behavior of the auxiliary estimate processes \( \{\hat{\mathbf{x}}_n(t)\}, \ n = 1, \cdots, N \), introduced in Section 3-A, see 7, we study a somewhat more general class of distributed estimate processes with time-varying local innovation gains. Other than establishing consistency of these estimates (see Theorem 4.1), we obtain pathwise convergence rate asymptotics of the estimate processes to \( \theta^* \) (see Theorem 4.2). These latter convergence rate results will be used to analyze the impact of the auxiliary estimates in the adaptive gain computation (9)-(10).

**Theorem 4.1.** For each \( n \), let \( \{\mathbf{z}_n(t)\} \) be an \( \mathbb{R}^M \)-valued \( \{\mathcal{F}_t\} \)-adapted process (estimator) evolving as follows:

\[
\mathbf{z}_n(t + 1) = \mathbf{z}_n(t) - \beta_t \sum_{n \in \Omega_n(t)} (\mathbf{z}_n(t) - \mathbf{z}_n(t)) + \alpha_t \mathbf{K}_n(t) (g_{n}(\mathbf{y}_n(t)) - h_{n}(\mathbf{z}_n(t))).
\]

Suppose Assumptions 2.2, 3.3 and 3.4 on the network system model hold, and the weight sequences \( \{\beta_t\} \) and \( \{\alpha_t\} \) satisfy Assumption 4.2 Additionally, let the matrix gain processes \( \{\mathbf{K}_n(t)\} \) be \( \mathbb{S}^M_+ \)-valued \( \{\mathcal{F}_t\} \)-adapted, and there exist a positive definite matrix \( \mathcal{K} \) and a constant \( \tau_3 > 0 \), such that the gain processes \( \{\mathbf{K}_n(t)\} \) converge uniformly to \( \mathcal{K} \) at rate \( \tau_3 \), i.e., for each \( \delta > 0 \), there exists a deterministic time \( t_\delta \), such that for all \( n \)

\[
\mathbb{P}_{\theta^*} \left( \sup_{t \geq t_\delta} \| \mathbf{K}_n(t) - \mathcal{K} \| \leq \delta \right) = 1.
\]

Then, for each \( n \), \( \{\mathbf{z}_n(t)\} \) is a consistent estimator of \( \theta^* \), i.e., \( \mathbf{z}_n(t) \to \theta^* \) as \( t \to \infty \) a.s.

The proof of Theorem 4.1 is accomplished in steps, the key intermediate ingredients being Lemma 4.1 and Proposition 4.1 concerning the boundedness of the processes \( \{\mathbf{z}_n(t)\}, \ n = 1, \cdots, N \), and a Lyapunov type-construction, respectively.

**Lemma 4.1.** Let the hypotheses of Theorem 4.1 hold. Then, for each \( n \), the process \( \{\mathbf{z}_n(t)\} \) is bounded a.s., i.e.,

\[
\mathbb{P}_{\theta^*} \left( \sup_{t \geq 0} \| \mathbf{z}_n(t) \| < \infty \right) = 1.
\]

**Proof:** Let \( \tilde{z}_n(t) = \mathbf{z}_n(t) - \theta^* \) and denote by \( \mathbf{z}_t, \hat{\mathbf{z}}_t \) and \( \tilde{\theta} \) the \( \mathbb{R}^N \)-valued \( \text{Vec}(\mathbf{z}_n(t)) \), \( \text{Vec}(\hat{\mathbf{z}}_n(t)) \), and \( \mathbf{1}_N \otimes \theta^* \), respectively. Noting that \( (L_t \otimes I_M) (\mathbf{1}_N \otimes \theta^*) = 0 \), the process \( \{\tilde{z}_t\} \) is seen to satisfy

\[
\tilde{z}_{t+1} = \tilde{z}_t - \beta_t (L_t \otimes I_M) \tilde{z}_t - \alpha_t \mathbf{K}_t \left( \tilde{\mathbf{h}}_t - \tilde{\mathbf{h}}(\tilde{\theta}) \right) + \alpha_t \mathbf{K}_t \left( \tilde{\mathbf{g}}(\mathbf{y}_t) - \tilde{\mathbf{h}}(\tilde{\theta}) \right),
\]

where \( \tilde{\mathbf{h}}(\tilde{\theta}) \) and \( \tilde{\mathbf{h}}_t \) denote the Gaussian distribution and weak convergence, respectively.
where
\[
\overline{h}(z_t) = \text{Vec} \left( h_n(z_n(t)) \right), \quad \overline{h}(\theta^*) = \text{Vec} \left( h_n(\theta^*) \right), \quad \overline{g}(y_t) = \text{Vec} \left( g_n(y_n(t)) \right),
\] (17)
and \( \overline{K}_t = \text{Diag}(K_n(t)) \). Note that, by hypothesis, \( \overline{K}_t \in S_{++}^{N \times N} \) and define the \( \mathbb{R}_+ \)-valued \( \{ F_t \} \)-adapted process \( \{ V_t \} \) by
\[
V_t = \tilde{z}_t^\top \left( I_N \otimes \mathcal{K}^{-1} \right) \tilde{z}_t
\] (18)
for all \( t \). Note that by (16) we obtain
\[
( I_N \otimes \mathcal{K} )^{-1} \tilde{z}_{t+1} = ( I_N \otimes \mathcal{K} )^{-1} \tilde{z}_t - \beta_t \left( L_t \otimes \mathcal{K}^{-1} \right) \tilde{z}_t
\]
\[
- \alpha_t ( I_N \otimes \mathcal{K} )^{-1} \overline{K}_t \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right) + \alpha_t ( I_N \otimes \mathcal{K} )^{-1} \overline{K}_t \left( \overline{g}(y_t) - \overline{h}(\theta^*) \right).
\]
By (3) we have for all \( t \geq 0 \)
\[
\mathbb{E}_{\theta^*} \left[ \overline{g}(y_t) - \overline{h}(\theta^*) \right] = 0,
\]
and using the temporal independence of the Laplacian sequence we obtain
\[
\mathbb{E}_{\theta^*} \left[ V_{t+1} \mid F_t \right] = V_t - 2 \beta_t \tilde{z}_t^\top ( \mathcal{L} \otimes \mathcal{K}^{-1} ) \tilde{z}_t - 2 \alpha_t \tilde{z}_t^\top ( I_N \otimes \mathcal{K}^{-1} ) K_t \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right)
\]
\[
+ \beta_t^2 \tilde{z}_t^\top \mathbb{E}_{\theta^*} \left[ ( \mathcal{L} \otimes I_M ) \left( I_N \otimes \mathcal{K}^{-1} \right) ( \mathcal{L} \otimes I_M ) \right] \tilde{z}_t
\]
\[
+ 2 \alpha_t \beta_t \tilde{z}_t^\top \left( \mathcal{L} \otimes I_M \right) ( I_N \otimes \mathcal{K}^{-1} ) K_t \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right)
\]
\[
+ \alpha_t^2 \tilde{z}_t^\top \left( \overline{g}(y_t) - \overline{h}(\theta^*) \right)^\top K_t ( I_N \otimes \mathcal{K}^{-1} ) K_t \left( \overline{g}(y_t) - \overline{h}(\theta^*) \right)
\]
\[
+ \alpha_t^2 \mathbb{E}_{\theta^*} \left[ \left( \overline{g}(y_t) - \overline{h}(\theta^*) \right)^\top K_t ( I_N \otimes \mathcal{K}^{-1} ) K_t \left( \overline{g}(y_t) - \overline{h}(\theta^*) \right) \right]
\] (19)
for all \( t \geq 0 \).

Recall the definition of consensus subspace in Definition[11] and note that by using the properties of the Laplacian \( \mathcal{L} \) and matrix Kronecker products we have
\[
\tilde{z}_t^\top \left( \mathcal{L} \otimes \mathcal{K}^{-1} \right) \tilde{z}_t = ( \tilde{z}_t )^\top \left( \mathcal{L} \otimes \mathcal{K}^{-1} \right) ( \tilde{z}_t ) \geq \lambda_2(\mathcal{L}) \lambda_1(\mathcal{K}^{-1}) \| ( \tilde{z}_t ) \|^2
\] (20)
for all \( t \geq 0 \), where \( \lambda_1(\mathcal{K}^{-1}) > 0 \) denotes the smallest eigenvalue of the positive definite matrix \( \mathcal{K}^{-1} \).

Now consider the inequality
\[
\tilde{z}_t^\top \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right) = \sum_{n=1}^{N} ( z_n(t) - \theta^* )^\top ( h_n(z_n(t)) - h_n(\theta^*) ) \geq 0
\]
(where the non-negativity of the terms in the summation follows from Proposition[21]), and note that, by Assumption[33] and hypothesis[15], there exist positive constants \( c_1 \) and \( t_1 \) large enough such that
\[
\tilde{z}_t^\top \left( I_N \otimes \mathcal{K}^{-1} \right) K_t \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right)
\]
\[
\geq \tilde{z}_t^\top \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right) - \tilde{z}_t^\top \left( I_N \otimes \mathcal{K}^{-1} \right) \left( K_t - ( I_N \otimes \mathcal{K}^{-1} ) \right) \left( \overline{h}(z_t) - \overline{h}(\theta^*) \right)
\]
\[
\geq - \| \tilde{z}_t \| \| I_N \otimes \mathcal{K}^{-1} \| \| K_t - ( I_N \otimes \mathcal{K}^{-1} ) \| \| \overline{h}(z_t) - \overline{h}(\theta^*) \| \)
\[
\geq - c_1 \left( 1/(t+1)^{\tau_3} \right) \left( 1 + \| \tilde{z}_t \|^2 \right)
\]
for all \( t \geq t_1 \), where we also use the inequality \( \| \tilde{z}_t \| \leq \| \tilde{z}_t \|^2 + 1 \). Similarly, by invoking the boundedness of the matrices involved and the linear growth condition on the \( h_n(\cdot) \)-s and making \( c_1 \) and \( t_1 \) larger if necessary, we obtain the following sequence of inequalities for all \( t \geq t_1 \):

\[
\begin{align*}
\mathbb{E}_{\theta^*} \left[ (L \otimes I_M) \left( I_N \otimes K^{-1} \right) \left( L \otimes I_M \right) \tilde{z}_t \right] &= (\tilde{z}_t)_{c=1} \mathbb{E}_{\theta^*} \left[ (L \otimes I_M) \left( I_N \otimes K^{-1} \right) \left( L \otimes I_M \right) \right] (\tilde{z}_t)_{c=1} \leq c_1 \| (\tilde{z}_t)_{c=1} \|^2, \\
\tilde{z}_t^T (L \otimes I_M) \left( I_N \otimes K^{-1} \right) K_t \left( \tilde{h} (\tilde{z}_t) - \tilde{h} (\tilde{\theta}) \right) &\leq c_1 \left( 1 + \| \tilde{z}_t \|^2 \right), \\
(\tilde{h} (\tilde{z}_t) - \tilde{h} (\tilde{\theta}))^T K_t \left( I_N \otimes K^{-1} \right) K_t \left( \tilde{h} (\tilde{z}_t) - \tilde{h} (\tilde{\theta}) \right) &\leq c_1 \left( 1 + \| \tilde{z}_t \|^2 \right),
\end{align*}
\]

and

\[
\mathbb{E}_{\theta^*} \left[ (\tilde{g}(y_t) - \tilde{h}(\tilde{\theta}))^T K_t \left( I_N \otimes K^{-1} \right) K_t \left( \tilde{g}(y_t) - \tilde{h}(\tilde{\theta}) \right) \right] \leq c_1,
\]

where the last inequality uses the fact that \( \tilde{g}(y_t) \) possesses moments of all orders due to the exponential statistics.

Noting that there exist positive constants \( c_2 \) and \( c_3 \) such that

\[
c_2 \| \tilde{z}_t \|^2 \leq \tilde{z}_t^T \left( I_N \otimes K^{-1} \right) \tilde{z}_t = V_t \leq c_3 \| \tilde{z}_t \|^2
\]

for all \( t \), by (19)-(21) we have for all \( t \geq t_1 \)

\[
\begin{align*}
\mathbb{E}_{\theta^*} [V_{t+1} | \mathcal{F}_t] &\leq \left( 1 + c_4 \alpha_t \left( \frac{1}{(t+1)^\gamma_3} + \beta_t + \alpha_t \right) \right) V_t \\
&\quad - c_5 (\beta_t - \beta_t^2) \| (\tilde{z}_t)_{c=1} \|^2 + c_6 \left( \frac{\alpha_t}{(t+1)^\gamma_3} + \alpha_t \beta_t + \alpha_t^2 \right)
\end{align*}
\]

for some positive constants \( c_4 \), \( c_5 \) and \( c_6 \). Since \( \beta_t \to 0 \) as \( t \to \infty \) by (11), we may choose \( t_2 \) large enough (larger than \( t_1 \)) such that \( (\beta_t - \beta_t^2) \geq 0 \) for all \( t \geq t_2 \). Further, the hypotheses on the weight sequences (11) confirm the existence of constants \( \tau_4 \) and \( \tau_5 \) strictly greater than 1, and positive constants \( c_7 \) and \( c_8 \), such that

\[
c_4 \alpha_t \left( \frac{1}{(t+1)^\gamma_3} + \beta_t + \alpha_t \right) \leq \frac{c_7}{(t+1)^\gamma_4} = \gamma_t
\]

and

\[
c_6 \left( \frac{\alpha_t}{(t+1)^\gamma_3} + \alpha_t \beta_t + \alpha_t^2 \right) \leq \frac{c_8}{(t+1)^\gamma_5} = \gamma'_t
\]

for all \( t \geq t_2 \) (by making \( t_2 \) larger if necessary). By the above construction we then obtain

\[
\mathbb{E}_{\theta^*} [V_{t+1} | \mathcal{F}_t] \leq (1 + \gamma_t) V_t + \gamma'_t
\]

for all \( t \geq t_2 \) with the positive weight sequences \( \{ \gamma_t \} \) and \( \{ \gamma'_t \} \) being summable, i.e.,

\[
\sum_{t \geq 0} \gamma_t < \infty \quad \text{and} \quad \sum_{t \geq 0} \gamma'_t < \infty.
\]
Note that, by (25), the product $\prod_{s=t}^{\infty} (1 + \gamma_s)$ exists for all $t$, and define by $\{W_t\}$ the $\mathbb{R}_+^+$-valued $\{\mathcal{F}_t\}$-adapted process such that

$$W_t = \left(\prod_{s=t}^{\infty} (1 + \gamma_s)\right) V_t + \sum_{s=t}^{\infty} \gamma'_s, \quad \forall t. \quad (26)$$

By (24), the process $\{W_t\}$ may be shown to satisfy

$$\mathbb{E}_{\theta^*} [W_{t+1} | \mathcal{F}_t] \leq W_t$$

for all $t \geq t_2$. Being a non-negative supermartingale the process $\{W_t\}$ converges a.s. to a bounded random variable $W^*$ as $t \to \infty$. It then follows readily by (26) that $V_t \to V^*$ a.s. as $t \to \infty$. In particular, we conclude that the process $\{V_t\}$ is bounded a.s., which establishes the desired boundedness of the sequences $\{z_n(t)\}$ for all $n$.

