Distributed Parameter Estimation in Sensor Networks: Nonlinear Observation Models and Imperfect Communication

Soummya Kar, José M. F. Moura and Kavita Ramanan

Abstract

The paper studies distributed static parameter (vector) estimation in sensor networks with nonlinear observation models and noisy inter-sensor communication. It introduces *separably estimable* observation models that generalize the observability condition in linear centralized estimation to nonlinear distributed estimation. It studies two distributed estimation algorithms in separably estimable models, the $\mathcal{NU}$ (with its linear counterpart $\mathcal{LU}$) and the $\mathcal{NLU}$. Their update rule combines a *consensus* step (where each sensor updates the state by weight averaging it with its neighbors' states) and an *innovation* step (where each sensor processes its local current observation.) This makes the three algorithms of the *consensus* + *innovations* type, very different from traditional consensus. The paper proves consistency (all sensors reach consensus almost surely and converge to the true parameter value,) efficiency, and asymptotic unbiasedness. For $\mathcal{LU}$ and $\mathcal{NU}$, it proves asymptotic normality and provides convergence rate guarantees. The three algorithms are characterized by appropriately chosen decaying weight sequences. Algorithms $\mathcal{LU}$ and $\mathcal{NU}$ are analyzed in the framework of stochastic approximation theory; algorithm $\mathcal{NLU}$ exhibits mixed time-scale behavior and biased perturbations, and its analysis requires a different approach that is developed in the paper.

Keywords: Asymptotic normality, consensus, consensus + innovations, consistency, distributed parameter estimation, Laplacian, separable estimable, spectral graph theory, stochastic approximation, unbiasedness.

I. Introduction

A. Background and Motivation

The paper studies distributed inference, in particular, distributed estimation, as *consensus+innovations* algorithms that generalize distributed consensus by combining, at each time step, cooperation among agents (*consensus*) with assimilation of their observations (*innovations*). Our *consensus* + *innovations* algorithms contrast with: i) standard consensus, see the extensive literature, e.g., [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], where in each time step only local averaging of the neighbors' states occurs, and no observations are processed; and ii) distributed estimation algorithms, see recent literature,e.g., [15], [16], [17], [18], [19], where between measurement updates a large number of consensus steps (theoretically, an infinite number of steps) is taken. *Combined* consensus+innovations algorithms are natural when a distributed network estimates a spatially varying random field defined at M spatial locations, say, for simplicity, a temperature field. The goal is to reconstruct at each and every sensor an accurate image of the *entire* spatial distribution of the M-dimensional field, assuming that at each time step each sensor makes a noisy measurement of the temperature at its single location. Without cooperation (no consensus step,) the processing of the successive temperature readings at each sensor (successive innovation steps) leads to a reliable estimate of the temperature at the sensor location–but provides no clue about the temperature distribution at the other $M-1$ locations. On the other hand, if sensors cooperate (consensus iterates,) but only process the initial measurement, as in traditional consensus, they converge to the average temperature across

Manuscript received August 29, 2008; revised February 05, 2011; accepted November 15, 2011. The work of Soummya Kar and José M. F. Moura was supported by NSF under grants CCF-1011903 and CCF-1018509, and by AFOSR grant FA95501010291. The work of Kavita Ramanan was supported by the NSF under grants DMS 0405343 and CMMI 0728064.

Soummya Kar and José M. F. Moura are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA 15213 (e-mail: soummyak@andrew.cmu.edu, moura@ece.cmu.edu, ph: (412) 268-6341, fax: (412) 268-3890.)

Kavita Ramanan is with the Division of Applied Mathematics, Brown University, Providence, RI 02912 (e-mail: Kavita_Ramanan@brown.edu.)

Communicated by M. Lops, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2012.219450

the field, not to an estimate of the M-dimensional temperature distribution. The distributed consensus+innovations algorithms that we introduce achieve both; each sensor converges to an estimate of the entire M-dimensional field by combining *consensus* and *processing* of the sensors measurements. Subsequent to this paper, analysis of detection consensus+innovations type algorithms is, e.g., in [20], [21].

Important questions that arise with consensus+innovations algorithms include: i) *convergence*: do the algorithms converge and if so in what sense; ii) *consensus*: do the agents reach a consensus on their field estimates; iii) *distributed versus centralized*: how good is the distributed field estimate at each sensor when compared with the *centralized* estimate obtained by a fusion center, in other words are the distributed estimate sequences consistent, and asymptotically unbiased, efficient, or normal; and iv) *rate of convergence*: what is the rate at which the distributed estimators converge. These questions are very distinct from the convergence issues considered in the "consensus only" literature.

We present three distributed consensus+innovations inference algorithms: $\mathcal{LU}$ for linear observation models (as when each sensor makes a noisy reading of the temperature at its location, see Section II-E;) and two algorithms, $\mathcal{NU}$ and $\mathcal{NLU}$, for nonlinear observation models (like in power grids when each sensor measures a phase differential through a sinusoidal modulation, see Section IV-D.) The paper introduces the conditions on the sensor observations model (separable estimability that we define) *and* on the communication network (connectedness on average) for the distributed estimates to converge. The *separable estimability*, akin to global observability, and *connectedness*, is an intuitively pleasing condition and in a sense minimal–distributed estimation cannot do better than (the optimal) centralized estimator, hence, the model better be (globally) observable (but not necessarily locally observable;) and if the sensors need to cooperate to assimilate the data collected by the distributed network, the network should be connected (on average,) or the sensors will work in isolation.

In contrast with other settings, e.g., *linear* distributed least-mean-square (LMS) approaches to parameter estimation, e.g, [22], [23], [24], [25], we study distributed estimation in the usual framework of linear or nonlinear observations of a (vector) parameter in noise[1], when the dimension, M_n, of the observation at each sensor n is $M_n \ll M$, and the parameter estimation model is locally unobservable, i.e., each individual sensor cannot recover the entire M-dimensional parameter from its M_n-dimensional observation, even if noiseless. Through cooperation (consensus) a (local) sensor estimator may converge to an estimate of the entire M-dimensional field, by simultaneously combining at each time i, its estimate, its observation (innovation), and the estimates received from the sensors with which it communicates. We show in the paper conditions under which this holds.

We extend this distributed estimation model to include sensor and link or communication channel failures, random communication protocols, and quantized communication. These conditions make the problem more realistic when a large number of agents are involved since inexpensive sensors are bounded to fail at random times, packet loss in wireless digital communications cause links to fail intermittently, agents can communicate asynchronously via a random protocol like gossip or one of its variants, and the agents may be resource constrained and have a limited bit budget for communication. We make no distributional assumptions on the sensors and link failures, they can be spatially correlated, [28], but are temporally uncorrelated[2]. We show that, under these broad conditions, the three *distributed* estimation algorithms, $\mathcal{LU}$, $\mathcal{NU}$, and $\mathcal{NLU}$, are consistent if the observation model is separable estimable (see Section III-A) and the network is connected on average.

Algorithms $\mathcal{LU}$, $\mathcal{NU}$, and $\mathcal{NLU}$: $\mathcal{LU}$ applies when the noisy observations are linear on the parameter. For the linear model, the separably estimable condition reduces to a rank condition on the global observability Grammian. $\mathcal{LU}$ combines at each time iteration the consensus term with the innovations associated with the new observation. Note that, in this algorithm, as well as with the other two nonlinear algorithms, the dimension of the local observation for sensor n, M_n, is much smaller than the dimension M of the field, i.e., $M_n \ll M$. The algorithm $\mathcal{NU}$ generalizes $\mathcal{LU}$ to nonlinear separably estimable models. It is very important to note that, in both algorithms, $\mathcal{LU}$ and $\mathcal{NU}$, the same asymptotically decaying to zero time-varying weight sequence is associated with the consensus and innovation updates; in other words, both the consensus and innovation terms of the algorithm exhibit the same decay rate. Because of this, it is enough to resort to stochastic approximation techniques to prove consistency, asymptotic unbiasedness, and asymptotic normality for both algorithms, $\mathcal{LU}$ and $\mathcal{NU}$. For a treatment of general distributed stochastic algorithms see [29], [30], [31], [32]. Beyond consistency, we characterize explicitly for the $\mathcal{LU}$ algorithm

[1] In this paper, we restrict attention to static parameter (fields). Time-varying parameters/signals are considered elsewhere, see, for example, [26],[27] for estimation/filtering of fading (non-stationary) parameters and general time-varying linear dynamical systems, respectively.

[2] This dynamic network is more general and subsumes the erasure network model, where the link failures are independent over space *and* time.

the asymptotic variance and compare it with the asymptotic variance of the centralized optimal scheme. For the $\mathcal{NU}$ algorithm, and general models, it is difficult to find explicitly a Lyapounov function (as needed by stochastic approximation). However, with weaker assumptions on the nonlinear observation model (Lipschitz continuity and certain growth properties,) we guarantee the existence of a Lyapounov function; hence, demonstrate asymptotic normality of the $\mathcal{NU}$ estimates, see Theorems 18 and 19 in Section III-C. These conditions are much easier to verify than guessing the form of a Lyapounov function. Also, in the proof of Theorem 18, we actually show how to use these conditions to determine a Lyapounov function explicitly, which can then be used to analyze convergence rates.