The following useful convergence may be extracted as a corollary to Lemma 4.1.

**Corollary 4.1.** Under the hypotheses of Lemma 4.1 there exists a finite random variable $V^*$ such that $V_t \to V^*$ a.s. as $t \to \infty$, where $V_t = \hat{z}_t^T \left(I_N \otimes K^{-1}\right) \hat{z}_t$ as in (18).

The following Lyapunov-type construction, whose proof is relegated to Appendix A, will be critical to the subsequent development.

**Proposition 4.1.** Let $\varepsilon \in (0, 1)$ and $\Gamma_\varepsilon$ denote the set

$$\Gamma_\varepsilon = \left\{z \in \mathbb{R}^{NM} : \varepsilon \leq \left\| z - \overline{\theta} \right\| \leq 1/\varepsilon \right\}. \quad (27)$$

For each $t \geq 0$, denote by $\mathcal{H}_t : \mathbb{R}^{NM} \to \mathbb{R}$ the function given by

$$\mathcal{H}_t(z) = \frac{b_\beta}{a_t} \left( z - \overline{\theta} \right)^T \left( \bar{T} \otimes K^{-1}\right) \left( z - \overline{\theta} \right) + \left( z - \overline{\theta} \right)^T \left( \bar{h}(z) - \bar{h}(\overline{\theta}) \right)$$

for all $z \in \mathbb{R}^{NM}$, where the matrix $K^{-1} \in S^{M}_{++}$ and $b_\beta > 0$ is a constant. Then, there exist $t_\varepsilon > 0$ and a constant $\overline{b}_\varepsilon > 0$ such that for all $t \geq t_\varepsilon$

$$\mathcal{H}_t(z) \geq \overline{b}_\varepsilon \left\| z - \overline{\theta} \right\|^2, \quad \forall z \in \Gamma_\varepsilon.$$

We now complete the proof of Theorem 4.1.

**Proof of Theorem 4.1.** In what follows we use the notation and definitions formulated in the proof of Lemma 4.1. Let us consider $\varepsilon \in (0, 1)$ and let $\rho_\varepsilon$ denote the $\{\mathcal{F}_t\}$ stopping time

$$\rho_\varepsilon = \inf \left\{ t \geq 0 : z_t \notin \Gamma_\varepsilon \right\},$$

where $\Gamma_\varepsilon$ is defined in (27). Let $\{V_t\}$ be the $\{\mathcal{F}_t\}$-adapted process defined in (18) and denote by $\{V_t^\varepsilon\}$ the stopped process

$$V_t^\varepsilon = V_{t \wedge \rho_\varepsilon}, \quad \forall t,$$
which is readily seen to be \( \{ F_t \} \) adapted. Noting that
\[
V_{t+1}^\varepsilon = V_{t+1}^\varepsilon (\rho_\varepsilon > t) + V_{t+1}^\varepsilon (\rho_\varepsilon \leq t)
\]
and the fact that the indicator function \( \mathbb{I}(\rho_\varepsilon > t) \) and the random variable \( V_{t+1}^\varepsilon (\rho_\varepsilon \leq t) \) are adapted to \( F_t \) for all \( t \) (\( \rho_\varepsilon \) being an \( \{ F_t \} \) stopping time), we have
\[
\mathbb{E}_{\theta^*} [V_{t+1}^\varepsilon \mid F_t] = \mathbb{E}_{\theta^*} [V_{t+1}^\varepsilon \mid F_t] \mathbb{I}(\rho_\varepsilon > t) + V_{t+1}^\varepsilon \mathbb{I}(\rho_\varepsilon \leq t)
\]
(29)
for all \( t \).

Recall the function \( \mathcal{H}_t(\cdot) \) defined in (28), setting \( b_\beta = 1/2 \) in the definition of \( \mathcal{H}_t(\cdot) \) we obtain
\[
2\beta t \tilde{z}_t^\top (T \otimes K^{-1}) \tilde{z}_t + 2\alpha t \tilde{z}_t^\top (I_N \otimes K^{-1}) K_t \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)
= 2\alpha t \mathcal{H}_t(z_t) + \beta t \tilde{z}_t^\top (T \otimes K^{-1}) \tilde{z}_t
+ 2\alpha t \tilde{z}_t^\top (I_N \otimes K^{-1}) (K_t - (I_N \otimes K^{-1})) \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right).
\]
A slight rearrangement of the terms in the expansion (19) then yields
\[
\mathbb{E}_{\theta^*} [V_{t+1} \mid F_t] = V_t - 2\alpha t \mathcal{H}_t(z_t) - \beta t \tilde{z}_t^\top (T \otimes K^{-1}) \tilde{z}_t
- 2\alpha t \tilde{z}_t^\top (I_N \otimes K^{-1}) (K_t - (I_N \otimes K^{-1})) \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)
+ \beta t \tilde{z}_t^\top \mathbb{E}_{\theta^*} [(T \otimes I_M) (I_N \otimes K^{-1}) (T \otimes I_M)] \tilde{z}_t
+ 2\alpha t \beta t \tilde{z}_t^\top (T \otimes I_M) (I_N \otimes K^{-1}) K_t \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)
+ \alpha t \tilde{z}_t^\top \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)^\top K_t (I_N \otimes K^{-1}) K_t \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)
+ \alpha t \tilde{z}_t^\top \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)^\top K_t (I_N \otimes K^{-1}) K_t \left( \tilde{h}(z_t) - \tilde{h}(\theta^*) \right)
\]
for all \( t \geq 0 \), where \( \mathcal{H}_t(\cdot) \) is defined in (28). The inequalities in (20)-(22) then show that there exist positive constants \( b_1, b_2 \) and \( b_3 \), and a deterministic time \( t_1 \) (large enough), such that,
\[
\mathbb{E}_{\theta^*} [V_{t+1} \mid F_t] \leq (1 + b_1 (\alpha t (t+1)^{-\gamma_3} + \alpha_2^2 + \alpha t \beta_3)) V_t - 2\alpha t \mathcal{H}_t(z_t)
- b_2 (\beta t - \beta_2^2) \| \tilde{z}_t \|_2^2 + b_3 (\alpha t (t+1)^{-\gamma_3} + \alpha_2^2 + \alpha t \beta_3)
\]
(31)
for all \( t \geq t_1 \). Note that, by definition, on the event \( \{ \rho_\varepsilon > t \} \) we have \( z_t \in \Gamma_\varepsilon \), and hence, an immediate application of Proposition 4.1 establishes the existence of a positive constant \( \overline{b}_t \) and a large enough deterministic time \( t_\varepsilon > 0 \), such that,
\[
\mathcal{H}_t(z_t) \mathbb{I}(\rho_\varepsilon > t) \geq \overline{b}_t \| \tilde{z}_t \|_2^2 \mathbb{I}(\rho_\varepsilon > t)
\]
for all $t \geq t_\varepsilon$. By (22) and (30)-(31) and making $t_\varepsilon$ larger if necessary, it then follows that there exist a constant $b_4(\varepsilon) > 0$ such that

$$
\mathbb{E}_{\theta^*} [V_{t+1} | \mathcal{F}_t] \mathbb{I} (\rho_\varepsilon > t) \leq \left[ (1 - b_4(\varepsilon) \alpha_t + b_1 (\alpha_t (t+1)^{-\tau_3} + \alpha_t^2 + \alpha_t \beta_t) ) V_t \right.
-b_2 \left( \beta_t - \beta_t^2 \right) \| \hat{z}_t \| + b_3 (\alpha_t (t+1)^{-\tau_3} + \alpha_t^2 + \alpha_t \beta_t) \right] \mathbb{I} (\rho_\varepsilon > t)
$$

for all $t \geq t_\varepsilon$. Since $\alpha_t \to 0$ and $\beta_t \to 0$ as $t \to \infty$, by choosing $t_\varepsilon$ large enough we may assert

$$
\beta_t - \beta_t^2 \geq 0, \quad \forall t \geq t_\varepsilon,
$$

and the existence of positive constants $b_5$ and $\tau_4$ such that

$$
b_3 (\alpha_t (t+1)^{-\tau_3} + \alpha_t^2 + \alpha_t \beta_t) \leq b_5 \alpha_t (t+1)^{-\tau_4}, \quad \forall t \geq t_\varepsilon.
$$

We thus obtain for $t \geq t_\varepsilon$

$$
\mathbb{E}_{\theta^*} [V_{t+1} | \mathcal{F}_t] \mathbb{I} (\rho_\varepsilon > t) \leq \left[ (1 - (b_4(\varepsilon)/2) \alpha_t ) V_t + b_5 \alpha_t (t+1)^{-\tau_4} \right] \mathbb{I} (\rho_\varepsilon > t).
$$

(32)

Note that, by definition of $\Gamma_\varepsilon$,

$$
\| \hat{z}_t \|^2 \geq \varepsilon^2 \text{ on } \{ \hat{z}_t \in \Gamma_\varepsilon \},
$$

and, hence, by (22) we conclude that there exists a constant $b_6(\varepsilon) > 0$ such that

$$
V_t \geq b_6(\varepsilon) \text{ on } \{ \rho_\varepsilon > t \}.
$$

By (32) we then have for all $t \geq t_\varepsilon$

$$
\mathbb{E}_{\theta^*} [V_{t+1} | \mathcal{F}_t] \mathbb{I} (\rho_\varepsilon > t) \leq \left[ V_t - b_7(\varepsilon) \alpha_t + b_5 \alpha_t (t+1)^{-\tau_4} \right] \mathbb{I} (\rho_\varepsilon > t)
$$

with $b_7(\varepsilon)$ being another positive constant. Finally, the observation that $(b_7(\varepsilon)/2) \alpha_t \geq b_7 \alpha_t (t+1)^{-\tau_4}$ eventually leads to

$$
\mathbb{E}_{\theta^*} [V_{t+1} | \mathcal{F}_t] \mathbb{I} (\rho_\varepsilon > t) \leq \left[ V_t - (s_7(\varepsilon)/2) \alpha_t \right] \mathbb{I} (\rho_\varepsilon > t)
= V_t \mathbb{I} (\rho_\varepsilon > t) - b_8(\varepsilon) \alpha_t \mathbb{I} (\rho_\varepsilon > t)
$$

for all $t \geq t_\varepsilon$ (making $t_\varepsilon$ larger if necessary), where $b_8(\varepsilon) = b_7(\varepsilon)/2$.

By (29) we then obtain

$$
\mathbb{E}_{\theta^*} \left[ V^\varepsilon_{t+1} | \mathcal{F}_t \right] \leq V_t \mathbb{I} (\rho_\varepsilon > t) + V_t \mathbb{I} (\rho_\varepsilon \leq t) - b_8(\varepsilon) \alpha_t \mathbb{I} (\rho_\varepsilon > t)
= V^\varepsilon_t - b_8(\varepsilon) \alpha_t \mathbb{I} (\rho_\varepsilon > t)
$$

(33)
for all $t \geq t_\varepsilon$. Note that the $\{F_t\}$-adapted process $\{V_t^\varepsilon\}_{t \geq t_\varepsilon}$ satisfies $E_{\theta^*}[V_{t+1}^\varepsilon | F_t] \leq V_t^\varepsilon$ for all $t \geq t_\varepsilon$; hence, being a (non-negative) supermartingale it converges, i.e., there exists a finite random variable $V_\varepsilon^*$ such that $V_t^\varepsilon \rightarrow V_\varepsilon^*$ a.s. as $t \rightarrow \infty$. Now consider the $\{F_t\}$-adapted $\mathbb{R}_+$-valued process $\{W_t^\varepsilon\}$ given by

$$W_t^\varepsilon = V_t^\varepsilon + b_8(\varepsilon) \sum_{s=0}^{t-1} \alpha_s 1(\rho_s > s),$$

and note that, by (33) we obtain

$$E_{\theta^*}[W_{t+1}^\varepsilon | F_t] \leq V_t^\varepsilon - b_8(\varepsilon) \alpha_t 1(\rho_t > t) + b_8(\varepsilon) \sum_{s=0}^{t} \alpha_s 1(\rho_s > s) = W_t^\varepsilon$$

for all $t \geq t_\varepsilon$; hence $\{W_t^\varepsilon\}_{t \geq t_\varepsilon}$ is a non-negative supermartingale and there exists a finite random variable $W_\varepsilon^*$ such that $W_t^\varepsilon \rightarrow W_\varepsilon^*$ a.s. as $t \rightarrow \infty$. We then conclude by (34) that the following limit exists:

$$\lim_{t \rightarrow \infty} b_8(\varepsilon) \sum_{s=0}^{t-1} \alpha_s 1(\rho_s > s) = W_\varepsilon^* - V_\varepsilon^* < \infty \text{ a.s.}$$

(35)

Given that $\sum_{s=0}^{t-1} \alpha_s \rightarrow \infty$ as $t \rightarrow \infty$, the limit condition in (35) is fulfilled only if the summation terminates at a finite time a.s., i.e., we must have $\rho_\varepsilon < \infty$ a.s.