The third algorithm, $\mathcal{NLU}$, applies when the observation models are only continuous and not Lipschitz continuous. $\mathcal{NLU}$ is however a mixed time-scale algorithm, where the consensus time-scale dominates the innovations time-scale, and consists of unbiased perturbations (detailed explanation is provided in the paper.) Because of this mixed time-scales, the $\mathcal{NLU}$ algorithm does not fall under the purview of standard stochastic approximation theory, and to show its consistency requires an altogether different framework as developed in the paper, see Theorems 21 and 22, in Section IV.

The two algorithms $\mathcal{NLU}$ and $\mathcal{NU}$ represent different tradeoffs. We show consistency for $\mathcal{NLU}$ under weaker assumptions (observation model continuity) than for $\mathcal{NU}$ (Lipschitz continuity plus growth conditions.) On the other hand, when these more stringent conditions hold, $\mathcal{NU}$ provides convergence rate guarantees and asymptotic normality; these follow from standard stochastic approximation theory that apply to $\mathcal{NU}$ but not to $\mathcal{NLU}$.

Brief comment on the literature. We contrast our work with relevant recent literature on distributed estimation. Papers [15], [17], [33], [18] study estimation in *static* networks, where either the sensors take a single snapshot of the field at the start and then initiate distributed consensus protocols (or more generally distributed optimization, as in [17]) to fuse the initial estimates, or the observation rate of the sensors is assumed to be much slower than the inter-sensor communicate rate, thus permitting a separation of the two time-scales. On the contrary, our consensus+innovations algorithm combines fusion (consensus) and observation (innovation) updates in the same iteration. The network is *dynamic* with channel failures, the protocols are *random*, and the sensors fail. Unlike [15], [17], [33], [18], our approach does not require distributional assumptions on the observation noise, and we make explicit the structural assumptions on the observation model (separable estimability) and network connectivity needed to guarantee consistent parameter estimates at every sensor. These structural assumptions are quite weak and are necessary for centralized estimators to obtain consistency.

There is considerable work in *linear* distributed least-mean-square (LMS) approaches to parameter estimation in *static* networks, e.g., [22], [23], [24], [25]. While LMS is also a consensus+innovation type algorithm, we show how our linear algorithm $\mathcal{LU}$ and LMS are quite distinct, with a very different setup and goal. In LMS, for example, for channel estimation or channel equalization, [34], in adaptive filtering, [35], or in system identification, see [36], the observations $z_n(i)$ are the output of a noisy finite impulse response channel (or a linear system to be identified) excited by a random input sequence $u(i)$ (these random input sequences are the regressors.) The unknown channel impulse response θ is to be estimated by a stochastic gradient type algorithm that has available (in the channel estimation or training phase) *both* the random inputs *and* the regressors. In contrast, in the distributed estimation problem we study, for example, for the $\mathcal{LU}$, the observations at sensor n and time i are

$$\mathbf{z}_n(i) = H_n(i)\theta^* + \zeta_n(i) \tag{1}$$

For example, in (1), the observation matrices $H_n(i)$ could be of the form,

$$H_n(i) = \frac{1}{p}\delta_n(i)\overline{H}_n$$

where $\delta_n(i)$ is a zero-one Bernoulli variable to account for sensor failures, $p > 0$ denotes the sensing probability, and the mean value $\overline{H}_n$ models the normal operation of the sensor, e.g., measuring the local temperature, or an average of local temperatures. Equation (1) is the usual model in parameter estimation or waveform filtering, while (I-A) extends this model in a significant way to random intermittent measurements. In our distributed estimation algorithms, we do *not* know the random observation matrix $H_n(i)$ (only its first and second moment), while in the LMS where they play the role of the regressors, they are usually known to the LMS algorithm.

We contrast further our *linear distributed* $\mathcal{LU}$ with linear distributed LMS. References [22], [24], [25] use non-decaying combining weights that lead to a residual tracking error; under appropriate assumptions, these algorithms can be adapted to certain time-varying tracking scenarios; we consider time varying processes in other work, [26], [27]. Reference [23] considers decaying weight sequences as we do in $\mathcal{LU}$, thereby establishing also $\mathcal{L}_2$ convergence to the desired parameter value. All these works deal with distributed *linear* problems, while our work emphasizes distributed estimators for linear and nonlinear sensor observation models and establishes their convergence properties. We present the necessary (minimal) structural conditions that the distributed sensing model (given) and the inter-sensor communication network should satisfy to guarantee the existence of *successful* distributed estimators. Also, apart from treating generic separably estimable nonlinear observation models, in the linear case, our algorithms $\mathcal{NU}$ and $\mathcal{LU}$ lead to asymptotic normality in addition to consistency and asymptotic unbiasedness in random time-varying networks with quantized inter-sensor communication and sensor failures.

Remark. We noted that the $\mathcal{NLU}$ algorithm is mixed time scale; this means a stochastic algorithm where two potentials act in the same update step with different weight or gain sequences. This should not be confused with (centralized) stochastic algorithms with coupling (see [37]), where a quickly switching parameter influences the relatively slower dynamics of another state, leading to *averaged* dynamics. We note further in this context that [38] (and references therein) develops methods to analyze mixed time scale (centralized) algorithms in the context of simulated annealing. In [38], the role of our innovation (new observation) potential is played by a martingale difference term. However, in our study, the innovation is not a martingale difference process, and a key step in the analysis is to derive pathwise strong approximation results to characterize the rate at which the innovation process converges to a martingale difference process.

We briefly comment on the organization of the remaining of the paper. The rest of this section introduces notation and preliminaries to be adopted throughout the paper. To motivate the generic nonlinear problem, we study the linear case (algorithm $\mathcal{LU}$) in Section II. Section III studies the generic separably estimable models and the algorithm $\mathcal{NU}$, whereas algorithm $\mathcal{NLU}$ is presented in Section IV. Finally, Section V concludes the paper.

B. Notation

For completeness, this subsection sets notation and presents preliminaries on algebraic graph theory, matrices, and dithered quantization to be used in the sequel.

Preliminaries: We adopt the following notation. $\mathbb{R}^k$: the k-dimensional Euclidean space; I_k: $k\times k$ identity matrix; $\mathbf{1}_k, \mathbf{0}_k$: column vector of ones and zeros in $\mathbb{R}^k$, respectively; $P_k=\frac{1}{k}\mathbf{1}_k\mathbf{1}_k^T$: the rank one $k\times k$ projector, whose only non-zero eigenvalue is one, and the corresponding normalized eigenvector is $\left(1/\sqrt{k}\right)\mathbf{1}_k$; $\|\cdot\|$: the standard Euclidean 2-norm when applied to a vector and the induced 2-norm when applied to matrices, which is equivalent to the matrix spectral radius for symmetric matrices; $\theta\in\mathcal{U}\subset\mathbb{R}^M$: the parameter to be estimated; θ^*: the true (but unknown) value of the parameter θ; $\mathbf{x}_n(i)\in\mathbb{R}^M$: the estimate of θ^* at time i at sensor n–without loss of generality (wlog), the initial estimate, $\mathbf{x}_n(0)$, at time 0 at sensor n is a non-random quantity; $(\Omega,\mathcal{F})$: common measurable space where all the random objects are defined; $\mathbb{P}_{\theta^*}[\cdot]$ and $\mathbb{E}_{\theta^*}[\cdot]$: the probability and expectation operators when the true (but unknown) parameter value is θ^*–when the context is clear, we abuse notation by dropping the subscript. All inequalities involving random variables are to be interpreted a.s. (almost surely.)

Spectral graph theory: For the *undirected* graph $G=(V,E)$, $V=[1\cdots N]$ is the set of nodes or vertices, $|V|=N$, and E is the set of edges. The unordered pair $(n,l)\in E$ if there exists an edge between nodes n and l. The graph G is simple if devoid of self-loops and multiple edges and connected if there exists a path[3] between each pair of nodes. The neighborhood of node n is $\Omega_n=\{l\in V\,|\,(n,l)\in E\}$. The degree $d_n=|\Omega_n|$ of node n is the number of edges with n as one end point, and $D=\operatorname{diag}(d_1\cdots d_N)$ is the degree matrix, the diagonal matrix with diagonal entries the degrees d_n. 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. The graph Laplacian matrix, L, is $L=D-A$; it is a a positive semidefinite matrix whose eigenvalues can be ordered as $0=\lambda_1(L)\leq\lambda_2(L)\leq\cdots\leq\lambda_N(L)$. The smallest eigenvalue $\lambda_1(l)$ is zero, with $\left(1/\sqrt{N}\right)\mathbf{1}_N$ being the corresponding normalized eigenvector. 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 [39], [40], [41] for detailed treatment of graphs and their spectral theory.

Kronecker product: Since we are dealing with vector parameters, most of the matrix manipulations will involve Kronecker products. For example, the Kronecker product of the $N\times N$ matrix L and I_M will be an $NM\times NM$ matrix, denoted by $L\otimes I_M$. We will deal often with matrices of the form $C=[I_{NM}-bL\otimes I_M-aI_{NM}-P_N\otimes I_M]$. It follows from the properties of Kronecker products and of the matrices L and P_N that the eigenvalues of the matrix C are $-a$ and $1-b\lambda_i(L)-a$, $2\leq i\leq N$, each being repeated M times.

[3] A path between nodes n and l of length m is a sequence $(n=i_0,i_1,\cdots,i_m=l)$ of vertices, such that $(i_k,i_{k+1})\in E\,\forall\, 0\leq k\leq m-1$.

We now review results from statistical quantization theory.