To summarize, we have for each $\varepsilon \in (0, 1)$, $\rho_\varepsilon < \infty$ a.s., i.e., the process $\{\hat{Z}_t\}$ exits the set $\Gamma_\varepsilon$ in finite time a.s. In particular, for each positive integer $r > 1$, let $\rho_{1/r}$ be the stopping time obtained by choosing $\varepsilon = 1/r$ and consider the sequence $\{\hat{Z}_{\rho_{1/r}}\}$ (which is well defined due to the a.s. finiteness of each $\rho_{1/r}$) and note that, by definition,

$$\|\hat{Z}_{\rho_{1/r}}\| \in [0, 1/r) \cup (r, \infty) \text{ a.s.}$$

(36)

However, the a.s. boundedness of the sequence $\{\hat{Z}_t\}$ (see Lemma 4.1) implies that

$$\mathbb{P}_{\theta^*}(\|\hat{Z}_{\rho_{1/r}}\| > r \text{ i.o.}) = 0,$$

where i.o. stands for infinitely often as $r \rightarrow \infty$. Hence, by (36) we conclude that there exists a finite random integer valued random variable $r^*$ such that $\|\hat{Z}_{\rho_{1/r}}\| < 1/r$ for all $r \geq r^*$. This, in turn implies that $\|\hat{Z}_{\rho_{1/r}}\| \rightarrow 0$ as $r \rightarrow \infty$ a.s., and, in particular, we obtain

$$\mathbb{P}_{\theta^*}(\liminf_{t \rightarrow \infty}\|\hat{Z}_t\| = 0) = 1.$$

By (22) we may also conclude that $\liminf_{t \rightarrow \infty} V_t = 0$ a.s. Noting that the limit of $\{V_t\}$ exists a.s. (see Corollary 4.1) we further obtain $V_t \rightarrow 0$ as $t \rightarrow \infty$ a.s., from which, by another application of (22), we conclude that $\hat{Z}_t \rightarrow 0$ as $t \rightarrow \infty$ a.s. and the desired consistency assertion follows.

The other major result of this section concerns the pathwise convergence rate of the processes $\{z_n(t)\}$ to $\theta^*$, stated as follows:
Theorem 4.2. Let the processes $\{z_n(t)\}$ be defined as in (44) and the assumptions and hypotheses of Theorem 4.1 hold. Then, there exists a constant $\mu > 0$ such that for all $n$ we have

$$
P_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^{\mu} \|z_n(t) - \theta^*\| = 0 \right) = 1. $$

In order to obtain Theorem 4.2, we will first quantify the rate of agreement among the individual agent estimates. Specifically, we have the following (see Appendix A for a proof):

Lemma 4.2. Let the hypotheses of Lemma 4.1 hold. Then, for each pair of agents $n$ and $l$, we have

$$
P_{\theta^*} \left( \lim_{t \to \infty} (t + 1)\tau \|z_n(t) - z_l(t)\| = 0 \right) = 1,$$

for all $\tau \in (0, 1 - \tau_2)$.

We now complete the proof of Theorem 4.2.

Proof of Theorem 4.2. In what follows we reuse the notation and intermediate processes constructed in the proofs of Lemma 4.1 and Theorem 4.1. Recall $\{V_t\}$ to be the $\{F_t\}$-adapted process defined in (18). By (31) (and the development preceding it) we note that there exist positive constants $b_1$, $b_2$, and $b_3$, and a deterministic time $t_1$ (large enough), such that,

$$
E_{\theta^*} [V_{t+1} | F_t] \leq (1 + b_1 (\alpha_t (t + 1)^{-\tau_3} + \alpha_t^2 + \alpha_t \beta_t)) V_t - 2\alpha_t \mathcal{H}_t(z_t) - b_2 (\beta_t - \beta_t^2) \|z_t\|_t^2 + b_3 (\alpha_t (t + 1)^{-\tau_3} + \alpha_t^2 + \alpha_t \beta_t)
$$

for all $t \geq t_1$, where the function $\mathcal{H}_t(\cdot)$ is defined in (28). By (34) we obtain

$$
\mathcal{H}_t(z) \geq \left(z - \hat{\theta}^*\right)^T \left( \mathcal{L}(z) - \mathcal{L}(\hat{\theta}^*) \right) + \left(z - \hat{\theta}^*\right)^T \left( \bar{h}(z) - \bar{h}(\hat{\theta}^*) \right)
$$

and

$$
= \left(z - \theta^*\right)^T \left(h(z) - h(\theta^*)\right) + \left(z - \hat{\theta}^*\right)^T \left( \bar{h}(z) - \bar{h}(\hat{\theta}^*) \right) + \left(z - \hat{\theta}^*\right)^T \left( \bar{h}(z) - \bar{h}(\hat{\theta}^*) \right)
$$

for all $z \in \mathbb{R}^{NM}$.

Note that, by Proposition 2.1, $h(\cdot)$ is continuously differentiable with positive definite gradient $\nabla h(\theta^*) = I(\theta^*)$ at $\theta^*$; hence, by the mean-value theorem, there exists $\varepsilon_0 > 0$ such that for all $\theta \in B_{\varepsilon_0}(\theta^*)$ we have

$$
h(\theta) - h(\theta^*) = (I(\theta^*) + R(\theta, \theta^*)) (\theta - \theta^*),
$$

where $R(\cdot, \theta^*)$ is a measurable $\mathbb{R}^{M \times M}$-valued function of $\theta$ such that

$$
\|R(\theta, \theta^*)\| \leq \frac{\lambda_1 (I(\theta^*))}{2} \quad \forall \theta \in B_{\varepsilon_0}(\theta^*),
$$
with \( \lambda_1(I(\theta^*)) > 0 \) denoting the smallest eigenvalue of \( I(\theta^*) \). Also, observing that the function \( \overline{h}(\cdot) \) is locally Lipschitz, we may conclude that there exists a constant \( \ell_{\varepsilon_0} \) such that

\[
\left\| \overline{h}(\mathbf{z}) - \overline{h}(\mathbf{z}') \right\| \leq \ell_{\varepsilon_0} \left\| \mathbf{z} - \mathbf{z}' \right\| \quad \forall \mathbf{z}, \mathbf{z} \in \mathbb{B}_{\varepsilon_0}(\overline{\theta}).
\]

(41)

Now note that, by Theorem 4.1, \( \mathbf{z}_t \to 0 \) a.s. as \( t \to \infty \), and, by Lemma 4.2, there exists a constant \( \tau > 0 \) such that

\[
\mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^\tau \left\| (\mathbf{z}_t)_{C^+} \right\| = 0 \right) = 1.
\]

Now consider \( \delta > 0 \) (arbitrarily small) and note that by Egorov’s theorem there exists a (deterministic) time \( t_{\delta} > 0 \) (chosen to be larger than \( t_1 \) in (37)), such that \( \mathbb{P}_{\theta^*} (A_{\delta}) \geq 1 - \delta \), where \( A_{\delta} \) denotes the event

\[
\mathcal{A}_\delta = \left\{ \sup_{t \geq t_{\delta}} \left\| \mathbf{z}_t - \overline{\theta} \right\| \leq \varepsilon_0 \right\} \bigcup \left\{ \sup_{t \geq t_{\delta}} (t + 1)^\tau \left\| (\mathbf{z}_t)_{C^+} \right\| \leq \varepsilon_0 \right\}.
\]

Consequently, denoting by \( \rho_{\delta} \) the \( \{ \mathcal{F}_t \} \) stopping time

\[
\rho_{\delta} = \inf \left\{ t \geq t_{\delta} : \left\| \mathbf{z}_t - \overline{\theta} \right\| > \varepsilon_0 \text{ or } (t + 1)^\tau \left\| (\mathbf{z}_t)_{C^+} \right\| > \varepsilon_0 \right\},
\]

we have that

\[
\mathbb{P}_{\theta^*} (\rho_{\delta} = \infty) \geq 1 - \delta.
\]

(42)

Now consider \( t \in [t_{\delta}, \rho_{\delta}] \); noting that

\[
\left\| \mathbf{z}_t^a - \theta^* \right\| \leq \left\| \mathbf{z}_t - \overline{\theta} \right\| \leq \varepsilon_0,
\]

we have by the construction in (39)-(40)

\[
\left( (\mathbf{z}_t)_{C} - \overline{\theta} \right)^\top \left( \overline{h}(\mathbf{z}_t) - \overline{h}(\overline{\theta}) \right) = \left( \mathbf{z}_t^a - \theta^* \right)^\top (h(\mathbf{z}_t^a - h(\theta^*))
\]

\[
\geq \left( \left\| I(\theta^*) \right\| - \left\| R(\mathbf{z}_t^a, \theta^*) \right\| \right) \left\| \mathbf{z}_t^a - \theta^* \right\|^2 \geq (1/2)\lambda_1(I(\theta^*)) \left\| \mathbf{z}_t^a - \theta^* \right\|^2 \geq b_4 V_t
\]

for some constant \( b_4 > 0 \).

Similarly, using (41), we have the following inequalities for \( t \in [t_{\delta}, \rho_{\delta}] \):

\[
\left( (\mathbf{z}_t)_{C} - \overline{\theta} \right)^\top \left( \overline{h}(\mathbf{z}_t) - \overline{h}(\overline{\theta}) \right) \leq \left\| (\mathbf{z}_t)_{C} \right\| \left\| \overline{h}(\mathbf{z}_t) - \overline{h}(\overline{\theta}) \right\|
\]

\[
\leq \varepsilon_0 (t + 1)^{-\tau} \ell_{\varepsilon_0} \left\| \mathbf{z}_t - \overline{\theta} \right\| \leq \varepsilon_0^2 (t + 1)^{-\tau} \ell_{\varepsilon_0},
\]

and

\[
\left( (\mathbf{z}_t)_{C} - \overline{\theta} \right)^\top \left( \overline{h}(\mathbf{z}_t) - \overline{h}(\mathbf{z}_t)_{C} \right) \leq \left\| (\mathbf{z}_t)_{C} - \overline{\theta} \right\| \ell_{\varepsilon_0} \left\| \mathbf{z}_t - (\mathbf{z}_t)_{C} \right\|
\]

\[
\leq \varepsilon_0 \ell_{\varepsilon_0} \left\| (\mathbf{z}_t)_{C} \right\| \leq \varepsilon_0^2 \ell_{\varepsilon_0} (t + 1)^{-\tau}.
\]

(44)
Hence, from (38) and (43)-(44), we conclude that for $t \in [\delta, \rho_\delta)$ we have

$$\mathcal{H}_t(z_t) \geq b_4 V_t - 2\varepsilon_0^2 \epsilon z_0 (t + 1)^{-\tau}.$$ \hfill (45)

Let $\{V_t^\delta\}$ be the $\mathbb{R}_+$-valued $\{\mathcal{F}_t\}$-adapted process such that $V_t^\delta = V_t \mathbb{I}(t < \rho_\delta)$ for all $t$. Noting that

$$V_{t+1}^\delta = V_{t+1}\mathbb{I}(t+1 < \rho_\delta) \leq V_{t+1}\mathbb{I}(t < \rho_\delta),$$

we have

$$\mathbb{E}_{\theta^*}[V_{t+1}^\delta | \mathcal{F}_t] = \mathbb{I}(t < \rho_\delta) \mathbb{E}_{\theta^*}[V_{t+1}^\delta | \mathcal{F}_t] \quad \forall t.$$ \hfill (46)

For $t \geq t_\delta$ we have by (45)

$$\mathcal{H}_t(z_t)\mathbb{I}(t < \rho_\delta) \geq (b_4 V_t - 2\varepsilon_0^2 \epsilon z_0 (t + 1)^{-\tau}) \mathbb{I}(t < \rho_\delta),$$

hence, it follows from (37) and (46) that

$$\mathbb{E}_{\theta^*}[V_{t+1}^\delta | \mathcal{F}_t] \leq (1 + b_1 (\alpha(t+1)^{-\tau_3} + \alpha_2^2 + \alpha_t \beta_t)) V_t \mathbb{I}(t < \rho_\delta)$$

$$-2\alpha_t (b_4 V_t - 2\varepsilon_0^2 \epsilon z_0 (t + 1)^{-\tau}) \mathbb{I}(t < \rho_\delta)$$

$$-b_2 (\beta_t - \beta_t^2) \|\langle \nu_t \rangle_{\mathcal{C}_+}\|^2 + b_3 (\alpha_t (t + 1)^{-\tau_3} + \alpha_2^2 + \alpha_t \beta_t)$$

$$\leq (1 - \alpha_t (2b_1 - b_1(t+1)^{-\tau_3} - b_1 \alpha_t - b_1 \beta_t)) V_t^\delta - b_2 (\beta_t - \beta_t^2) \|\langle \nu_t \rangle_{\mathcal{C}_+}\|^2$$

$$+ \alpha_t (b_2 (t + 1)^{-\tau_3} + b_3 \alpha_t + b_3 \beta_t + 4\varepsilon_0^2 \epsilon z_0 (t + 1)^{-\tau})$$

for all $t \geq t_\delta$. Observing the decay rates of the various coefficients (see [11]), we conclude that there exist a deterministic time $t'_\delta \geq t_\delta$, and positive constants (independent of $\delta$) $b_5$, $b_6$ and $\tau_4$ such that

$$\mathbb{E}_{\theta^*}[V_{t+1}^\delta | \mathcal{F}_t] \leq (1 - b_5 \alpha_t) V_t^\delta + b_6 \alpha_t (t + 1)^{-\tau_4}$$ \hfill (47)

and $b_5 \alpha_t < 1$, for all $t \geq t'_\delta$.