Dithered quantization: We assume that all sensors are equipped with identical uniform, dithered quantizers $\mathbf{q}(\cdot) : \mathbb{R}^M \to \mathcal{Q}^M$ applied componentwise, with countable alphabet $\mathcal{Q}^M = \left\{ [k_1\Delta, \cdots, k_M\Delta]^T \;\middle|\; k_i \in \mathbb{Z},\ \forall i \right\}$. We assume the dither satisfies the Schuchman conditions (see [42], [43], [44], [45],) so that the error sequence for subtractively dithered systems ([43]) $\{\varepsilon(i)\}_{i\geq 0}$

$$\varepsilon(i) = q(y(i) + \nu(i)) - (y(i) + \nu(i)) \tag{2}$$

is an i.i.d. sequence of uniformly distributed random variables on $[-\Delta/2, \Delta/2)$, which is independent of the input sequence $\{y(i)\}_{i\geq 0}$. In (2), the dither sequence $\{\nu(i)\}_{i\geq 0}$ is i.i.d. uniformly distributed random variables on $[-\Delta/2, \Delta/2)$, independent of the input sequence $\{y(i)\}_{i\geq 0}$; we refer to [46] where we use this model and make further relevant comments.

Consistency and asymptotic unbiasedness: We recall standard definitions from sequential estimation theory (see, for example, [47]).

Definition 1 (Consistency) : A sequence of estimates $\{\mathbf{x}^\bullet(i)\}_{i\geq 0}$ is called consistent if

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty} \mathbf{x}^\bullet(i) = \theta^*\right] = 1,\ \forall \theta^* \in \mathcal{U}$$

or, in other words, if the estimate sequence converges a.s. to the true parameter value. The above definition of consistency is also called strong consistency. When the convergence is in probability, we get weak consistency. In this paper, we use the term consistency to mean strong consistency, which implies weak consistency.

Definition 2 (Asymptotic Unbiasedness) :

A sequence of estimates $\{\mathbf{x}^\bullet(i)\}_{i\geq 0}$ is called asymptotically unbiased if

$$\lim_{i\to\infty} \mathbb{E}_{\theta^*}\left[\mathbf{x}^\bullet(i)\right] = \theta^*, \quad \forall \theta^* \in \mathcal{U}$$

II. Distributed Linear Parameter Estimation: Algorithm $\mathcal{LU}$

In this section, we consider the algorithm $\mathcal{LU}$ for *distributed* parameter estimation when the observation model is linear. This problem motivates the generic *separably estimable* nonlinear observation models considered in Sections III and IV. Section II-A sets up the distributed linear estimation problem and presents the algorithm $\mathcal{LU}$. Section II-B establishes the consistency and asymptotic unbiasedness of the $\mathcal{LU}$ algorithm, where we show that, under the $\mathcal{LU}$ algorithm, all sensors converge a.s. to the true parameter value, θ^*. Convergence rate analysis (asymptotic normality) is carried out in Section II-C, while Section II-E illustrates $\mathcal{LU}$ with an example.

A. Problem Formulation: Algorithm $\mathcal{LU}$

Let $\theta^* \in \mathbb{R}^M$ be an M-dimensional parameter to be estimated by a network of N sensors. Sensor n makes the observations:

$$\mathbf{z}_n(i) = H_n(i)\theta^* + \zeta_n(i)\ \in \mathbb{R}^{M_n} \tag{3}$$

Each sensor observes only a subset of M_n of the components of θ^*, or M_n linear combinations of a few components of θ^*, with $M_n \ll M$. We make the following assumptions.

(A.1)Observation Noise: The observation noise process, $\left\{\zeta(i) = \left[\zeta_1^T(i), \cdots, \zeta_N^T(i)\right]^T\right\}_{i\geq 0}$ is i.i.d. zero mean, with finite second moment. In particular, the observation noise covariance is bounded and independent of i

$$\mathbb{E}\left[\zeta(i)\zeta^T(j)\right] = S_\zeta \delta_{ij},\ \forall i, j \geq 0 \tag{4}$$

where the Kronecker symbol $\delta_{ij} = 1$ if $i = j$ and zero otherwise. Note that the observation noises at different sensors may be correlated during a particular iteration. Eqn. (4) states only temporal independence. The spatial correlation of the observation noise makes our model applicable to practical sensor network problems, for instance, for distributed target localization, where the observation noise is generally correlated across sensors.

(A.2)Sensor Failures: The observation matrices, $\{[H_1(i), \cdots, H_N(i)]\}_{i\geq 0}$, form an i.i.d. sequence with mean $\left[\overline{H}_1, \cdots, \overline{H}_N\right]$ and finite second moment. In particular, we have

$$H_n(i) = \overline{H}_n + \widetilde{H}_n(i),\ \forall i, n$$

where, $\overline{H}_n = \mathbb{E}\left[H_n(i)\right], \forall i, n$ and the sequence $\left\{\left[\widetilde{H}_1(i), \cdots, \widetilde{H}_N(i)\right]\right\}_{i\geq 0}$ is zero mean i.i.d. with finite second moment. Here, also, we require only temporal independence of the observation matrices, but allow them to be spatially correlated. For example, $H_n(i) = \delta_n(i)\overline{H}_n$, with $\delta_n(i)$ an iid sequence of Bernoulli variables modeling intermittent sensor failures.

Remark 3 The $\mathcal{LU}$ update does not use the instantaneous observation matrices, $H_n(i)$, only their ensemble averages. This is a distinction between $\mathcal{LU}$ and LMS. LMS (for example, in adaptive filtering) assumes the random matrices $H_n(i)$ are, together with the observations, also available (see Chapter 4 of [48]). In parameter estimation, the $H_n(i)$ model sensor failures and $\mathcal{LU}$ has no control over their instantiations. Hence, while in LMS it may be reasonable to use the instantaneous values of the random regressors $H_n(i)$, in the setting we consider, the instantaneous realizations of the observation matrices are not available.

(A.3)Mean Connectedness, Link Failures, and Random Protocols: The graph Laplacians

$$L(i) = \overline{L} + \widetilde{L}(i),\ \forall i \geq 0$$

are a sequence of i.i.d. matrices with mean $\overline{L} = \mathbb{E}\left[L(i)\right]$. We make no distributional assumptions on the $\{L(i)\}$. Although independent at different times, during the same iteration, the link failures can be spatially dependent, i.e., correlated. This is more general and subsumes the erasure network model, where the link failures are independent over space *and* time. Wireless sensor networks motivate this model since interference among the wireless communication channels correlates the link failures over space, while, over time, it is still reasonable to assume that the channels are memoryless or independent. Connectedness of the graph is an important issue. The random instantiations $G(i)$ of the graph need not be connected; in fact, all these instantiations may be disconnected. We only require the graph to be connected on *average*. This is captured by requiring that $\lambda_2\left(\overline{L}\right) > 0$, enabling us to capture a broad class of asynchronous communication models; for example, the random asynchronous gossip protocol analyzed in [49] satisfies $\lambda_2\left(\overline{L}\right) > 0$ and hence falls under this framework.

(A.4) Independence: The sequences $\{L(i)\}_{i\geq 0}$, $\{\zeta_n(i)\}_{1\leq n\leq N,\ i\geq 0}$, $\{H_n(i)\}_{1\leq n\leq N, i\geq 0}$, $\{\nu_{nl}^m(i)\}$ are mutually independent.

We introduce the distributed observability condition required for convergence of the $\mathcal{LU}$ linear estimation algorithm.

Definition 4(Distributed observability) The observation system (3) is *distributedly observable* if the matrix G is full rank

$$G = \sum_{n=1}^{N} \overline{H}_n^T \overline{H}_n \tag{5}$$

This distributed observability extends the observability condition for a centralized estimator that is needed to get a consistent estimate of the parameter θ^*. We note that the information available to the n-th sensor at any time i about the corresponding observation matrix is just the mean $\overline{H}_n$, and *not* the random $H_n(i)$. Hence, the state update equation uses only the $\overline{H}_n$'s, as given in (6) below.

(A.5) Observability: The distributed observation system (3) is *distributedly observable* in the sense of definition 4.

Algorithm $\mathcal{LU}$: We consider now the algorithm $\mathcal{LU}$ for distributed parameter estimation in the linear observation model (3). Starting from some initial deterministic estimate of the parameters[4], $\mathbf{x}_n(0) \in \mathbb{R}^M$, each sensor n

[4] The initial states may be random, we assume deterministic for notational simplicity.

generates a sequence of estimates, $\{\mathbf{x}_n(i)\}_{i\geq 0}$ by the following distributed iterative algorithm:

$$\mathbf{x}_n(i+1) = \mathbf{x}_n(i) - \alpha(i)\left[b\sum_{l\in\Omega_n(i)}\left(\mathbf{x}_n(i) - \mathbf{q}\left(\mathbf{x}_l(i)\right.\right.\right. \tag{6}$$

$$\left.\left.\left. +\nu_{nl}(i)\right)\right) - \overline{H}_n^T\left(\mathbf{z}_n(i) - \overline{H}_n\mathbf{x}_n(i)\right)\right]$$

where $\{\mathbf{q}(\mathbf{x}_l(i)+\nu_{nl}(i))\}_{l\in\Omega_n(i)}$ is the dithered quantized exchanged data. In (6), the sequence of weights $\{\alpha(i)\}$ satisfies the persistence condition **B5** given in the Appendix A; $b>0$ is a constant and $\{\alpha(i)\}_{i\geq 0}$ is a sequence of weights with properties to be defined below. The algorithm (6) is distributed because for sensor n it involves only the data from the sensors in its neighborhood $\Omega_n(i)$. Using (2), the state update can be written as

$$\mathbf{x}_n(i+1) = \mathbf{x}_n(i) - \alpha(i)\left[b\sum_{l\in\Omega_n(i)}\left(\mathbf{x}_n(i) - \mathbf{x}_l(i)\right)\right.$$

$$\left. - \overline{H}_n^T\left(\mathbf{z}_n(i) - \overline{H}_n\mathbf{x}_n(i)\right) - b\nu_{nl}(i) - b\varepsilon_{nl}(i)\right] \tag{7}$$