Let us now choose a constant $\overline{\tau}$ (independently of $\delta$) such that $\overline{\tau} \in (0, b_5 \wedge \tau_4 \wedge 1)$. Then, using the inequality

$$(t + 1)^{\overline{\tau}} - t^{\overline{\tau}} \leq \overline{\tau} \overline{\tau}^{-1}$$

we have for all $t > t_\delta$

$$(t + 1)^{\overline{\tau}} (1 - b_5 \alpha_{t_{\delta - 1}}) \leq t^{\overline{\tau}} (1 + \overline{\tau} t^{-1}) (1 - b_5 t^{-1}) \leq t^{\overline{\tau}} (1 - (b_5 - \overline{\tau}) t^{-1})$$

and

$$(t + 1)^{\overline{\tau}} t^{-1 - \tau_4} = (1 + t^{-1})^{\overline{\tau}} t^{-1 - (\tau_4 - \overline{\tau})} \leq (1 + (t'_\delta)^{-1})^{\overline{\tau}} t^{-1 - (\tau_4 - \overline{\tau})}.$$
Thus, from (47) we obtain
\[
\mathbb{E}_{\theta^*} \left[ (t+1)^{\sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} V_t^\delta} \mid \mathcal{F}_{t-1} \right] \leq \left( 1 - \left( b_0 - \mathfrak{m} \right) t^{-1} \right) t^{\sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} V_{t-1}^\delta} + b_0 \left( 1 + (t_{\delta}')^{-1} \right) t^{\sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} V_{t-1}^\delta}
\]
for all \( t > t_{\delta}' \) for some constant \( b_0^\delta > 0 \). Since \( \tau_k > \mathfrak{m} \), we have \( \sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} < \infty \); denoting by \( \{ W_t^\delta \} \) the non-negative \( \mathcal{F}_t \)-adapted process such that
\[
W_t^\delta = (t+1)^{\sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} V_t^\delta} + b_0 \left( 1 + (t_{\delta}')^{-1} \right) t^{\sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} V_{t-1}^\delta},
\]
we have that \( \mathbb{E}_{\theta^*} [W_t^\delta \mid \mathcal{F}_{t-1}] \leq W_t^\delta \) for all \( t > t_{\delta}' \). Hence, the process \( \{ W_t^\delta \}_{t \geq t_{\delta}'} \) is a non-negative supermartingale and converges a.s. to a finite non-negative random variable \( W_\infty^\delta \). By (48) we further conclude that \( (t+1)^{\sum_{s=1}^{\infty} s^{-1-(\tau_s-\mu)} V_t^\delta} \rightarrow W_\infty^\delta \) a.s. as \( t \rightarrow \infty \). Now let \( \mu \in (0, \mathfrak{m}) \) be another constant (chosen independently of \( \delta \)); noting that the limit \( W_\infty^\delta \) is finite, the above convergence leads to
\[
\mathbb{P}_{\theta^*} \left( \lim_{t \rightarrow \infty} (t+1)^{\mu} V_t^\delta = 0 \right) = 1.
\]
By (49) and the fact that \( V_t^\delta = \mathbb{V}_t \mathbb{I}(t < \rho_\delta) \) for all \( t \), we conclude that,
\[
\lim_{t \rightarrow \infty} (t+1)^{\mu} V_t = 0 \text{ a.s. on } \{ \rho_\delta = \infty \}.
\]
Hence, by (42) we obtain
\[
\mathbb{P}_{\theta^*} \left( \lim_{t \rightarrow \infty} (t+1)^{\mu} V_t = 0 \right) \geq 1 - \delta.
\]
Since \( \delta > 0 \) was chosen arbitrarily and \( \mu > 0 \) is independent of \( \delta \), we have, in fact, \( (t+1)^{\mu} V_t \rightarrow 0 \) a.s. as \( t \rightarrow \infty \) by taking \( \delta \) to zero. The desired assertion follows immediately by noting the correspondence between the processes \( \{ \mathbb{z}_t \} \) and \( \{ V_t \} \) (see (23)).

The assertions of Theorem 4.2 may readily be extended to the case of non-uniform (over sample paths) convergence of the matrix gain sequences \( \{ K_n(t) \} \) to their designated limit \( K \) as follows (see Appendix A for a proof):

**Corollary 4.2.** Let the sequences \( \{ z_n(t) \} \) be defined as in (14). Let Assumptions 2.2, 3.2, and 3.2 hold as in the hypotheses of Theorem 4.2 and the matrix gain sequences \( \{ K_n(t) \} \) be such that \( (t+1)^{\tau_3} \| K_n(t) - K \| \rightarrow 0 \) a.s. as \( t \rightarrow \infty \) for all \( n \). Then, the assertions of Theorem 4.2 continue to hold, i.e., there exists \( \mu > 0 \) such that \( (t+1)^{\mu} \| z_n(t) - \theta^* \| \rightarrow 0 \) a.s. as \( t \rightarrow \infty \) for all \( n \).

Note that Corollary 4.2 is in fact a restatement of Theorem 4.2 under the relaxed assumption that the convergence of the matrix gain sequences \( \{ K_n(t) \} \) need not be uniform over sample paths.

5. **Proofs of Main Results**

Throughout this section, Assumption 2.2 and Assumptions 3.1, 3.3 are assumed to hold.
A. Convergence of Auxiliary Estimates and Adaptive Gains

The first result concerns the consistency of the auxiliary estimate sequence \( \{ \hat{x}_n(t) \} \) at each agent. To this end, noting that the evolution of the auxiliary estimates, see (7), corresponds to a specific instantiation of the generic estimator dynamics analyzed in Theorem 4.2 (with \( K_n(t) = I_M \) for all \( n \) and \( t \)), we immediately have the following:

**Lemma 5.1.** For each \( n \), the auxiliary estimate sequence \( \{ \hat{x}_n(t) \} \) (see Section 3-A) is a strongly consistent estimate of \( \theta^* \). In particular, there exists a positive constant \( \mu_0 \) such that
\[
(t + 1)^{\mu_0} \| \hat{x}_n(t) - \theta^* \| \to 0 \quad \text{as} \quad t \to \infty \quad \text{a.s. for all} \quad n.
\]

Lemma 5.1 and local Lipschitz continuity of the functions \( h_n(\cdot) \) lead to the following characterization of the adaptive gain sequences \( \{ K_n(t) \} \) (9) driving the local innovation terms of the agent estimates \( \{ x_n(t) \} \) (8) (see Appendix B for a proof):

**Lemma 5.2.** There exists a positive constant \( \tau' \) such that, for each \( n \), the adaptive gain sequence \( \{ K_n(t) \} \), see (9), converges a.s. to \( N.I^{-1}(\theta^*) \) at rate \( \tau' \), i.e.,
\[
P \left( (t + 1)^{\tau'} \| K_n(t) - N.I^{-1}(\theta^*) \| = 0 \right) = 1
\]
where \( I(\cdot) \) denotes the centralized Fisher information, see (6).

As an immediate consequence of the above development, we have the following consistency of the distributed agent estimates \( \{ x_n(t) \} \):

**Corollary 5.1.** For each \( n \), the estimate sequence \( \{ x_n(t) \} \) (see Section 3-A) is a strongly consistent estimate of \( \theta^* \), i.e., \( x_n(t) \to \theta^* \) as \( t \to \infty \) a.s.

**Proof:** Note that, by Lemma 5.2, there exists \( \tau' > 0 \) such that \( (t + 1)^{\tau'} \| K_n(t) - N.I^{-1}(\theta^*) \| \to 0 \) as \( t \to \infty \) a.s. Thus, the sequences \( \{ x_n(t) \} \) fall under the purview of Theorem 4.2 (with \( K = N.I^{-1}(\theta^*) \)) and the assertion follows.

B. Proofs of Theorem 3.1 and Theorem 3.2

The key idea in proving Theorem 3.1 and Theorem 3.2 consists of comparing the nonlinear estimate recursions, see (8), to a suitably linearized recursion. To this end, we consider the following result on distributed linear stochastic recursions developed in [21] in the context of asymptotically efficient distributed parameter estimation in linear multi-agent models. The result to be stated below is somewhat less general than the development in [21], but serves the current scenario.

**Theorem 5.1** (Theorem 3.2 and Theorem 3.3 in [21]). For each \( n \), let \( \{ v_n(t) \} \) be an \( \mathbb{R}^M \)-valued \( \{ F_t \} \)-adapted
Then the following hold for the processes evolving in a distributed fashion as follows:

\[ v_n(t+1) = v_n(t) - \beta_t \sum_{t \in \Omega_n(t)} (v_n(t) - v(t)) + \alpha_t D_n(t) (B_n (\theta^* - v_n(t)) + \zeta_n(t)), \]

where \( B_n \), for each \( n \), is an \( M_n \times M \) matrix (for some positive integer \( M_n \)) such that

1. the matrix \( A = \sum_{n=1}^{N} B_n^T B_n \) is positive definite;
2. for each \( n \), the \( M \times M_n \) matrix valued process \( \{D_n(t)\} \) is \( \mathcal{F}_t \)-adapted with \( D_n(t) \to N(A^{-1}B_n^T \) as \( t \to \infty \) a.s.;
3. for each \( n \), the \( \{F_{t+1}\} \)-adapted sequence is such that \( \{\zeta_n(t)\} \) is independent of \( \mathcal{F}_t \) for all \( t \), the sequence \( \{\zeta_n(t)\} \) is i.i.d. with zero mean and covariance \( I_{M_n} \), and there exists a constant \( \varepsilon > 0 \) such that \( E[\|\zeta_n(t)\|^{2+\varepsilon}] < \infty \);
4. the Laplacian sequence \( \{L_t\} \) representing the random communication neighborhoods \( \Omega_n(t), n = 1, \cdots, N \), satisfies Assumption 5.2, and
5. the weight sequences \( \{\alpha_t\} \) and \( \{\beta_t\} \) satisfy Assumption 5.2

Then the following hold for the processes \( \{v_n(t)\}, n = 1, \cdots, N \):

1. for each \( n \) and \( \tau \in [0, 1/2] \), we have
   \[ \mathbb{P} \left( \lim_{t \to \infty} (t+1)^{\tau} \|v_n(t) - \theta^*\| = 0 \right) = 1; \]
2. for each \( n \), the sequence \( \{v_n(t)\} \), viewed as an estimate of \( \theta^* \), is asymptotically normal with asymptotic covariance \( A^{-1} \), i.e.,
   \[ \sqrt{t+1} (v_n(t) - \theta^*) \Rightarrow \mathcal{N} (0, A^{-1}) . \]

The following corollary to Theorem 5.1 will be used in the sequel.

**Corollary 5.2.** For each \( n \), let \( \tilde{v}_n(t) \) be the \( \mathcal{F}_t \)-adapted \( \mathbb{R}^M \)-valued process evolving in a distributed fashion as

\[ \tilde{v}_n(t+1) = \tilde{v}_n(t) - \beta_t \sum_{t \in \Omega_n(t)} (\tilde{v}_n(t) - \tilde{v}(t)) + \alpha_t K_n(t) (I_n(\theta^*) (\theta^* - \tilde{v}_n(t)) + w_n(t)), \]

where

1. for each \( n \), \( \{w_n(t)\} \) is the \( \mathcal{F}_{t+1} \)-adapted sequence given by \( w_n(t) = g_n(y_n(t)) - h_n(\theta^*) \) for all \( t \);
2. for each \( n \), \( \{K_n(t)\} \) denotes the \( \mathcal{F}_t \)-adapted innovation gain sequence defined as in (9);
3. the Laplacian sequence \( \{L_t\} \) representing the random communication neighborhoods \( \Omega_n(t), n = 1, \cdots, N \), satisfies Assumption 5.2, and
4. the weight sequences \( \{\alpha_t\} \) and \( \{\beta_t\} \) satisfy Assumption 5.2

Then the following hold for the processes \( \{\tilde{v}_n(t)\}, n = 1, \cdots, N \):
(1) for each \( n \) and \( \tau \in [0, 1/2) \), we have

\[
P_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^\tau \| \tilde{\nu}_n(t) - \theta^* \| = 0 \right) = 1;
\]

(2) for each \( n \), the sequence \( \{ \tilde{\nu}_n(t) \} \), viewed as an estimate of \( \theta^* \), is asymptotically efficient as per Definition 2.1, i.e.,

\[
\sqrt{t+1} (\tilde{\nu}_n(t) - \theta^*) \Rightarrow N \left( 0, I^{-1}(\theta^*) \right).
\]

**Proof:** Note that, by Proposition 5.2, for each \( n \) the local Fisher information matrix \( I_n(\theta^*) \) is positive semidefinite; hence, there exists (for example, by a Cholesky factorization) a positive integer \( M_n \) and an \( M_n \times M \) matrix \( B_n \) such that \( I_n(\theta^*) = B_n^T B_n \). By Proposition 2.1, for each \( n \), the sequence \( \{ w_n(t) \} \) possesses moments of all orders, is zero-mean with covariance \( I_n(\theta) \). Since \( I_n(\theta^*) = B_n^T B_n \), there exists another \( \{ F_{t+1} \} \) adapted sequence \( \{ \zeta_n(t) \} \) (not necessarily unique depending on the rank of the matrix \( B_n \)) satisfying condition (3) in the hypothesis of Theorem 5.1 such that \( B_n^T \zeta_n(t) = w_n(t) \) for all \( t \) a.s.

Also, for each \( n \), denote by \( \{ D_n(t) \} \) the \( M \times M_n \) matrix-valued \( \{ F_t \} \)-adapted process such that \( D_n(t) = K_n(t) B_n^\top \) for all \( t \); since, by Lemma 5.2, \( K_n(t) \to N.I^{-1}(\theta^*) \) as \( t \to \infty \) a.s., we have that \( D_n(t) \to N.I^{-1}(\theta^*) B_n^\top \) as \( t \to \infty \) a.s.

It is now clear, that the evolution of the sequences \( \{ \tilde{\nu}_n(t) \} \) may be rewritten as follows in terms of the newly introduced variables:

\[
\tilde{\nu}_n(t + 1) = \tilde{\nu}_n(t) - \beta_t \sum_{l \in \Omega_n(t)} (\tilde{\nu}_n(t) - \tilde{\nu}_l(t)) + \alpha_t D_n(t) (B_n(\theta^* - \tilde{\nu}_n(t)) + \zeta_n(t)).
\]

Finally noting that, by construction and Proposition 2.2,

\[
I(\theta^*) = \sum_{n=1}^N I_n(\theta^*) = \sum_{n=1}^N B_n^\top B_n,
\]

we conclude that the evolution in (51) falls under the purview of Theorem 5.1 (with the identification that \( A = I(\theta^*) \)) and the desired assertions follow.

The processes \( \{ \tilde{\nu}_n(t) \} \) as introduced and analyzed in Corollary 5.2 may, in fact, be viewed as linearizations of the nonlinear estimator dynamics, see (8), the linearizations being performed in the vicinity of the true parameter value \( \theta^* \). Clearly, in order for such linearization to provide meaningful insight into the actual nonlinear dynamics of the estimators \( \{ x_n(t) \} \)'s, it is necessary that the latter approach stay close to \( \theta^* \) (around which the linearization is carried out) asymptotically, which, in turn, is guaranteed by the consistency of the estimators \( \{ x_n(t) \} \)'s, see Corollary 5.1. The consistency allows us to obtain insight into the detailed dynamics of the estimators \( \{ x_n(t) \} \)'s by characterizing the pathwise deviations of the former from their linearized counterparts. These ideas are formalized below in Lemma 5.3 (see Appendix B for a proof) leading to the main results of the paper as presented in Section 3-B.
Lemma 5.3. For each $n$, let $\{x_n(t)\}$ be the estimate sequence at agent $n$ as defined in (8), and $\{\hat{v}_n(t)\}$ denote the process defined in (50) under the hypotheses of Corollary 5.2. Then, there exists a constant $\tau > 1/2$ such that

$$\mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^\tau \| x_n(t) - \hat{v}_n(t) \| = 0 \right) = 1$$

for all $n$.

With the above development, we may now complete the proofs of Theorem 3.1 and Theorem 3.2 as follows.