We rewrite (7) in compact form. Define the random vectors, $\boldsymbol{\Upsilon}(i)$ and $\boldsymbol{\Psi}(i)\in\mathbb{R}^{NM}$ with vector components

$$\boldsymbol{\Upsilon}_n(i) = -\sum_{l\in\Omega_n(i)}\nu_{nl}(i) \text{ and } \boldsymbol{\Psi}_n(i) = -\sum_{l\in\Omega_n(i)}\varepsilon_{nl}(i) \tag{8}$$

It follows from the Schuchman conditions on the dither, see Section I-B and [46], that

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{\Upsilon}(i)\right] &= \mathbb{E}\left[\boldsymbol{\Psi}(i)\right] = \mathbf{0},\ \forall i\\ \sup_i\mathbb{E}\left[\|\boldsymbol{\Upsilon}(i)\|^2\right] &= \sup_i\mathbb{E}\left[\|\boldsymbol{\Psi}(i)\|^2\right]\\ &\leq \frac{N(N-1)M\Delta^2}{12}\end{aligned} \tag{9}$$

from which we then have

$$\begin{aligned}\sup_i\mathbb{E}\left[\|\boldsymbol{\Upsilon}(i)+\boldsymbol{\Psi}(i)\|^2\right] &\leq 2\sup_i\mathbb{E}\left[\|\boldsymbol{\Upsilon}(i)\|^2\right] +\\ +2\sup_i\mathbb{E}\left[\|\boldsymbol{\Psi}(i)\|^2\right] &\leq \frac{N(N-1)M\Delta^2}{3} = \eta_q\end{aligned} \tag{10}$$

The iterations in (6) can be written in compact form. Stack all sensors state estimates in a long state vector estimate $\mathbf{x}(i) = \left[\mathbf{x}_1^T(i)\cdots\mathbf{x}_N^T(i)\right]^T$ and define the matrices

$$\begin{aligned}\overline{D}_{\overline{H}} &= \text{diag}\left[\overline{H}_1^T\cdots\overline{H}_N^T\right]\\ D_{\overline{H}} &= \overline{D}_{\overline{H}}\overline{D}_{\overline{H}}^T = \text{diag}\left[\overline{H}_1^T\overline{H}_1\cdots\overline{H}_N^T\overline{H}_N\right]\end{aligned}$$

Then, the compact vector form of the $\mathcal{LU}$ algorithm is

$$\begin{aligned}\mathbf{x}(i+1) = \mathbf{x}(i) - \alpha(i)&\left[b(L(i)\otimes I_M)\mathbf{x}(i)\overline{D}_{\overline{H}}^T\right.\\ &\left.-\overline{D}_{\overline{H}}\left(\mathbf{z}(i) - \overline{D}_{\overline{H}}^T\mathbf{x}(i)\right) + b\boldsymbol{\Upsilon}(i) + b\boldsymbol{\Psi}(i)\right]\end{aligned} \tag{11}$$

In the $\mathcal{LU}$ algorithm (11), the covariance matrix of the noise is defined as

$$S_q = \mathbb{E}\left[\left(\boldsymbol{\Upsilon}(i)+\boldsymbol{\Psi}(i)\right)\left(\boldsymbol{\Upsilon}(i)+\boldsymbol{\Psi}(i)\right)^T\right] \tag{12}$$

Markov: We characterize the state vector estimate $\mathbf{x}(i)$. Consider the filtration, $\{\mathcal{F}_i^{\mathbf{x}}\}_{i\geq 0}$, given by

$$\mathcal{F}_i^{\mathbf{x}} = \sigma\left(\mathbf{x}(0), \{L(j), \mathbf{z}(j), \boldsymbol{\Upsilon}(j), \boldsymbol{\Psi}(j)\}_{0\leq j<i}\right) \tag{13}$$

From **(A1)–(A4)**, $L(i), \mathbf{z}(i)$, $\boldsymbol{\Upsilon}(i)$, $\boldsymbol{\Psi}(i)$ are independent of $\mathcal{F}_i^{\mathbf{x}}$; so, $\{\mathbf{x}(i), \mathcal{F}_i^{\mathbf{x}}\}_{i\geq 0}$ is a Markov process.

B. Consistency of $\mathcal{LU}$

We consider consistency and asymptotic unbiasedness.

Lemma 5 Consider $\mathcal{LU}$ under Assumptions **(A.1)-(A.5)**. Then, $\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]$ is symmetric positive definite.

Proof: Symmetricity is obvious. It also follows from the properties of Laplacian matrices and the structure of $D_{\overline{H}}$ that these matrices are positive semidefinite. Then the matrix $\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]$ is positive semidefinite, being the sum of two positive semidefinite matrices. To prove positive definiteness, assume, on the contrary, that the matrix $\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]$ is not positive definite. Then, there exists, $\mathbf{x}\in\mathbb{R}^{NM}$, such that $\mathbf{x}\neq\mathbf{0}$ and

$$\mathbf{x}^T\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]\mathbf{x} = \mathbf{0}$$

From the positive semidefiniteness of $\overline{L}\otimes I_M$ and $D_{\overline{H}}$, and the fact that $b>0$, it follows

$$\mathbf{x}^T\left[\overline{L}\otimes I_M\right]\mathbf{x} = 0,\ \ \mathbf{x}^T D_{\overline{H}}\mathbf{x} = 0 \tag{14}$$

Partition $\mathbf{x}$ as $\mathbf{x} = \left[\mathbf{x}_1^T\cdots\mathbf{x}_N^T\right]^T, \mathbf{x}_n\in\mathbb{R}^M, \forall 1\leq n\leq N$. It follows from the properties of Laplacian matrices and the fact that $\lambda_2(\overline{L})>0$, that (14) holds *iff*

$$\mathbf{x}_n = \mathbf{a},\ \forall n \tag{15}$$

where $\mathbf{a}\in\mathbb{R}^M$, and $\mathbf{a}\neq\mathbf{0}$. Also, (14) implies

$$\sum_{n=1}^{N}\mathbf{x}_n^T\overline{H}_n^T\overline{H}_n\mathbf{x}_n = 0 \tag{16}$$

Let G be as in (5). Equations (16) and (15) imply

$$\mathbf{a}^T G\mathbf{a} = 0,$$

a contradiction, since $G>0$ by Assumption **(A.5)** and $\mathbf{a}\neq\mathbf{0}$. Thus, $\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]>0$. ■

Theorem 6 ($\mathcal{LU}$: Asymptotic unbiasedness) Let the $\mathcal{LU}$ algorithm under **(A.1)-(A.5)**. Then, $\{\mathbf{x}_n(i)\}_{i\geq 0}$, at sensor n is asymptotically unbiased

$$\lim_{i\to\infty}\mathbb{E}\left[\mathbf{x}_n(i)\right] = \theta^*,\ \ 1\leq n\leq N$$

Proof: Taking expectations on both sides of (11) and by the independence assumption **(A.4)**,

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}(i+1)\right] = \mathbb{E}\left[\mathbf{x}(i)\right] - \alpha(i)\left[b\left(\overline{L}\otimes I_M\right)\mathbb{E}\left[\mathbf{x}(i)\right] + \right.\\ \left. + D_{\overline{H}}\mathbb{E}\left[\mathbf{x}(i)\right] - \overline{D}_{\overline{H}}\mathbb{E}\left[\mathbf{z}(i)\right]\right]\end{aligned} \tag{17}$$

Subtracting $\mathbf{1}_N\otimes\theta^*$ from both sides of (17), noting that

$$\begin{aligned}\left(\overline{L}\otimes I_M\right)\left(\mathbf{1}_N\otimes\theta^*\right) = \mathbf{0},\\ \overline{D}_{\overline{H}}\mathbb{E}\left[\mathbf{z}(i)\right] = D_{\overline{H}}\left(\mathbf{1}_N\otimes\theta^*\right)\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}(i+1)\right] - \mathbf{1}_N\otimes\theta^* = \left[I_{NM} - \alpha(i)\left(b\overline{L}\otimes I_M + \right.\right.\\ \left.\left. +\ D_{\overline{H}}\right)\right]\left[\mathbb{E}\left[\mathbf{x}(i)\right] - \mathbf{1}_N\otimes\theta^*\right]\end{aligned} \tag{18}$$

Let $\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)$, $\lambda_{\max}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)$ be the smallest and largest eigenvalues of the positive definite matrix $\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]$ (Lemma 5.) Since $\alpha(i)\to 0$ (Assumption **(B.5)**, Appendix A),

$$\exists i_0 \ni:\ \alpha(i_0) \le \frac{1}{\lambda_{\max}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)},\ \forall i \ge i_0 \tag{19}$$

Continuing the recursion in (18), we have, for $i > i_0$,

$$\mathbb{E}\left[\mathbf{x}(i)\right] - \mathbf{1}_N\otimes\theta^* = \left(\prod_{j=i_0}^{i-1}\left[I_{NM} - \alpha(j)\left(b\overline{L}\otimes I_M + \right.\right.\right.$$
$$\left.\left.\left. + \ D_{\overline{H}}\right)\right]\right)\left[\mathbb{E}\left[\mathbf{x}(i_0)\right] - \mathbf{1}_N\otimes\theta^*\right] \tag{20}$$

Eqn. (20) implies

$$\left\|\mathbb{E}\left[\mathbf{x}(i)\right] - \mathbf{1}_N\otimes\theta^*\right\| \le \left(\prod_{j=i_0}^{i-1}\left\|I_{NM} - \alpha(j)\right.\right.$$
$$\left.\left.\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)\right\|\right)\left\|\mathbb{E}\left[\mathbf{x}(i_0)\right] - \mathbf{1}_N\otimes\theta^*\right\|,\quad i > i_0$$

It follows from (19)

$$\left\|I_{NM} - \alpha(j)\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)\right\| =$$
$$1 - \alpha(j)\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right),\ j \ge i_0$$

Eqns. (II-B,II-B) now give for $i > i_0$

$$\left\|\mathbb{E}\left[\mathbf{x}(i)\right] - \mathbf{1}_N\otimes\theta^*\right\| \le \left(\prod_{j=i_0}^{i-1}\left(1 - \alpha(j)\right.\right.$$
$$\left.\left.\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)\right)\right)\left\|\mathbb{E}\left[\mathbf{x}(i_0)\right] - \mathbf{1}_N\otimes\theta^*\right\|,$$

Finally, from the inequality $1 - a \le e^{-a}$, $0 \le a \le 1$, get

$$\left\|\mathbb{E}\left[\mathbf{x}(i)\right] - \mathbf{1}_N\otimes\theta^*\right\| \le e^{-\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)\sum_{j=i_0}^{i-1}\alpha(j)}$$
$$\left\|\mathbb{E}\left[\mathbf{x}\left(i_0\right)\right] - \mathbf{1}_N\otimes\theta^*\right\|,\quad i > i_0$$

Since, $\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right) > 0$ and the weight sequence sums to infinity, the theorem follows since

$$\lim_{i\to\infty}\left\|\mathbb{E}\left[\mathbf{x}(i)\right] - \mathbf{1}_N\otimes\theta^*\right\| = 0$$

■

Before proceeding to Theorems 7 and 10 establishing the consistency and asymptotic normality of the $\mathcal{LU}$, the reader may refer to Appendix A, where useful results on stochastic approximation are discussed.

Theorem 7 ($\mathcal{LU}$: Consistency) Consider $\mathcal{LU}$ under (**A.1**)–(**A.5**). Then, the estimate sequence $\left\{\mathbf{x}_n(i)\right\}_{i\ge 0}$ at sensor n is consistent

$$\mathbb{P}\left[\lim_{i\to\infty}\mathbf{x}_n(i) = \theta^*,\ \forall n\right] = 1$$

Proof: The proof follows by showing that $\left\{\mathbf{x}(i)\right\}_{i\ge 0}$ satisfies the Assumptions **(B.1)-(B.5)** of Theorem 29 (Appendix A). Recall the filtration, $\left\{\mathcal{F}_i^{\mathbf{x}}\right\}_{i\ge 0}$, in (13). Rewrite (11) by adding and subtracting the vector $\mathbf{1}_N\otimes\theta^*$

and noting that

$$\begin{aligned}
&\left(\overline{L}\otimes I_M\right)\left(\mathbf{1}_N\otimes\theta^*\right)=\mathbf{0}\\
&\mathbf{x}(i+1)=\mathbf{x}(i)-\alpha(i)\Bigg[b\left(\overline{L}\otimes I_M\right)(\mathbf{x}(i)-\\
&\qquad -\mathbf{1}_N\otimes\theta^*)+b\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x}(i)+\\
&\quad +D_{\overline{H}}\left(\mathbf{x}(i)-\mathbf{1}_N\otimes\theta^*\right)-\overline{D}_{\overline{H}}\left(\mathbf{z}(i)-\overline{D}_{\overline{H}}^{T}\mathbf{1}_N\otimes\theta^*\right)+\\
&\qquad\qquad +\ b\boldsymbol{\Upsilon}(i)+b\boldsymbol{\Psi}(i)\Bigg]
\end{aligned}\tag{21}$$

In the notation of Theorem 29, Appendix A, let $R\left(\mathbf{x}\right)$ and $\Gamma\left(i+1,\mathbf{x},\omega\right)$ as in (22) and (23) below and rewrite (21)

$$\mathbf{x}(i+1)=\mathbf{x}(i)+\alpha(i)\left[R(\mathbf{x}(i))+\Gamma\left(i+1,\mathbf{x}(i),\omega\right)\right]$$

$$R\left(\mathbf{x}\right)=-\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\tag{22}$$

$$\begin{aligned}
\Gamma\left(i+1,\mathbf{x},\omega\right)=&-\Bigg[b\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x}-\\
&-\overline{D}_{\overline{H}}\left(\mathbf{z}(i)-\overline{D}_{\overline{H}}^{T}\mathbf{1}_N\otimes\theta^*\right)+b\boldsymbol{\Upsilon}(i)+b\boldsymbol{\Psi}(i)\Bigg]
\end{aligned}\tag{23}$$

Under the Assumptions **(A.1)-(A.5)**, for fixed $i+1$, the random family, $\left\{\Gamma\left(i+1,\mathbf{x},\omega\right)\right\}_{\mathbf{x}\in\mathbb{R}^{NM}}$, is $\mathcal{F}_{i+1}^{\mathbf{x}}$ measurable, zero-mean and independent of $\mathcal{F}_{i}^{\mathbf{x}}$. Hence, the assumptions **(B.1)-(B.2)** of Theorem 29 are satisfied.

We now show the existence of a stochastic potential function $V(\cdot)$ satisfying the remaining Assumptions **(B.3)-(B.4)** of Theorem 29. To this end, define

$$\begin{aligned}
V\left(\mathbf{x}\right)=\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T&\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]\\
&\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)
\end{aligned}$$

Clearly, $V\left(\mathbf{x}\right)\in\mathbb{C}_2$ with bounded second order partial derivatives. It follows from the positive definiteness of $\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]$ (Lemma 5), that

$$V\left(\mathbf{1}_N\otimes\theta^*\right)=0,\quad V\left(\mathbf{x}\right)>0,\ \mathbf{x}\neq\mathbf{1}_N\otimes\theta^*$$

Since the matrix $\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]$ is positive definite, the matrix $\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]^2$ is also positive definite and hence, there exists a constant $c_1>0$, such that

$$\begin{aligned}
\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]^2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\geq\\
\geq c_1\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|^2,\quad\forall\mathbf{x}\in\mathbb{R}^{NM}
\end{aligned}$$

It then follows that

$$\begin{aligned}
&\sup_{\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|>\epsilon}\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)=\\
&-2\inf_{\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|>\epsilon}\Big\{\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]^2\\
&\qquad\qquad\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\Big\}\\
&\leq-2\inf_{\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|>\epsilon}c_1\left\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right\|^2\\
&\leq-2c_1\epsilon^2<0
\end{aligned}$$

Thus, Assumption **(B.3)** is satisfied. From (22)

$$\begin{aligned}&\|R\left(\mathbf{x}\right)\|^2=\\&=\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]^2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\\&=-\frac{1}{2}\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)\end{aligned}\tag{24}$$

From (23) and the independence Assumption **(A.4)**

$$\begin{aligned}&\mathbb{E}\left[\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2\right]=\\&=\mathbb{E}\left[\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left(b\widetilde{L}(i)\otimes I_M\right)^2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\right]\\&\qquad+\mathbb{E}\left[\left\|\overline{D}_{\overline{H}}\left(\mathbf{z}(i)-\overline{D}_{\overline{H}}^T\mathbf{1}_N\otimes\theta^*\right)\right\|^2\right]+\\&\qquad+b^2\mathbb{E}\left[\|\boldsymbol{\Upsilon}(i)+\boldsymbol{\Psi}(i)\|^2\right]\end{aligned}\tag{25}$$

Since the random matrix $\widetilde{L}(i)$ takes values in a finite set, there exists a constant $c_2>0$, such that, $\forall\mathbf{x}\in\mathbb{R}^{NM}$,

$$\begin{aligned}&\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left(b\widetilde{L}(i)\otimes I_M\right)^2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\le\\&\qquad\le c_2\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|^2\end{aligned}\tag{26}$$

Again, since $\left(b\overline{L}\otimes I_M+D_{\overline{H}}\right)$ is positive definite, there exists a constant $c_3>0$, such that, $\forall\mathbf{x}\in\mathbb{R}^{NM}$,

$$\begin{aligned}&\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\ge\\&\qquad\ge c_3\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|^2.\end{aligned}\tag{27}$$

We then have from (26)-(27)