Proof of Theorem 3.1. Let $\tau \in (0, 1/2)$ and note that, for each $n$,

$$(t + 1)^\tau \| x_n(t) - \theta^* \| \leq (t + 1)^\tau \| x_n(t) - \hat{v}_n(t) \| + (t + 1)^\tau \| \hat{v}_n(t) - \theta^* \| , \quad (52)$$

where $\{\hat{v}_n(t)\}$ is the (linearized) approximation introduced and analyzed in Corollary 5.2. By Lemma 5.3 (first assertion) and Corollary 5.2, since $\tau < 1/2$, we have $(t + 1)^\tau \| x_n(t) - \hat{v}_n(t) \| \to 0$ and $(t + 1)^\tau \| \hat{v}_n(t) - \theta^* \| \to 0$ respectively as $t \to \infty$ a.s. Hence, by (52) we obtain $(t + 1)^\tau \| x_n(t) - \theta^* \| \to 0$ as $t \to \infty$ a.s., thus establishing Theorem 3.1.

Proof of Theorem 3.2. Note that by Lemma 5.3 (first assertion), for each $n$

$$\mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} \sqrt{I + 1} (x_n(t) - \theta^*) - \sqrt{I + 1} (\hat{v}_n(t) - \theta^*) \right) = 0$$

$$= \mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} \sqrt{I + 1} \| x_n(t) - \hat{v}_n(t) \| = 0 \right) = 1.$$ 

Hence, in particular, the sequences $\{\sqrt{I + 1}(x_n(t) - \theta^*)\}$ and $\{\sqrt{I + 1}(\hat{v}_n(t) - \theta^*)\}$ possess the same weak limit (if the latter exists for one of the sequences); the asymptotic normality (efficiency) in Theorem 3.2 then follows immediately by the corresponding for the $\{\hat{v}_n(t)\}$ sequence in Corollary 5.2 (second assertion).

6. CONCLUSIONS

We have developed distributed estimators of the consensus + innovations type for multi-agent scenarios with general exponential family observation statistics that yield consistent and asymptotically efficient parameter estimates at all agents. Moreover, the above estimator properties and optimality hold as long as the aggregate or global sensing model is observable and the inter-agent communication network is connected in the mean (otherwise, irrespective of the network sparsity). Along the way, we have characterized analogues of classical system and information theoretic notions such as observability to the distributed-information setting.

An important future research question arises naturally: in this paper we have assumed that the parametrization is continuous unconstrained, i.e., $\theta$ may take values over the entire space $\mathbb{R}^M$. It would be of interest to extend the approach to account for constrained parametrization – the parameter $\theta$ could belong to a restricted subset $\Theta \subset \mathbb{R}^M$ either because of direct physical constraints or due to constrained natural parameterizations of the local exponential families involved, i.e., the domains of definition of the functions $\lambda_n(\cdot)$ in (1) being strict subsets of $\mathbb{R}^M$. A specific instance being finite classification or detection (hypothesis testing) problems in which $\theta$ may only...
assume a finite set of values\(^6\). The unconstrained estimation approach \(7)-(10\) may still be applicable to a subclass of such constrained cases by considering suitable analytical extensions of the various functions \(\lambda_n(\cdot)\)'s, \(h_n(\cdot)\)'s etc. over \(\mathbb{R}^M\); provided such extensions exist\(^7\) the proposed algorithm will lead to asymptotically efficient estimates at the network agents although the intermediate iterates may not belong to \(\Theta\). As a familiar example where such extension may be achievable by embedding, we may envision a binary hypothesis testing problem corresponding to the presence or absence of a signal observed in additive zero-mean Gaussian noise with known variance. In cases, where such analytical extensions may not be obtained, other modifications of the proposed scheme, for example by supplementing the local estimate update processes with a projection step onto the set \(\Theta\), may be helpful. In the interest of obtaining a unified distributed inference framework, it would be worthwhile to study such extensions and modifications of the proposed scheme.

APPENDIX A
PROOFS OF RESULTS IN SECTION 4

Proof of Proposition 4.1 \(\text{[27]}\) Let \(z \in \Gamma_\varepsilon\) and note that by reasoning along the lines of \(20\) we obtain
\[
(z - \theta^*)^T (I \otimes \mathcal{K}^{-1}) (z - \theta^*) = (z_{c\perp})^T (I \otimes \mathcal{K}^{-1}) z_{c\perp} \geq \lambda_2(T) \lambda_1(K^{-1}) \|z_{c\perp}\|^2,
\]
where \(z_{c\perp}\) denotes the projection of \(z\) onto the orthogonal complement of the consensus subspace (see Definition 4.1). We thus obtain
\[
\mathcal{H}_L(z) \geq \frac{b \beta}{\alpha_\varepsilon} \lambda_2(T) \lambda_1(K^{-1}) \|z_{c\perp}\|^2 + (z - \theta^*)^T (\mathcal{T}_L(z) - \mathcal{T}_L(\theta^*)) \geq \frac{b \beta}{\alpha_\varepsilon} \lambda_2(T) \lambda_1(K^{-1}) \|z_{c\perp}\|^2 + \left( z_{c\perp} - \lambda \right)^T (\mathcal{T}_L(z) - \mathcal{T}_L(\theta^*)) + (z_{c\perp})^T (\mathcal{T}_L(z) - \mathcal{T}_L(\theta^*)).
\]
In order to bound the last two terms in the above inequality, note that, for \(z \in \Gamma_\varepsilon\), by invoking Assumption 3.3 we obtain
\[
\left| (z_{c\perp})^T (\mathcal{T}_L(z) - \mathcal{T}_L(\theta^*)) \right| \leq c_1 \|z_{c\perp}\| \left( 1 + \|z - \theta^*\| \right) \leq c_1 (1/\varepsilon + 1) \|z_{c\perp}\|,
\]
where \(c_1\) is a positive constant. Also, by Proposition 4.1 the functions \(h_n(\cdot)\) are infinitely continuously differentiable and hence locally Lipschitz; in particular, noting that the set \(\Gamma'_\varepsilon\)
\[
\Gamma'_\varepsilon = \left\{ z \in \mathbb{R}^M : \|z - \theta^*\| \leq 1/\varepsilon \right\}
\]
\(^6\)We are somewhat abusing the notion of estimation which, to be precise, corresponds to inferring continuous parameters as pursued in this paper. However, by considering constrained parametrization, we are essentially expanding its usage to general inference problems including detection and classification.
\(^7\)An idea related to such analytical extensions is that of embedding into an exponential family (see [39] for some discussion in a related but centralized context), in which, broadly speaking, the objective is to obtain an unconstrained exponential family whose restriction to \(\Theta\) coincides with the given constrained family.
is compact, there exists a constant $\ell_{\Gamma_{\varepsilon}} > 0$, such that,

$$\|\overline{h}(z) - \overline{h}(z')\| \leq \ell_{\Gamma_{\varepsilon}} \|z - z'\|, \quad \forall z, z' \in \Gamma_{\varepsilon}.$$  \hspace{1cm} (56)

By observing that for $z \in \Gamma_{\varepsilon} \subset \Gamma_{\varepsilon}'$

$$\|z_c - \overline{\theta}\| = \|z - \overline{\theta}\| - \|z_{c,\perp}\| \leq \|z - \overline{\theta}\| \leq 1/\varepsilon,$$

we obtain $z_c \in \Gamma_{\varepsilon}'$; hence, by (56), we may conclude that

$$\|\overline{h}(z) - \overline{h}(z_c)\| \leq \ell_{\Gamma_{\varepsilon}} \|z - z_c\| = \ell_{\Gamma_{\varepsilon}} \|z_{c,\perp}\|$$

for all $z \in \Gamma_{\varepsilon}$. Thus, for $z \in \Gamma_{\varepsilon}$, we have

$$\left( z_{c,\perp} \right)^\top (\overline{h}(z) - \overline{h}(z_c)) \leq \ell_{\Gamma_{\varepsilon}} \|z_{c,\perp}\| \|z - \overline{\theta}\| \leq (\ell_{\Gamma_{\varepsilon}}/\varepsilon) \|z_{c,\perp}\|. \hspace{1cm} (57)$$

Combining (56)-(58) and (57) we then obtain

$$\mathcal{H}_{\varepsilon}(z) \geq \left( \frac{b_2 \beta_1}{\alpha_l} \lambda_2(\overline{L}) \lambda_1(K^{-1}) \|z_{c,\perp}\| - c_1 \left( \frac{1}{\varepsilon} + 1 \right) - \frac{\ell_{\Gamma_{\varepsilon}}}{\varepsilon} \right) \|z_{c,\perp}\|$$

$$+ \left( z_c - \overline{\theta} \right)^\top (\overline{h}(z_c) - \overline{h}(\overline{\theta})) \hspace{1cm} (58)$$

for all $z \in \Gamma_{\varepsilon}$. By invoking standard properties of quadratic minimization, we note that there exist positive constants $\overline{c}_\varepsilon$, $c_3(\varepsilon)$ and $c_4(\varepsilon)$ such that for all $t \geq 0$

$$\left( \frac{b_2 \beta_1}{\alpha_l} \lambda_2(\overline{L}) \lambda_1(K^{-1}) \|z_{c,\perp}\| - c_1 \left( \frac{1}{\varepsilon} + 1 \right) - \frac{\ell_{\Gamma_{\varepsilon}}}{\varepsilon} \right) \|z_{c,\perp}\| > \overline{c}_\varepsilon \hspace{1cm} (59)$$

for all $z$ with $\|z_{c,\perp}\| > c_3(\varepsilon) \left( \alpha_l/\beta_l \right)$, and

$$\left( \frac{b_2 \beta_1}{\alpha_l} \lambda_2(\overline{L}) \lambda_1(K^{-1}) \|z_{c,\perp}\| - c_1 \left( \frac{1}{\varepsilon} + 1 \right) - \frac{\ell_{\Gamma_{\varepsilon}}}{\varepsilon} \right) \|z_{c,\perp}\| \geq -c_4(\varepsilon) \left( \alpha_l/\beta_l \right) \hspace{1cm} (60)$$

for all $z$, in particular, $z \in \Gamma_{\varepsilon}$.

Now, note that, by Proposition 2.1 (third assertion), for all $z \in \mathbb{R}^{N\times M}$ with $z_c \neq \overline{\theta}$ we have

$$\left( z_c - \overline{\theta} \right)^\top (\overline{h}(z_c) - \overline{h}(\overline{\theta})) = \sum_{n=1}^N (z_{c,\perp}^n - \theta^n)^\top (h_n(z_{c,\perp}^n) - h_n(\theta^n)) > 0$$

(see also Definition 1.1). Let us choose $\varepsilon'$ such that $\varepsilon' \in (0, \varepsilon)$; noting that the functions $h_n(\cdot)$ are continuous and the set $\Phi_{\varepsilon,\varepsilon'}$

$$\Phi_{\varepsilon,\varepsilon'} = \Gamma_{\varepsilon} \cap \left\{ z \in \mathbb{R}^{N\times M} : \|z_c - \overline{\theta}\| \geq \varepsilon' \right\}$$

is compact, we conclude that there exists $\delta_{\varepsilon} > 0$ such that

$$\inf_{z \in \Phi_{\varepsilon,\varepsilon'}} \left( z_c - \overline{\theta} \right)^\top (\overline{h}(z_c) - \overline{h}(\overline{\theta})) > \delta. \hspace{1cm} (61)$$

Further, since $\alpha_l/\beta_l \to 0$ as $t \to \infty$ (by hypothesis), there exist $t_{\varepsilon}$ large enough and a constant $\bar{\delta}_{\varepsilon} > 0$ such that

$$\varepsilon' < \varepsilon - c_3(\varepsilon) \left( \alpha_l/\beta_l \right) \quad \text{and} \quad c_4(\varepsilon) \left( \alpha_l/\beta_l \right) < \delta_{\varepsilon} - \bar{\delta}_{\varepsilon} \hspace{1cm} (62)$$
for all $t \geq t_\varepsilon$.

We now show that there exists $\delta_\varepsilon' > 0$ (independent of $t$) such that

$$\inf_{z \in \Gamma_\varepsilon} \mathcal{H}_t(z) > \delta_\varepsilon'$$  \hspace{1cm} (63)

for all $t \geq t_\varepsilon$. To this end, for $t \geq t_\varepsilon$, let $z \in \Gamma_\varepsilon$ and consider the two cases as to whether $\|z_{c, -}\| > c_3(\varepsilon)(\alpha_t/\beta_t)$ or not. Noting that by Proposition 2.1

$$\left(z_c - \bar{\theta}\right)^\top \left(\bar{\Pi}(z_c) - \bar{\Pi}(\bar{\theta})\right) \geq 0, \quad \forall z \in \mathbb{R}^{NM},$$

we have by (58)-(59) that

$$\mathcal{H}_t(z) > \bar{c}_\varepsilon$$  \hspace{1cm} (64)

for all $z \in \Gamma_\varepsilon$ with $\|z_{c, -}\| > c_3(\varepsilon)(\alpha_t/\beta_t)$. Now consider the other case, i.e., let $z \in \Gamma_\varepsilon$ with $\|z_{c, -}\| \leq c_3(\varepsilon)(\alpha_t/\beta_t)$; note that, since $t \geq t_\varepsilon$, we have for such $z$ by (62)

$$\|z_{c, -}\| = \|z - \bar{\theta}\| - \|z_{c, +}\| \geq \varepsilon - c_3(\varepsilon)(\alpha_t/\beta_t)\varepsilon'.$$

Hence, necessarily $z \in \Phi_{\varepsilon, \varepsilon'}$ and we have by (58)-(59), and (61)-(62)

$$\mathcal{H}_t(z) > \delta_\varepsilon - c_4(\varepsilon)(\alpha_t/\beta_t) > \bar{\delta}_\varepsilon.$$  \hspace{1cm} (65)

From (64) and (65) we then obtain for all $z \in \Gamma_\varepsilon$ and $t \geq t_\varepsilon$

$$\mathcal{H}_t(z) > \delta_\varepsilon' > 0,$$

where $\delta_\varepsilon' = \bar{c}_\varepsilon \wedge \bar{\delta}_\varepsilon$, thus establishing the assertion in (63).