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left(b\widetilde{L}(i)\otimes I_M\right)^2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\right]\le\\&\le\frac{c_2}{c_3}\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[b\overline{L}\otimes I_M+D_{\overline{H}}\right]\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\\&=c_4V\left(\mathbf{x}\right)\end{aligned}$$

for some constant $c_4=\frac{c_2}{c_3}>0$. The term

$$\mathbb{E}\left[\left\|\overline{D}_{\overline{H}}\mathbf{z}(i)-D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right\|^2\right]+b^2\mathbb{E}\left[\|\boldsymbol{\Upsilon}(i)+\boldsymbol{\Psi}(i)\|^2\right]$$

is bounded by a finite constant $c_5>0$, as it follows from Assumptions **(A.1)-(A.5)**. We then have from (24)-(25)

$$\begin{aligned}&\|R\left(\mathbf{x}\right)\|^2+\mathbb{E}\left[\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2\right]\le\\&\le-\frac{1}{2}\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)+c_4V\left(\mathbf{x}\right)+c_5\le\\&\le c_6\left(1+V\left(\mathbf{x}\right)\right)-\frac{1}{2}\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)\end{aligned}$$

where $c_6=\max\left(c_4,c_5\right)>0$. This verifies Assumption **(B.4)** of Theorem 29. Assumption **(B.5)** is satisfied by the choice of $\{\alpha(i)\}_{i\ge0}$. It then follows that the process $\{\mathbf{x}(i)\}_{i\ge0}$ converges a.s. to $\mathbf{1}_N\otimes\theta^*$. In other words,

$$\mathbb{P}[\lim_{i\to\infty}\mathbf{x}_n(i)=\theta^*,\ \forall n]=1$$

which establishes consistency of $\mathcal{LU}$. ■

The proof above can be modified to show $\mathcal{L}_2$ convergence of the sensor estimates to θ^*. Due to the fact that the $\mathcal{LU}$ update rule is linear, the driving noise terms are $\mathcal{L}_2$ bounded, and the stable (as shown in the proof) Lyapunov function $V(\cdot)$ assumes a positive definite quadratic form. Hence, by studying the recursion of the deterministic sequence $\{\mathbb{E}[V(\mathbf{x}(i))]\}$ and by similar arguments[5] as in [46] (Lemma 4), we conclude the following:

Lemma 8 (Mean square convergence) Let the hypotheses of Theorem 7 hold and, in addition, the weight sequence $\{\alpha(i)\}$ satisfy the following:

$$\alpha(i) = \frac{a}{(i+1)^{\tau}}$$

where $a > 0$ and $.5 < \tau \leq 1$. Then, the a.s. convergence in Theorem 7 holds in $\mathcal{L}_2$ also, i.e., for all n,

$$\lim_{i\to\infty} \mathbb{E}\left[\left\|\mathbf{x}_n(i) - \theta^*\right\|^2\right] = 0$$

C. Asymptotic Variance: $\mathcal{LU}$

In this subsection, we carry out a convergence rate analysis of the $\mathcal{LU}$ algorithm by studying its moderate deviation characteristics. We summarize here some definitions and terminology from the statistical literature, used to characterize the performance of sequential estimation procedures (see [47]).

Definition 9 (Asymptotic Normality) A sequence of estimates $\{\mathbf{x}^{\bullet}(i)\}_{i\geq 0}$ is asymptotically normal if for every $\theta^* \in \mathcal{U}$, there exists a positive semidefinite matrix $S(\theta^*) \in \mathbb{R}^{M\times M}$, such that,

$$\lim_{i\to\infty} \sqrt{i}\left(\mathbf{x}^{\bullet}(i) - \theta^*\right) \Longrightarrow \mathcal{N}\left(\mathbf{0}_M, S(\theta^*)\right)$$

The matrix $S(\theta^*)$ is called the asymptotic variance of the estimate sequence $\{\mathbf{x}^{\bullet}(i)\}_{i\geq 0}$.

In the following we prove the asymptotic normality of the $\mathcal{LU}$ algorithm and explicitly characterize the resulting asymptotic variance. To this end, define

$$\begin{aligned} S_H = \mathbb{E}\Bigg[& \left(\overline{D}_{\overline{H}} \begin{bmatrix} \widetilde{H}_1(i) & & \\ \ddots & \ddots & \ddots \\ & & \widetilde{H}_N(i) \end{bmatrix} \mathbf{1}_N \theta^*\right) \\ & \left(\overline{D}_{\overline{H}} \begin{bmatrix} \widetilde{H}_1(i) & & \\ \ddots & \ddots & \ddots \\ & & \widetilde{H}_N(i) \end{bmatrix} \mathbf{1}_N \theta^*\right)^T \Bigg] \end{aligned} \tag{28}$$

Let $\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)$ be the smallest eigenvalue of $\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]$ and recall S_{ζ}, S_q in (4) and (12).

We now state the main result of this subsection, establishing the asymptotic normality of the $\mathcal{LU}$ algorithm.

Theorem 10 ($\mathcal{LU}$: Asymptotic efficiency/ normality) Let the $\mathcal{LU}$ algorithm under **(A.1)-(A.5)** with link weight sequence, $\{\alpha(i)\}_{i\geq 0}$ that is given by:

$$\alpha(i) = \frac{a}{i+1}, \ \forall i$$

for some constant $a > 0$. Let $\{\mathbf{x}(i)\}_{i\geq 0}$ be the state sequence generated. Then, if $a > \frac{1}{2\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)}$, we have

$$\sqrt{(i)}\left(\mathbf{x}(i) - \mathbf{1}_N \otimes \theta^*\right) \Longrightarrow \mathcal{N}(\mathbf{0}, S(\theta^*)) \tag{29}$$

where

$$S(\theta^*) = a^2 \int_0^{\infty} e^{\Sigma v} S_0 e^{\Sigma v} dv, \tag{30}$$

$$\Sigma = -a\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right] + \frac{1}{2}I, \tag{31}$$

[5] Note, that Lemma 4 in [46] does not assume the additional term due to new observations at each iteration. However, this does not pose difficulties as the observation weights are the same as the consensus weights.

and

$$S_0 = S_H + \overline{D}_{\overline{H}} S_\zeta \overline{D}_{\overline{H}}^T + b^2 S_q$$

In particular, at any sensor n, the estimate sequence, $\{\mathbf{x}_n(i)\}_{i\geq 0}$ is asymptotically normal:

$$\sqrt(i)\left(\mathbf{x}_n(i) - \theta^*\right) \Longrightarrow \mathcal{N}(\mathbf{0}, S_{nn}(\theta^*))$$

where, $S_{nn}(\theta^*) \in \mathbb{R}^{M\times M}$ denotes the n-th principal block of $S(\theta^*)$.

Proof: The proof involves a step-by-step verification of Assumptions **(C.1)-(C.5)** of Theorem 29 (Appendix A), since the Assumptions **(B.1)-(B.5)** are already shown to be satisfied (see, Theorem 7.) Recall $R\left(\mathbf{x}\right)$ and $\Gamma\left(i+1, \mathbf{x}, \omega\right)$ from Theorem 7 ((22)-(23)). From (22), Assumption **(C.1)** of Theorem 29 is satisfied with

$$B = -\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right]$$

and $\delta\left(\mathbf{x}\right) \equiv 0$. Assumption **(C.2)** is satisfied by hypothesis, while the condition $a > \frac{1}{2\lambda_{\min}\left(b\overline{L}\otimes I_M + D_{\overline{H}}\right)}$ implies

$$\Sigma = -a\left[b\overline{L}\otimes I_M + D_{\overline{H}}\right] + \frac{1}{2}I_{NM} = aB + \frac{1}{2}I_{NM}$$

is stable, and hence Assumption **(C.3)**. To verify Assumption **(C.4)**, we have from Assumption **(A.4)**

$$\begin{aligned} A\left(i, \mathbf{x}\right) &= \mathbb{E}\left[\Gamma\left(i+1, \mathbf{x}, \omega\right)\Gamma^T\left(i+1, \mathbf{x}, \omega\right)\right] \\ &= b^2\mathbb{E}\left[\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x}\mathbf{x}^T\left(\widetilde{L}(i)\otimes I_M\right)^T\right] \\ &\quad + \mathbb{E}\left[\left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right)\right. \\ &\qquad\qquad \left.\left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right)^T\right] \\ &\quad + b^2\mathbb{E}\left[\left(\mathbf{\Upsilon}(i) + \mathbf{\Psi}(i)\right)\left(\mathbf{\Upsilon}(i) + \mathbf{\Psi}(i)\right)^T\right] \end{aligned} \tag{32}$$

From the i.i.d. assumptions, we note that all the three terms on the R.H.S. of (32) are independent of i, and, in particular, the last two terms are constants. For the first term, we note that

$$\lim_{\mathbf{x}\to\mathbf{1}_N\otimes\theta^*}\mathbb{E}\left[\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x}\mathbf{x}^T\left(\widetilde{L}(i)\otimes I_M\right)^T\right] = \mathbf{0}$$

from the bounded convergence theorem, as the entries of $\left\{\widetilde{L}(i)\right\}_{i\geq 0}$ are bounded and

$$\left(\widetilde{L}(i)\otimes I_M\right)\left(\mathbf{1}_N\otimes\theta^*\right) = \mathbf{0} \tag{33}$$

For the second term on the R.H.S. of (32), we have

$$\begin{aligned} &\mathbb{E}\left[\left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right)\left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right)^T\right] \\ &= \mathbb{E}\left[\left(\overline{D}_{\overline{H}}\begin{bmatrix} \widetilde{H}_1(i) & & \\ \ddots & \ddots & \ddots \\ & & \widetilde{H}_N(i) \end{bmatrix}\mathbf{1}_N\theta^*\right)\right. \\ &\qquad \left.\left(\overline{D}_{\overline{H}}\begin{bmatrix} \widetilde{H}_1(i) & & \\ \ddots & \ddots & \ddots \\ & & \widetilde{H}_N(i) \end{bmatrix}\mathbf{1}_N\theta^*\right)^T\right] + \\ &\qquad + \mathbb{E}\left[\overline{D}_{\overline{H}}\zeta\zeta^T\overline{D}_{\overline{H}}^T\right] \\ &= S_H + \overline{D}_{\overline{H}}S_\zeta\overline{D}_{\overline{H}}^T \end{aligned} \tag{34}$$

where the last step follows from (28),(4). Finally, we note the third term on the R.H.S. of (32) is $b^2 S_q$, see (12). We thus have from (32)-(34)

$$\lim_{i\to\infty,\ \mathbf{x}\to\mathbf{x}^*} A\left(i,\mathbf{x}\right) = S_H + \overline{D}_{\overline{H}} S_\zeta \overline{D}_{\overline{H}}^T + b^2 S_q = S_0$$

We now verify Assumption **(C.5)**. Consider a fixed $\epsilon > 0$. We note that (83) is a restatement of the uniform integrability of the random family, $\left\{\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2\right\}_{i\geq 0,\ \|\mathbf{x}-\theta^*\|<\epsilon}$. From (23), we have

$$\begin{aligned}
\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2 &= \Big\|b\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x} - \\
&\quad - \left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right) + b\boldsymbol{\Upsilon}(i) + b\boldsymbol{\Psi}(i)\Big\|^2 \\
&= \Big\|b\left(\widetilde{L}(i)\otimes I_M\right)\left(\mathbf{x}-\theta^*\right) - \\
&\quad - \left(\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right) + b\boldsymbol{\Upsilon}(i) + b\boldsymbol{\Psi}(i)\Big\|^2 \\
&\leq 9\Big[\left\|\left(b\widetilde{L}(i)\otimes I_M\right)\left(\mathbf{x}-\theta^*\right)\right\|^2 + \\
&\quad + \left\|\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right\|^2 + b^2\left\|\boldsymbol{\Upsilon}(i) + \boldsymbol{\Psi}(i)\right\|^2\Big]
\end{aligned}$$

where we used (33) and the inequality, $\|\mathbf{y}_1+\mathbf{y}_2+\mathbf{y}_3\|^2 \leq 9\left[\|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2 + \|\mathbf{y}_3\|^2\right]$, for vectors $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$. From (26) we note that, if $\|\mathbf{x}-\theta^*\| < \epsilon$,

$$\left\|\left(b\widetilde{L}(i)\otimes I_M\right)\left(\mathbf{x}-\theta^*\right)\right\|^2 \leq c_2\epsilon^2$$

From (II-C), the family [defined in (35) below]

$$\left\{\widetilde{\Gamma}\left(i+1,\mathbf{x},\omega\right)\right\}_{i\geq 0,\|\mathbf{x}-\theta^*\|<\epsilon}$$

dominates the family

$$\left\{\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2\right\}_{i\geq 0,\ \|\mathbf{x}-\theta^*\|<\epsilon},$$

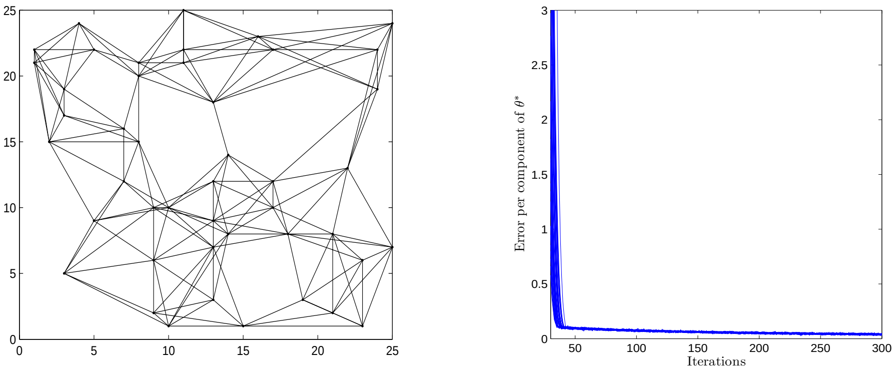
where

$$\begin{aligned}
\widetilde{\Gamma}\left(i+1,\mathbf{x},\omega\right) = 9\Big[c_2\epsilon^2 &+ \left\|\overline{D}_{\overline{H}}\mathbf{z}(i) - D_{\overline{H}}\mathbf{1}_N\otimes\theta^*\right\|^2 \\
&+ b^2\left\|\boldsymbol{\Upsilon}(i) + \boldsymbol{\Psi}(i)\right\|^2\Big]
\end{aligned} \tag{35}$$

The family $\left\{\widetilde{\Gamma}\left(i+1,\mathbf{x},\omega\right)\right\}_{i\geq 0,\ \|\mathbf{x}-\theta^*\|<\epsilon}$ is i.i.d. and hence uniformly integrable (see [50]). Then the family $\left\{\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2\right\}_{i\geq 0,\ \|\mathbf{x}-\theta^*\|<\epsilon}$ is also uniformly integrable since it is dominated by the uniformly integrable family $\left\{\widetilde{\Gamma}\left(i+1,\mathbf{x},\omega\right)\right\}_{i\geq 0,\ \|\mathbf{x}-\theta^*\|<\epsilon}$ (see [50]). Thus **(C.1)-(C.5)** are verified and the theorem follows. $\blacksquare$

D. A Simulation Example

Fig. 1 (b) shows the performance of $\mathcal{LU}$ for the network of $N = 45$ sensors in Fig. 1 (a), where the sensors are deployed randomly on a 25×25 grid. The sensors communicate in a fixed radius and are further constrained to have a maximum of 6 neighbors per node. The true parameter $\theta^* \in \mathbb{R}^{45}$. Each node is associated with a single component of θ^*, i.e., $\overline{H}_n = \mathbf{e}_n^T$, the unit vector of zeros, except entry n that is 1. For the experiment, each component of θ^* is generated by an instantiation of a zero mean Gaussian random variable of variance 25. The parameter θ^* represents the state of the field to be estimated. In this example, the field is white, stationary, and hence each sample of the field has the same Gaussian distribution and is independent of the others. More generally, the components of θ^* may correspond to random field samples, as dictated by the sensor deployment, that can possibly arise from the discretization of a field governed by a PDE. Each sensor observes the corresponding field component in additive Gaussian noise. For example, sensor 1 observes $z_1(t) = \theta_1^* + \zeta_1(t)$, where $\zeta_1(t) \sim \mathcal{N}(0,1)$.



(a)

(b)

Fig. 1. Illustration of distributed linear parameter estimation. (a) Example network deployment of 45 nodes. (b) Convergence of normalized estimation error at each sensor.

Clearly, such a model satisfies the distributed observability condition

$$G = \sum_{n=1}^{N} \overline{H}_n^T \overline{H}_n = I_{45} = G^{-1}$$

Fig. 1(b) shows the normalized error at every sensor plotted against the iteration index i for an instantiation of the algorithm. The normalized error for the n-th sensor at time i is given by the quantity $\|\mathbf{x}_n(i) - \theta^*\| / 45$, i.e., the estimation error normalized by the dimension of θ^*. We note that the errors converge to zero as established by the theoretical findings. The decrease is rapid at the beginning and slows down at i increases. This is a standard property of stochastic approximation based algorithms, consequence of the decreasing weight sequence $\alpha(i)$ required for convergence. From the plots, although the individual sensors are low rank observations of the true parameter, by collaborating, each sensor reconstructs the true parameter value, as desired.

E. An Example

From Theorem 10 and (28), we note that the asymptotic variance is independent of θ^*, if the observation matrices are non-random. In that case, it is possible to optimize (minimize) the asymptotic variance over the weights a and b. In the following, we study a special case permitting explicit computations and that leads to interesting results. Consider a scalar parameter ($M = 1$) and let each sensor n have the same i.i.d. observation model,

$$z_n(i) = h\theta^* + \zeta_n(i)$$

where $h \neq 0$ and $\{\zeta_n(i)\}_{i \geq 0,\ 1 \leq n \leq N}$ is a family of independent zero mean Gaussian random variables with variance σ^2. In addition, assume unquantized inter-sensor exchanges. We define the average asymptotic variance per sensor attained by the algorithm $\mathcal{LU}$ as

$$S_{\mathcal{LU}} = \frac{1}{N} \mathrm{Tr}\,(S)$$

where S is given by (30) in Theorem 10. From Theorem 10, we have $S_0 = \sigma^2 h^2 I_N$ and, hence, from (30)

$$\begin{aligned} S_{\mathcal{LU}} &= \frac{a^2\sigma^2 h^2}{N} \mathrm{Tr}\left(\int_0^\infty e^{2\Sigma v} dv \right) \\ &= \frac{a^2\sigma^2 h^2}{N} \int_0^\infty \mathrm{Tr}\left(e^{2\Sigma v} \right) dv \end{aligned}$$

From (31) the eigenvalues of $2\Sigma v$ are $\left[-2ab\lambda_n(\overline{L}) - \left(2ah^2-1\right)\right] v$ for $1 \leq n \leq N$ and we have

$$\begin{aligned}
S_{\mathcal{LU}} &= \frac{a^2\sigma^2h^2}{N}\sum_{n=1}^{N}\int_0^{\infty} e^{\left[-2ab\lambda_n(\overline{L})-\left(2ah^2-1\right)\right]v}dv \\
&= \frac{a^2\sigma^2h^2}{N}\sum_{n=1}^{N}\frac{1}{2ab\lambda_n(\overline{L})+(2ah^2-1)} \\
&= \frac{a^2\sigma^2h^2}{N\left(2ah^2-1\right)} + \frac{a^2\sigma^2h^2}{N} \\
&\quad \sum_{n=2}^{N}\frac{1}{2ab\lambda_n(\overline{L})+(2ah^2-1)}
\end{aligned} \tag{36}$$

In this case, the constraint $a > \frac{1}{2\lambda_{\min}(b\overline{L}\otimes I_M + D_{\overline{H}})}$ in Theorem 10 reduces to $a > \frac{1}{2h^2}$, and hence the problem of optimum a, b design to minimize $S_{\mathcal{LU}}$ is given by

$$S^{*}_{\mathcal{LU}} = \inf_{a>\frac{1}{2h^2},\; b>0} S_{\mathcal{LU}}$$

It is to be noted, that the first term on the last step of (36) is minimized at $a = \frac{1}{h^2}$ and the second term (always non-negative under the constraint) goes to zero as $b \to \infty$ for any fixed $a > 0$. Hence, we have

$$S^{*}_{\mathcal{LU}} = \frac{\sigma^2}{Nh^2} \tag{37}$$

The above shows that, by setting $a = \frac{1}{h^2}$ and b sufficiently large in $\mathcal{LU}$, $S_{\mathcal{LU}}$ is arbitrarily close to $S^{*}_{\mathcal{LU}}$.