Finally, let $\bar{b}_\varepsilon = \varepsilon^2 \bar{\delta}_\varepsilon$, and note that the desired claim follows by

$$\mathcal{H}_t(z) > \delta_\varepsilon' \geq \delta_\varepsilon' \left(\varepsilon^2 \|z - \bar{\theta}\|^2\right) = \bar{b}_\varepsilon \|z - \bar{\theta}\|^2$$

for all $z \in \Gamma_\varepsilon$ and $t \geq t_\varepsilon$, where we used that fact that $\varepsilon^2 \|z - \bar{\theta}\|^2 \leq 1$ on $\Gamma_\varepsilon$. $\blacksquare$

**Proof of Lemma 4.2** Before proceeding to the proof of Lemma 4.2 we state the following approximation results obtained in [21] on convergence estimates of stochastic recursions (Lemma A.1) and certain attributes of time-varying stochastic Laplacian matrices (Lemma A.2), to be used as intermediate ingredients in the proof.

**Lemma A.1** (Lemma 4.3 in [21]). Let $\{w_t\}$ be an $\mathbb{R}_+\text{-valued} \\{\mathcal{F}_t\}$ adapted process that satisfies

$$w_{t+1} \leq (1 - r_1(t))w_t + r_2(t)U_t(1 + J_t).$$

In the above, $\{r_1(t)\}$ is an $\{\mathcal{F}_{t+1}\}$ adapted process, such that for all $t$, $r_1(t)$ satisfies $0 \leq r_1(t) \leq 1$ and

$$\frac{a_1}{(t + 1)^{a_1}} \leq \mathbb{E}[r_1(t) | \mathcal{F}_t] \leq 1$$

for all $t \geq t_\varepsilon$. To this end, for $t \geq t_\varepsilon$, let $z \in \Gamma_\varepsilon$ and consider the two cases as to whether $\|z_{c, -}\| > c_3(\varepsilon)(\alpha_t/\beta_t)$ or not. Noting that by Proposition 2.1

$$\left(z_c - \bar{\theta}\right)^\top \left(\bar{\Pi}(z_c) - \bar{\Pi}(\bar{\theta})\right) \geq 0, \quad \forall z \in \mathbb{R}^{NM},$$

we have by (58)-(59) that

$$\mathcal{H}_t(z) > \bar{c}_\varepsilon$$  \hspace{1cm} (64)

for all $z \in \Gamma_\varepsilon$ with $\|z_{c, -}\| > c_3(\varepsilon)(\alpha_t/\beta_t)$. Now consider the other case, i.e., let $z \in \Gamma_\varepsilon$ with $\|z_{c, -}\| \leq c_3(\varepsilon)(\alpha_t/\beta_t)$; note that, since $t \geq t_\varepsilon$, we have for such $z$ by (62)

$$\|z_{c, -}\| = \|z - \bar{\theta}\| - \|z_{c, +}\| \geq \varepsilon - c_3(\varepsilon)(\alpha_t/\beta_t)\varepsilon'.$$

Hence, necessarily $z \in \Phi_{\varepsilon, \varepsilon'}$ and we have by (58)-(59), and (61)-(62)

$$\mathcal{H}_t(z) > \delta_\varepsilon - c_4(\varepsilon)(\alpha_t/\beta_t) > \bar{\delta}_\varepsilon.$$  \hspace{1cm} (65)

From (64) and (65) we then obtain for all $z \in \Gamma_\varepsilon$ and $t \geq t_\varepsilon$

$$\mathcal{H}_t(z) > \delta_\varepsilon' > 0,$$

where $\delta_\varepsilon' = \bar{c}_\varepsilon \wedge \bar{\delta}_\varepsilon$, thus establishing the assertion in (63).

Finally, let $\bar{b}_\varepsilon = \varepsilon^2 \bar{\delta}_\varepsilon$, and note that the desired claim follows by

$$\mathcal{H}_t(z) > \delta_\varepsilon' \geq \delta_\varepsilon' \left(\varepsilon^2 \|z - \bar{\theta}\|^2\right) = \bar{b}_\varepsilon \|z - \bar{\theta}\|^2$$

for all $z \in \Gamma_\varepsilon$ and $t \geq t_\varepsilon$, where we used that fact that $\varepsilon^2 \|z - \bar{\theta}\|^2 \leq 1$ on $\Gamma_\varepsilon$. $\blacksquare$
with $a_1 > 0$ and $0 < \delta_1 < 1; \{r_2(t)\}$ is a deterministic sequence satisfying $r_2(t) \leq a_2(t + 1)^{-\delta_2}$ for all $t \geq 0$, where $a_2$ and $\delta_2$ are positive constants. Further, let $\{U_t\}$ and $\{J_t\}$ be $\mathbb{R}_+$ valued $\mathcal{F}_t$ and $\{\mathcal{F}_{t+1}\}$ adapted processes respectively with $\sup_{t \geq 0} U_t < \infty$ a.s. The process $\{J_t\}$ is i.i.d. with $J_t$ independent of $\mathcal{F}_t$ for each $t$ and satisfies the moment condition (Lemma 4.4 in [21])

\[ \text{Lemma A.2} \]

Let $\{\mathcal{F}_t\}$ be an $\mathbb{R}^{NM}$-valued $\{\mathcal{F}_t\}$ adapted process such that $w_t \in C^1_t$ for all $t$, where $C_t^1$ denotes the orthogonal complement of the consensus subspace $C$, see Definition [17]. Also, let $\{L_t\}$ be an $\{\mathcal{F}_t\}$-adapted sequence of Laplacians satisfying Assumption [37]. Then there exists an $\{\mathcal{F}_{t+1}\}$ adapted $\mathbb{R}_+$-valued process $\{r_t\}$ (depending on $\{w_t\}$ and $\{L_t\}$), a deterministic time $t_r$ (large enough), and a constant $c_r > 0$, such that $0 \leq r_t \leq 1$ a.s. and

\[ \| (I_{NM} - \beta_t L_t \otimes I_M) \mathbf{w}_t \| \leq (1 - r_t) \| \mathbf{w}_t \| \]

with

\[ \mathbb{E} [r_t \ | \ \mathcal{F}_t] \geq \frac{c_r}{(t + 1)^{\tau_2}} \ a.s. \]

for all $t \geq t_r$, where the weight sequence $\{\beta_t\}$ and $\tau_2$ are defined in (11).

**Proof of Lemma [4.2]** Recall the $\{\mathcal{F}_t\}$-adapted process $\{z_t\}$ with $z_t = \text{Vec}(z_n(t))$ for all $t$, and note that by (14) we have

\[ z_{t+1} = (I_{NM} - \beta_t L_t \otimes I_M) z_t - \alpha_t \overline{K}_t \left( \overline{h}(z_t) - \overline{h}(\overline{\theta}) \right) + \alpha_t \overline{K}_t \left( \overline{y}_t - \overline{h}(\overline{\theta}) \right), \]

the functions $\overline{h}(\cdot)$ and $\overline{y}(\cdot)$ being defined in (17). For each $n$, let $\tilde{z}_n(t) = z_n(t) - z_{0,n}$, and denote by $\{\tilde{z}_t\}$ the $\{\mathcal{F}_t\}$-adapted process where $\tilde{z}_t = \text{Vec}(\tilde{z}_n(t))$ for all $t$. Using the fact $(L_t \otimes I_M)(1_N \otimes z_{0,t}) = 0$, we have

\[ \tilde{z}_{t+1} = (I_{NM} - \beta_t L_t \otimes I_M) \tilde{z}_t - \alpha_t \overline{U}_t + \alpha_t \overline{J}_t, \]

where $\{\overline{U}_t\}$ and $\{\overline{J}_t\}$ are $\{\mathcal{F}_t\}$-adapted and $\{\mathcal{F}_{t+1}\}$-adapted processes given by

\[ \overline{U}_t = (I_{NM} - (1_N 1_N^T) \otimes I_M) \overline{K}_t \left( \overline{h}(z_t) - \overline{h}(\overline{\theta}) \right), \]

and

\[ \overline{J}_t = (I_{NM} - (1_N 1_N^T) \otimes I_M) \overline{K}_t \left( \overline{y}_t - \overline{h}(\overline{\theta}) \right) \]

respectively. Note that by hypothesis $\sup_t \| \overline{K}_t \| < \infty$ a.s., and, by Theorem [4.1], $\sup_t \| \tilde{z}_t \| < \infty$ a.s. Hence, by the linear growth condition on $\overline{h}(\cdot)$ (see Assumption [3.3]), there exists an $\{\mathcal{F}_t\}$-adapted process $\{U'_t\}$ such that

\[ \| U'_t \| \leq U'_t \| (I_{NM} - (1_N 1_N^T) \otimes I_M) \overline{K}_t \| \quad \forall t, \]

(68)
we have by (67)-(68)
\[
\|U_t\| + \|J_t\| \leq U_t (1 + J_t),
\]
with \(\{U_t\}\) being \(\{\mathcal{F}_t\}\)-adapted and \(\{J_t\}\) being the \(\{\mathcal{F}_{t+1}\}\)-adapted process, \(J_t = \|\overline{g}(y_t) - h(\overline{\theta})\|\) for all \(t\). Note that for every \(\varepsilon > 0\) we have
\[
\mathbb{E}_{\theta^*} \left[ J_t^{2+\varepsilon} \right] < \infty,
\]
which follows from the fact that \(\overline{g}(y_t)\) possesses moments of all orders (see Proposition 2.1). Hence, by (66) we obtain
\[
\|\tilde{z}_{t+1}\| \leq \|(I_{NM} - \beta_t L_t \otimes I_M) \tilde{z}_t\| + \alpha_t U_t (1 + J_t) \quad \forall t.
\]
Observe that, by construction, \(\tilde{z}_t \in C^+\) for all \(t\), and hence, by Lemma A.2 there exists an \(\{\mathcal{F}_{t+1}\}\) adapted \(\mathbb{R}_+\)-valued process \(\{r_t\}\) (depending on \(\{\tilde{z}_t\}\) and \(\{L_t\}\)), a deterministic time \(t_r\) (large enough), and a constant \(c_r > 0\), such that
\[
0 \leq r_t \leq 1 \quad \text{a.s. and}
\]
\[
\|(I_{NM} - \beta_t L_t \otimes I_M) \tilde{z}_t\| \leq (1 - r_t) \|\tilde{z}_t\|
\]
with
\[
\mathbb{E}[r_t | \mathcal{F}_t] \geq \frac{c_r}{(t + 1)^{\tau_2}} \quad \text{a.s.}
\]
for all \(t \geq t_r\). We then have by (70)-(71)
\[
\|\tilde{z}_{t+1}\| \leq (1 - r_t) \|\tilde{z}_t\| + \alpha_t U_t (1 + J_t) \quad \forall t.
\]
Now consider arbitrary \(\varepsilon > 0\) and note that, under the moment condition (69), the stochastic recursion in (72) falls under the purview of Lemma A.1 (by taking \(\delta_1\) and \(\delta_2\) in Lemma A.1 to be \(\tau_2\) and 1 respectively), and we conclude that \((t + 1)^{\tau} \|\tilde{z}_t\| \to 0\) as \(t \to \infty\) a.s. for each \(\tau \in (0, 1 - \tau_2 - 1/(2 + \varepsilon))\). Noting that
\[
\|z_n(t) - z_l(t)\| \leq \|z_n(t) - \tilde{z}_t^n\| + \|\tilde{z}_t^n - z_l(t)\| \leq 2 \|\tilde{z}_t\|
\]
for each pair \(n\) and \(l\) of agents, we may further conclude that
\[
P_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^{\tau} \|z_n(t) - z_l(t)\| = 0 \right) = 1,
\]
for all \(\tau \in (0, 1 - \tau_2 - 1/(2 + \varepsilon))\).

Since the above holds for arbitrary \(\varepsilon > 0\), the desired assertion follows by making \(\varepsilon\) tend to \(\infty\).

\textbf{Proof of Corollary 4.2:} Since \((t + 1)^{\tau_2} \|K_n(t) - K\| \to 0\) a.s. as \(t \to \infty\) for all \(n\), by Egorov’s theorem, for each \(\varepsilon > 0\), there exist a deterministic \(t_\varepsilon > 0\) and a positive constant \(c_\varepsilon\), such that
\[
P_{\theta^*} \left( \sup_{t \geq t_\varepsilon} (t + 1)^{\tau_2} \|K_n(t) - K\| > c_\varepsilon \right) < \varepsilon
\]
for all $n$. Now, for such an $\varepsilon > 0$, define, for each $n$, the following $\{\mathcal{F}_t\}$-adapted sequence $\{K_n^\varepsilon(t)\}$:

$$K_n^\varepsilon(t) = \begin{cases} 
K_n(t) & \text{if } t < t_\varepsilon \\
K_n(t) & \text{if } t \geq t_\varepsilon \text{ and } \|K_n(t) - K\| \leq c_\varepsilon(t + 1)^{-\tau_3} \\
K & \text{otherwise.}
\end{cases}$$

(75)

Note that, by the above construction, we have $\|K_n(t) - K\| \leq c_\varepsilon(t + 1)^{-\tau_3}$ for all $t \geq t_\varepsilon$; hence, choosing $\tau' \in (0, \tau_3)$ to be a constant (independent of $\varepsilon$), we have that

$$(t + 1)^{\tau'} \|K_n^\varepsilon(t) - K\| \leq c_\varepsilon(t + 1)^{-(\tau_3 - \tau')} \quad \forall t \geq t_\varepsilon$$

for all $n$ and each $\varepsilon > 0$. Thus, clearly, for each $\varepsilon > 0$ and all $n$, the sequence $\{K_n^\varepsilon(t)\}$ converges a.s. to $K$ uniformly (over sample paths) at rate $\tau' > 0$, i.e., for each $\delta > 0$, there exists (deterministic) $t_\delta > 0$ such that

$$\mathbb{P}_{\theta^*} \left( \sup_{t \geq t_\delta} (t + 1)^{\tau'} \|K_n^\varepsilon(t) - K\| \leq \delta \right) = 1.$$ 

Now, for each $\varepsilon > 0$, let us define the $\{\mathcal{F}_t\}$-adapted sequences $\{z_n^\varepsilon(t)\}$, $n = 1, \cdots, N$, evolving as

$$z_n^\varepsilon(t + 1) = z_n^\varepsilon(t) - \beta_t \sum_{l \in \Omega_n(t)} (z_n^\varepsilon(t) - z_l^\varepsilon(t)) + \alpha_t K_n^\varepsilon(t) (g_n(y_n(t)) - h_n(z_n^\varepsilon(t))).$$

Noting that

$$\left\{ \sup_{n,t} \|z_n^\varepsilon(t) - z_n(t)\| = 0 \right\} \quad \text{on} \quad \left\{ \sup_{n,t} \|K_n^\varepsilon(t) - K_n(t)\| = 0 \right\},$$

we have

$$\mathbb{P}_{\theta^*} \left( \sup_{n,t} \|z_n^\varepsilon(t) - z_n(t)\| = 0 \right) \geq 1 - N \varepsilon$$

(76)

by (74)-(75).

The uniform convergence of the gain sequences $\{K_n^\varepsilon(t)\}$ to $K$ at rate $\tau' > 0$ ensures that, for each $\varepsilon > 0$, the processes $\{z_n^\varepsilon(t)\}$ satisfy the hypotheses of Theorem 4.2, and, hence, there exists a positive constant $\mu$ (that depends on $\tau'$ but not $\varepsilon$), such that $(t + 1)^\mu \|z_n^\varepsilon(t) - \theta^\varepsilon\| \rightarrow 0$ as $t \rightarrow 0$ a.s. for each $n$. Hence, by (76) we have

$$\mathbb{P}_{\theta^*} \left( \lim_{t \rightarrow \infty} (t + 1)^\mu \|z_n(t) - \theta^\varepsilon\| = 0 \right) \geq 1 - N \varepsilon$$

(77)

for all $n$. Since $\varepsilon > 0$ is arbitrary and $\mu$ does not depend on $\varepsilon$, we may further conclude from (77) that $(t + 1)^\mu \|z_n(t) - \theta^\varepsilon\| \rightarrow 0$ as $t \rightarrow \infty$ a.s. for all $n$.

APPENDIX B

PROOFS IN SECTION 5

Proof of Lemma 5.2 The proof of Lemma 5.2 is accomplished in two steps: first, we show that the gain sequences reach consensus, and subsequently demonstrate that the limiting consensus value is indeed $I^{-1}(\theta^\varepsilon)$. To this end, consider the following:
Lemma B.1. Recall for each $n$, the $\{\mathcal{F}_t\}$-adapted sequence $\{G_n(t)\}$ evolving as in (10), and denote by $\{G^a_n\}$ their instantaneous network averages, i.e., $G^a_t = (1/N) \sum_{n=1}^N G_n^a(t)$ for all $t$. Then, for each $n$ and $\tau \in [0, 1 - \tau_2)$, we have

$$
\mathbb{P}_\theta \left( \lim_{t \to \infty} (t+1)^T \|G_n(t) - G^a(t)\| = 0 \right) = 1,
$$

where $\tau_2$ is the exponent associated with the weight sequence $\{\beta_t\}$, see Assumption 3.2.

Proof: We will show the desired convergence in the matrix Frobenius norm (denoted by $\|\cdot\|_F$ in the following), the convergence in the induced $\mathcal{L}_2$ sense following immediately. Note that, by Lemma 5.1, $\bar{x}_n(t) \to \theta^*$ as $t \to \infty$ a.s. for all $n$; hence, for each $n$, by the continuity of the local Fisher information matrix $I_n(\cdot)$, we have that $I_n(\bar{x}_n(t)) \to I_n(\theta^*)$ as $t \to \infty$ a.s.

Let $\bar{G}_n(t) = G_n(t) - G^a_t$ denote the deviation at agent $n$ from the instantaneous network average $G^a_t$ and $I^a_t = (1/N) \sum_{n=1}^N I_n(\bar{x}_n(t))$ the network average of the $I_n(\bar{x}_n(t))$'s. Also, let $\bar{G}_t$ and $\bar{I}_t$ denote the matrices $\text{Vec}(\bar{G}_n(t))$ and $\text{Vec}(\bar{I}_n(t))$ respectively, where $\bar{I}_n(t) = I_n(\bar{x}_n(t)) - I^a_t$ for all $n$. Using the following readily verifiable properties of the Laplacian $L_t$

$$
(1_N \otimes I_M)^\top (L_t \otimes I_M) = 0 \quad \text{and} \quad (L_t \otimes I_M) (1_N \otimes G^a_t) = 0,
$$

we have by (11)

$$
\bar{G}_{t+1} = (I_{NM} - \beta_t (L_t \otimes I_M) - \alpha_t I_{NM}) \bar{G}_t + \alpha_t \bar{I}_t
$$

for all $t \geq 0$.

Since for all $n$, $I_n(\bar{x}_n(t)) \to I_n(\theta^*)$ as $t \to \infty$ a.s., the sequences $\{I_n(\bar{x}_n(t))\}$ are bounded a.s. and, in particular, there exists an $\{\mathcal{F}_t\}$-adapted a.s. bounded process $\{U_t\}$ such that $\|\bar{I}_t\|_F \leq U_t$ for all $t$. For $m \in \{1, \ldots, M\}$, denote by $\bar{G}_{m,t}$ the $m$-th column of $\bar{G}_t$. Clearly, the process $\{\bar{G}_{m,t}\}$ is $\{\mathcal{F}_t\}$-adapted and $\bar{G}_{m,t} \in \mathcal{C} \perp$ for all $t$. Hence, by Lemma A.2 there exist a $[0, 1]$-valued $\{\mathcal{F}_{t+1}\}$-adapted process $\{r_{m,t}\}$ and a positive constant $c_{m,r}$ such that

$$
\left\| (I_{NM} - \beta_t L_t \otimes I_M) \bar{G}_{m,t} \right\| \leq (1 - r_{m,t}) \left\| \bar{G}_{m,t} \right\|$

and $\mathbb{E}_\theta^* [r_{m,t} | \mathcal{F}_t] \geq c_{m,r} / (t+1)^{r_2}$ a.s. for all $t \geq t_0$ sufficiently large. Noting that the square of the Frobenius norm is the sum of the squared column $\mathcal{L}_2$ norms, we have

$$
\left\| (I_{NM} - \beta_t L_t \otimes I_M) \bar{G}_t \right\|_F^2 \leq \sum_{m=1}^M (1 - r_{m,t})^2 \left\| \bar{G}_{m,t} \right\|^2 \leq (1 - r_t)^2 \left\| \bar{G}_t \right\|_F^2,
$$

where $\{r_t\}$ is the $\{\mathcal{F}_t\}$-adapted process given by $r_t = r_{1,t} \wedge \cdots \wedge r_{M,t}$ for all $t$. By the conditional Jensen’s inequality we obtain

$$
\mathbb{E}_\theta^* [r_t | \mathcal{F}_t] \geq \bigwedge_{m=1}^M \mathbb{E}_\theta^* [r_{m,t} | \mathcal{F}_t] \geq c_r / (t+1)^{r_2}.
$$
for some constant $c_r > 0$ and all $t \geq t_0$. Since $\beta_t / \alpha_t \to \infty$ as $t \to \infty$, by making $t_0$ larger if necessary, we obtain from (80)

$$
\left\| (I_{NM} - \beta_t L_t \otimes I_M - \alpha_t I_{NM}) \tilde{G}_t \right\|_F \leq \left\| (I_{NM} - \beta_t L_t \otimes I_M) \tilde{G}_t \right\|_F + \alpha_t \left\| \tilde{G}_t \right\|_F
\leq (1 - r_t) \left\| \tilde{G}_t \right\|_F + \alpha_t \left\| \tilde{G}_t \right\|_F \leq (1 - r_t/2) \left\| \tilde{G}_t \right\|_F
$$

(81)

for all $t \geq t_0$. It then follows from (79) and (81) that

$$
\left\| \tilde{G}_{t+1} \right\|_F \leq \left\| (I_{NM} - \beta_t L_t \otimes I_M - \alpha_t I_{NM}) \tilde{G}_t \right\|_F + \alpha_t U_t \leq (1 - r_t/2) \left\| \tilde{G}_t \right\|_F + \alpha_t U_t
$$

for all $t \geq t_0$. Clearly, the above recursion falls under the purview of Lemma A.1 (by setting $\delta_1, \delta_2$ and $J_t$ in Lemma A.1 to $\tau_2, 1$ and 0 respectively), and we conclude that $(t + 1)^{\tau} \left\| \tilde{G}_t \right\|_F \to 0$ as $t \to \infty$ a.s. for each $\tau \in [0, 1 - \tau_2]$. The assertion in Lemma B.1 follows immediately.

We state another approximation result from [40] regarding deterministic recursions to be used in the sequel.

**Proposition B.1** (Lemma 4.3 in [40]). Let $\{b_t\}$ be a scalar sequence satisfying

$$
b_{t+1} \leq \left(1 - \frac{c}{t+1}\right) b_t + d_t (t+1)^{-\tau}
$$

where $c > \tau, \tau > 0$, and the sequence $\{d_t\}$ is summable. Then $\limsup_{t \to \infty} (t + 1)^{\tau} b_t < \infty$.

We now complete the proof of Lemma 5.2.

**Proof of Lemma 5.2** Following the notation in the proof of Lemma B.1 and using properties of the graph Laplacian (78), the process $\{G^a_t\}$ (the instantaneous network average of the $G_n(t)$’s) may be shown to satisfy the following recursion for all $t$:

$$
G^a_{t+1} = (1 - \alpha_t) G^a_t + \alpha_t I^a_t.
$$

(82)

Noting that the local Fisher information matrices $I_n(\cdot)$ are locally Lipschitz in the argument and the fact that $\tilde{x}_n(t) \to \theta^*$ as $t \to \infty$ a.s. (see Lemma 5.1), we have that

$$
\left\| I^a_t - (1/N) I(\theta^*) \right\| = O \left( \sqrt{\frac{N}{n=1}} \left\| \tilde{x}_n(t) - \theta^* \right\| \right).
$$

(83)

Since, by Lemma 5.1 $(t + 1)^{\mu_0} \left\| \tilde{x}_n(t) - \theta^* \right\| \to 0$ as $t \to \infty$ a.s., we may further conclude from (83) that

$$
\left\| I^a_t - (1/N) I(\theta^*) \right\| = o \left( (t + 1)^{-\mu_0} \right).
$$

(84)

Now let $\tau_5$ be a positive constant such that $\tau_5 < (\mu_0 \wedge 1)$. Noting that $\alpha_t = (t+1)^{-1}$ by definition, by (84) we may then conclude that there exists an $\mathbb{R}_+$-valued $\{\mathcal{F}_t\}$-adapted stochastic process $\{\tilde{a}_t\}$, such that

$$
\alpha_t \left\| I^a_t - (1/N) I(\theta^*) \right\| \leq \tilde{a}_t (t + 1)^{-\tau_5}
$$

(85)

for all $t$, with $\{\tilde{a}_t\}$ satisfying

$$
\tilde{a}_t = o \left( (t + 1)^{-1 - \mu_0 + \tau_5} \right).
$$

(86)
By (82) and (85) we then obtain
\[
\left\| G_{t+1}^n - (1/N)I(\theta^*) \right\| \leq (1 - (t + 1)^{-1}) \left\| G_t^n - (1/N)I(\theta^*) \right\|
\]
\[+ \alpha_t \left\| I_t^n - (1/N)I(\theta^*) \right\| \]
\[\leq (1 - (t + 1)^{-1}) \left\| G_t^n - (1/N)I(\theta^*) \right\| + \tilde{d}_t(t + 1)^{-\tau_6}
\]
for all $t$. Further, by (86), we have $\sum_t \tilde{d}_t < \infty$ a.s. (since $\tau_5 < \mu_0$ by construction); also noting that $\tau_5 < 1$ (again by construction), a pathwise application of Proposition B.1 to the stochastic recursion (87) yields
\[
\limsup_{t \to \infty} (t + 1)^{\tau_6} \left\| G_t^n - (1/N)I(\theta^*) \right\| < \infty \text{ a.s.,}
\]
from which we may further conclude that $(t + 1)^{\tau_6} \left\| G_t^n - (1/N)I(\theta^*) \right\| \to 0$ as $t \to \infty$ a.s., where $\tau_6$ is another positive constant such that $\tau_6 < \tau_5$.

Now introducing another constant $\tau_7$ such that $0 < \tau_7 < (1 - \tau_2) \wedge \tau_6$, by Lemma B.1 it may be readily concluded that
\[
\mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^{\tau_7} \left\| G_n(t) - (1/N)I(\theta^*) \right\| = 0 \right) = 1
\]
for all $n$. Finally, noting that matrix inversion is a locally Lipschitz operator in a neighborhood of an invertible argument, we have by (9), (12), and (88) that
\[
\left\| K_n(t) - N.I^{-1}(\theta^*) \right\| = \left\| (G_n(t) + \varphi_t I_M)^{-1} - N.I^{-1}(\theta^*) \right\|
\]
\[= O \left( \left\| G_n(t) - (1/N)I(\theta^*) \right\| + \varphi_t \right) = O \left( (t + 1)^{-\tau_7} + (t + 1)^{-\mu_2} \right)
\]
\[= o \left( (t + 1)^{-\tau'} \right),
\]
where $\tau'$ may be taken to be an arbitrary positive constant satisfying $\tau' < \tau_7 \wedge \mu_2$. Hence, the desired assertion follows.

**Proof of Lemma 5.3** The following intermediate approximation will be used in the proof of Lemma B.3.

**Lemma B.2.** For each $n$, let \( \{x_n(t)\} \) and \( \{\tilde{v}_n(t)\} \) be as in the hypothesis of Lemma 5.3 and denote by \( \{\tilde{u}_n(t)\} \) the \( \{F_t\}\)-adapted process such that \( \tilde{u}_n(t) = x_n(t) - \tilde{v}_n(t) \) for all $t$. Then, for each $\gamma \in [0, 1 - \tau_2]$ (where $\tau_2$ is the exponent corresponding to $\{\beta_1\}$, see Assumption 3.2), we have
\[
\mathbb{P}_{\theta^*} \left( \lim_{t \to \infty} (t + 1)^{\gamma} \left\| \tilde{u}_n(t) - \tilde{u}_l(t) \right\| = 0 \right) = 1
\]
for all pairs $(n, l)$ of network agents.

**Proof:** By (8) and (50), the process \( \{\tilde{u}_n(t)\} \) is readily seen to satisfy the recursions
\[
\tilde{u}_n(t + 1) = \tilde{u}_n(t) - \beta_t \sum_{n=1}^N (\tilde{u}_n(t) - \tilde{u}_l(t)) - \alpha_t K_n(t) U_n(t),
\]
(89)
where

\[ \mathbb{U}_n'(t) = h_n(x_n(t)) - h_n(\theta^*) - I_n(\theta^*) (\alpha_n(t) - \theta^*) \]  \tag{90} \]

for all \( t \). Noting that the processes \{\(x_n(t)\)\} and \{\(\bar{v}_n(t)\)\} converge a.s. as \( t \to \infty \) (see Corollary 5.1 and Corollary 5.2), we conclude that the sequence \{\(U'_n(t)\)\}, thus defined, is bounded a.s. Denoting by \(U'_t\) and \(\bar{u}_t\) the block-vectors \(\text{Vec}(U'_n(t))\) and \(\text{Vec}(\bar{v}_n(t))\) respectively, from (89) we then have

\[ \bar{u}_{t+1} = (I_{NM} - \beta_t L_t \otimes I_M) \bar{u}_t - \alpha_t u_t, \]

where \(\{u_t\}\) is the \(\{\mathcal{F}_t\}\)-adapted process given by \(u_t = \text{Diag}(K_n(t)) \cdot U'_t\) for all \( t \). Noting that \(\{U'_t\}\) is bounded a.s. and the adaptive gain sequence \(\{K_n(t)\}\) converges a.s. as \( t \to \infty \) for all \( n \), the process \(\{u_t\}\) is readily seen to be bounded a.s. Further, denoting by \(\{\bar{u}_t\}\) and \(\{\bar{U}_t\}\) the processes, such that,

\[ \bar{u}_t = \left( I_{NM} - I_N \cdot (1_N \otimes I_M)^\top \right) u_t \quad \text{and} \quad \bar{U}_t = \left( I_{NM} - I_N \cdot (1_N \otimes I_M)^\top \right) u_t \]

for all \( t \), we have (using standard properties of the Laplacian)

\[ \bar{u}_t = (I_{NM} - \beta_t L_t \otimes I_M) \bar{u}_t - \alpha_t \bar{U}_t \]  \tag{91} \]

for all \( t \). Clearly, \(\bar{u}_t \in \mathcal{C}^1\) for all \( t \), and we may note that, at this point the evolution (91) resembles the dynamics analyzed in Lemma 4.2 (for the process \(\{\bar{z}_t\}\), see (70)). Following essentially similar arguments as in (70)-(73), we have \((t+1)\gamma \|\bar{u}_t\| \to 0\) as \( t \to \infty \) a.s. for all \( \gamma \in [0, 1 - \tau_2] \), from which the desired assertion follows.

\[ \text{Proof of Lemma 5.3} \]

In what follows we stick to the notation in the proof of Lemma B.2. By (89) we have that

\[ \bar{u}_{t+1}^a = \bar{u}_t - (1/N) \alpha_t \sum_{n=1}^N K_n(t) U'_n(t), \]

where \(\bar{u}_t^a = (1/N) \sum_{n=1}^N \bar{u}_n(t)\) for all \( t \). Now note that, for each \( n \), the function \(h_n(\cdot)\) is twice continuously differentiable with gradient \(I_n(\cdot)\) (see Proposition 2.1 and Proposition 2.2), and hence there exist positive constants \(\tau\) and \(R\), such that for each \( n \),

\[ \|h_n(z) - h_n(\theta^*) - I_n(\theta^*) (z - \theta^*)\| \leq \tau \|z - \theta^*\|^2 \]  \tag{92} \]

for all \( z \in \mathbb{R}^M \) with \(\|z - \theta^*\| \leq R\). Since \(x_n(t) \to \theta^*\) as \( t \to \infty \) a.s. (see Corollary 5.1) for each \( n \), there exists a finite random time \(t_R\) such that

\[ \sup_{n=1}^N \|x_n(t) - \theta^*\| \leq R \ \forall t \geq t_R \text{ a.s.} \]  \tag{93} \]

Hence, by (90) and (92), (93), we have that

\[ U'_n(t) = h_n(x_n(t)) - h_n(\theta^*) - I_n(\theta^*) (\bar{v}_n(t) - \theta^*) = I_n(\theta^*) (x_n(t) - \bar{v}_n(t)) + R_n(t) \]

\[ = I_n(\theta^*) \bar{u}_n(t) + R_n(t), \]
for all \( n \) and \( t \), where the residuals \( R_n(t) \), \( n = 1, \ldots, N \) satisfy

\[
\|R_n(t)\| \leq \tau \|x_n(t) - \theta^*\|^2 \quad \forall t \geq t_R \text{ a.s.}
\]

Standard algebraic manipulations further yield

\[
\|R_n(t)\| \leq \tau \|x_n(t) - \theta^*\|^2 \leq 2\tau \|x_n(t) - \bar{v}_n(t)\|^2 + 2\tau \|\bar{v}_n(t) - \theta^*\|^2
\]

\[
= 2\tau \|	ilde{u}_n(t)\|^2 + 2\tau \|\bar{v}_n(t) - \theta^*\|^2 \leq 4\tau \|	ilde{u}_n^p\|^2 + 4\tau \|	ilde{u}_n(t) - \tilde{u}_r^n\|^2 + 2\tau \|\bar{v}_n(t) - \theta^*\|^2
\]

for all \( t \geq t_R \) a.s.

Note that, the fact that \((t + 1)^\gamma \|\bar{v}_n(t) - \theta^*\| \to 0 \) as \( t \to \infty \) a.s. for all \( n \) and \( \tau \in [0, 1) \) implies that there exists a constant \( \gamma_1 > 1/2 \) such that

\[
\max_{n=1}^N \|\bar{v}_n(t) - \theta^*\|^2 = o \left( (t + 1)^{-\gamma_1} \right) \quad \text{a.s.}
\]

Also, by Lemma B.2 and the fact that \( \gamma_2 < 1/2 \) (see Assumption 3.2), we have that

\[
\max_{n=1}^N \|\tilde{u}_n(t) - \tilde{u}_r^n\|^2 = o \left( (t + 1)^{-\gamma_2} \right) \quad \text{a.s.}
\]

for some constant \( \gamma_2 > 1/2 \).

By the previous construction, the recursions for \( \{\tilde{u}_r^n\} \) may be written as

\[
\tilde{u}_{r+1}^n = \tilde{u}_r^n - \alpha_t Q_t \tilde{u}_r^n - \alpha_t \tilde{R}_t,
\]

where

\[
Q_t = (1/N) \sum_{n=1}^N K_n(t) I_n(\theta^*),
\]

and

\[
\tilde{R}_t = (1/N) \sum_{n=1}^N K_n(t) (I_n(\theta^*) (\tilde{u}_n(t) - \tilde{u}_r^n) + R_n(t))
\]

for all \( t \). By (94) and (98) we obtain

\[
\left\| \tilde{R}_t \right\| \leq (1/N) \sum_{n=1}^N \|K_n(t) I_n(\theta^*)\| \left\| \tilde{u}_n(t) - \tilde{u}_r^n \right\| + (4\tau/N) \sum_{n=1}^N \|K_n(t) I_n(\theta^*)\| \left\| \tilde{u}_r^n \right\|^2 + (4\tau/N) \sum_{n=1}^N \|K_n(t) I_n(\theta^*)\| \left\| \tilde{u}_n(t) - \tilde{u}_r^n \right\|^2 + (2\tau/N) \sum_{n=1}^N \|K_n(t) I_n(\theta^*)\| \left\| \bar{v}_n(t) - \theta^* \right\|^2
\]

for \( t \geq t_R \) a.s. Then, denoting by \( \{\tilde{X}_t\} \) the \( \{\tilde{F}_t\} \)-adapted process such that

\[
\tilde{X}_t = (4\tau/N) \sum_{n=1}^N \|K_n(t) I_n(\theta^*)\| \left\| \tilde{u}_r^n \right\|
\]
for all $t$, and observing that, by (95) and (96),
\[
(2\pi/N) \sum_{n=1}^{N} \| K_n(t) I_n(\theta^*) \| \| \hat{\nu}_n(t) - \theta^* \|^2 = o \left( \left( t + 1 \right)^{-\gamma_1} \right) \quad \text{a.s.},
\]
\[
(1/N) \sum_{n=1}^{N} \| K_n(t) I_n(\theta^*) \| \| \tilde{u}_n(t) - \tilde{u}_t^\alpha \|^2 = o \left( \left( t + 1 \right)^{-\gamma_2} \right) \quad \text{a.s.},
\]
and
\[
(4\pi/N) \sum_{n=1}^{N} \| K_n(t) I_n(\theta^*) \| \| \tilde{u}_n(t) - \tilde{u}_t^\alpha \|^2 = o \left( \left( t + 1 \right)^{-\gamma_2} \right) \quad \text{a.s.}
\]
(note that the gain sequences $\{ K_n(t) \}$’s converge a.s., hence, $K_n(t) = O(1)$ for all $n$), we obtain the following from (99):
\[
\left\| \hat{R}_t \right\| \leq \bar{T}_t \left\| \tilde{u}_t^\alpha \right\| + o \left( \left( t + 1 \right)^{-\gamma_3} \right) \quad \text{for some constant } \gamma_3 \text{ such that } 1/2 < \gamma_3 < \gamma_1 \land \gamma_2.
\]
Thus, by (97)-(98) and (100), and by making $t_R$ larger if necessary, we conclude that there exists a positive constant $b$ such that
\[
\left\| \tilde{u}_{t+1}^\alpha \right\| \leq \left\| I_M - \alpha_t Q_t + \alpha_t \bar{T}_t I_M \right\| \left\| \tilde{u}_t^\alpha \right\| + b \alpha_t (t + 1)^{-\gamma_3} \quad \text{(101)}
\]
for all $t \geq t_R$ a.s. Since $K_n(t) \to N.I^{-1}(\theta^*)$ as $t \to \infty$ a.s. for each $n$ and $\sum_{n=1}^{N} I_n(\theta^*) = I(\theta^*)$, we have $Q_t \to I_M$ as $t \to \infty$ a.s.; similarly, since for all $n$ the sequences $\{ x_n(t) \}$ and $\{ \hat{\nu}_n(t) \}$ converge to $\theta^*$ a.s. as $t \to \infty$ (see Corollary 5.3 and Corollary 5.2), it follows (from definition) that $\tilde{u}_n(t) \to 0$ as $t \to \infty$ a.s. for all $n$, and hence $\bar{T}_t \to 0$ as $t \to \infty$ a.s. The fact that, $Q_t \to I_M$ and $\bar{T}_t \to 0$ as $t \to \infty$ a.s., ensures that, by making $t_R$ larger if necessary, the following holds
\[
\left\| I_M - \alpha_t Q_t + \alpha_t \bar{T}_t I_M \right\| \leq 1 - (2/3).\alpha_t = 1 - (2/3).\left( t + 1 \right)^{-1} \quad \text{(102)}
\]
for all $t \geq t_R$ a.s. Let $\gamma_4$ be a constant such that $1/2 < \gamma_4 < \gamma_3 \land (2/3)$; then, by (101)-(102), we have
\[
\left\| \tilde{u}_{t+1}^\alpha \right\| \leq \left( 1 - (2/3).\left( t + 1 \right)^{-1} \right) \left\| \tilde{u}_t^\alpha \right\| + d_t \left( t + 1 \right)^{-\gamma_4}
\]
for all $t \geq t_R$ a.s., where $d_t = b \alpha_t (t + 1)^{\gamma_3 - \gamma_4}$. Since $\gamma_4 < 2/3$ and the sequence $\{ d_t \}$ is summable, a pathwise application of Proposition 5.1 yields
\[
\mathbb{P}_{\theta^*} \left( \limsup_{t \to \infty} \left( t + 1 \right)^{\gamma_4} \left\| \tilde{u}_t^\alpha \right\| < \infty \right) = 1.
\]
Hence, by choosing $\tau$, such that $1/2 < \tau < \gamma_2 \land \gamma_4$ (where $\gamma_2$ is defined in (96)), we have that $(t + 1)^{\tau} \tilde{u}_n(t) \to 0$ as $t \to \infty$ a.s. for all $n$ and the desired assertion follows.

**REFERENCES**

[1] D. Bertsekas, J. Tsitsiklis, and M. Athans, “Convergence theories of distributed iterative processes: A survey,” Technical Report for Information and Decision Systems, Massachusetts Inst. of Technology, Cambridge, MA, 1984.
[2] 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.

[3] 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.

[4] 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.

[5] J. Tsitsiklis, “Problems in decentralized decision making and computation,” PhD, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[6] 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.

[7] 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.

[8] S. Kar, J. 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, initial Submission: Aug. 2008. [Online]. Available: http://arxiv.org/abs/0809.0009

[9] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. Moura, “Distributed detection via gaussian running consensus: Large deviations asymptotic analysis,” *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 4381–4396, May 2011.

[10] D. Jakovetic, J. 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.

[11] 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.

[12] 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.

[13] S. Kar, J. Moura, and H. Poor, “QD-learning: a collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations,” *IEEE Transactions on Signal Processing*, initial Submission: April 2012. Revised: Oct. 2012. Accepted: Dec. 2012. [Online]. Available: http://arxiv.org/abs/1205.0047

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] S. Kar and J. 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.

[21] 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.
[23] 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.

[24] 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.

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

[26] M. Nevel’son and R. Has’minskii, *Stochastic Approximation and Recursive Estimation*. Providence, Rhode Island: American Mathematical Society, 1973.

[27] T. Li, M. Fu, L. Xie, and J.-F. Zhang, “Distributed consensus with limited communication data rate,” *IEEE Transactions on Automatic Control*, vol. 56, no. 2, pp. 279–292, Feb 2011.

[28] M. Huang and J. Manton, “Stochastic approximation for consensus seeking: mean square and almost sure convergence,” in *Proceedings of the 46th IEEE Conference on Decision and Control*, New Orleans, LA, USA, Dec. 12-14 2007.

[29] S. Kar and J. Moura, “Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise,” *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 355–369, Jan. 2009.

[30] S. B. Gelfand and S. K. Mitter, “Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$,” *SIAM J. Control Optim.*, vol. 29, no. 5, pp. 999–1018, September 1991.

[31] V. S. Borkar, *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge, UK: Cambridge University Press, 2008.

[32] G. Yin and Q. Zhang, *Discrete-time Markov Chains: Two-time-scale Methods and Applications*. New York, NY: Springer, 2005.

[33] V. Krishnamurthy, K. Topley, and G. Yin, “Consensus formation in a two-time-scale Markovian system,” *SIAM Journal: Multiscale Modeling and Simulation*, vol. 7, no. 4, pp. 1898–1927, July 2009.

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

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

[36] 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.

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

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

[39] A. Rukhin, “Recursive testing of multiple hypotheses: Consistency and efficiency of the Bayes rule,” *The Annals of Statistics*, vol. 22, no. 2, pp. 616–633, Jun. 1994.

[40] V. Fabian, “Stochastic approximation of minima with improved asymptotic speed,” *The Annals of Mathematical Statistics*, vol. 37, no. 1, pp. 191–200, Feb 1967.