We compare this optimum achievable asymptotic variance per sensor, $S^{*}_{\mathcal{LU}}$, attained by $\mathcal{LU}$ to that attained by a centralized scheme. In the centralized scheme, there is a central estimator, which receives measurements from all the sensors and computes an estimate based on all measurements. In this case, the sample mean estimator is an efficient estimator (in the sense of Cramér-Rao) and the estimate sequence $\{x_c(i)\}_{i\geq 0}$ is given by

$$x_c(i) = \frac{1}{Nih}\sum_{n,i} z_n(i)$$

and we have

$$\sqrt{i}\left(x_c(i) - \theta^*\right) \sim \left(0, \mathcal{S}_c\right)$$

where, S_c is the variance (which is also the one-step Fisher information in this case, see, [47]) and is given by

$$S_c = \frac{\sigma^2}{Nh^2}$$

From (37) we note that,

$$S^{*}_{\mathcal{LU}} = S_c$$

Thus the average asymptotic variance attainable by the distributed algorithm $\mathcal{LU}$ is the same as that of the optimum (in the sense of Cramér-Rao) centralized estimator having access to all information simultaneously. This is an interesting result, as it holds irrespective of the network topology. In particular, however sparse the inter-sensor communication graph is, the optimum achievable asymptotic variance is the same as that of the centralized efficient estimator. Note that weak convergence itself is a limiting result, and, hence, the rate of convergence in (29) in Theorem 10 will, in general, depend on the network topology.

F. Some generalizations

We discuss some generalizations of the basic $\mathcal{LU}$ scheme before proceeding to the nonlinear observation models addressed in the subsequent sections. We start by revisiting the scalar example in Section II-E for which the distributed $\mathcal{LU}$ is shown to achieve the performance of the optimal centralized estimator. Interestingly, the above example is not an isolated special case and has several important implications. The observation that by increasing $b > 0$ we can achieve asymptotic variance as close as desired to the centralized estimator hints to a more general time-scale separation in the case of unquantized transmissions. Intuitively, for a fixed $b > 0$, the weight associated to the consensus potential is $b\alpha(i)$, which goes to zero at the same rate as that of the innovation potential. Hence, in the long run, a non-negligible (in the scale $\{\alpha(i)\}$) amount of time is required to disseminate new information acquired by a sensor. In other words, the rate of uncertainty reduction in $\mathcal{LU}$ depends on both the rate of new information acquisition at the sensors and the rate of information dissemination in the network. On the contrary, in a centralized scenario, no additional time is incurred for information dissemination and the rate of uncertainty reduction is the same as the rate of information acquisition. This is manifested, in general, in the asymptotic variance of $\mathcal{LU}$, which is larger than its centralized counterpart due to the additional overhead of the mixing terms (Theorem 10). This suggests that, if the mixing can be carried out at a *faster* scale, the additional overhead due to the mixing time will not be observed at the time scale of observation acquisition and, in effect, the distributed scheme will lead to similar asymptotic variance as in the centralized setting. This is noted in Section II-E, where increasing the relative weight b of the consensus or mixing potential leads to a time scale separation between information dissemination and acquisition. Increasing b beyond bounds suggests that we replace the decreasing weight sequence $\{\alpha(i)\}$ from the consensus term and retain it with a constant weight, or more generally, a weight sequence $\{\beta(i)\}$, that asymptotically dominates $\{\alpha(i)\}$. Such a mixed time scale extension of the $\mathcal{LU}$ is introduced and analyzed in [26]. The results in [26] show that the conclusion in Section II-E for the scalar example (the distributed achieves the centralized performance in terms of asymptotic variance) holds in more general vector parameter settings by appropriately tuning the consensus and innovation weights. This is significant, as it justifies the applicability of distributed estimation schemes over centralized approaches.

The development in this paper assumes stationarity of the sensor observations over time. While this is applicable and is a commonly used assumption in many statistical models, some scenarios inherently lead to non-stationary observation time series. For example, consider a distributed sensor network monitoring a target that fades over time. In this example, the sensor observation models are no longer stationary as the SNR (signal to noise ratio) decays over time, the decay rate being a function of the fading characteristics. Treating nonstationarity requires modification of the algorithm (intuitively, the update rules are no longer stationary) and is pursued in [26].

The above did not exploit the physical significance of the parameter θ. That θ may itself come from a spatially distributed random field was only implicit in the distributed observation model. Typical examples include instrumenting a spatially distributed random field (say a temperature surface) with a sensor network. Another example is of cyberphysical systems, where a network of physical entities equipped with sensors are deployed over a large geographical region. A well known example in this setting is the power grid, a large distributed network of generators and loads. Our results imply that, under appropriate observability conditions, the physical field θ may be reconstructed completely at each node (sensor).[6] However, for such systems, the parameter θ representing the physical field is quite large dimensional, may be of the order of 10^3 or more, as exemplified by the power grid.[7] It is then impractical and unnecessary to reconstruct the high dimensional parameter in its entirety at each node. On the other hand, the node may be interested only in its state, or those of its close neighbors. In general, the observation at each sensor reflects the coupling of a few local physical states and hence, acting alone, a node may not be able to recover its state uniquely. In [51], we develop approaches to address this problem, where each node wants to reconstruct a few components[8] of the large state vector. The estimation approach would lead to low dimensional data exchanges between neighboring sensors (nodes) and local estimate updates would involve only those components, the node wants to reconstruct. Due to the partial information exchange between sensors and the fact that sensors may have different goals, the distributed observability no longer culminates to the sum of network connectivity and global connectivity, but requires more subtle relations between the observation model and the network topology. In general, the scope of such problems of distributed estimation with partial inter-sensor information exchange is quite broad and challenging, and we refer the reader to [51] (Chapter 5) for an exposition.

[6] A node, in this context, refers to the physical entity at a geographical location, for example, a generator in a power grid. The sensing or measurement unit associated to a node is referred to as a sensor. The state of a node represents the field intensity at that point, for example, the phase of a generator.

[7] For problems involving infinite dimensional systems, such as the temperature distribution over a domain in the Euclidean space, any reasonable discretization would lead to a large dimensional θ.

[8] These components may vary from node to node.

III. Nonlinear Observation Models: Agorithm $\mathcal{NU}$

The previous section developed the algorithm $\mathcal{LU}$ for distributed parameter estimation when the observation model is linear. In this section, we extend the previous development to accommodate more general classes of nonlinear observation models. We comment briefly on the organization of this section. In Section III-B, we introduce notation and setup the problem, and in Section III-C we present the $\mathcal{NU}$ algorithm for distributed parameter estimation for nonlinear observation models and establish conditions for its consistency.

A. Nonlinear Observation Models

Similar to Section II, let $\theta^* \in \mathcal{U} \subset \mathbb{R}^M$ be the true but unknown parameter value. We assume that the domain $\mathcal{U}$ is an open set in $\mathbb{R}^M$. In the general case, the observation model at each sensor n consists of an i.i.d. sequence $\{\mathbf{z}_n(i)\}_{i\geq 0}$ in $\mathbb{R}^{M_N}$ with

$$\mathbb{P}_{\theta^*}[\mathbf{z}_n(i) \in \mathcal{D}] = \int_{\mathcal{D}} dF_{n,\theta^*}, \quad \forall\ \mathcal{D} \in \mathbb{B}^{M_N} \tag{38}$$

where F_{n,θ^*} denotes the distribution function of the random vector $\mathbf{z}_n(i)$. For consistent parameter estimates, even in centralized settings, some form of observability needs to be imposed on the nonlinear model. In the following, we assume that the distributed observation model is *separably estimable*, a notion which we introduce now.

Definition 11 (Separably Estimable) Let $\{\mathbf{z}_n(i)\}_{i\geq 0}$ be the i.i.d. observation sequence at sensor n, where $1 \leq n \leq N$. We call the parameter estimation problem to be separably estimable, if there exist functions $g_n(\cdot) : \mathbb{R}^{M_N} \longmapsto \mathbb{R}^{\overline{M}}$, $\forall 1 \leq n \leq N$, such that the function $h(\cdot) : \mathcal{U} \longmapsto \mathbb{R}^{\overline{M}}$ is continuous and invertible on $\mathcal{U}$

$$h(\theta) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_\theta\left[g_n(\mathbf{z}_n(i))\right] \tag{39}$$

Remark 12 Before providing examples of separably estimable observation models and demonstrating the applicability of the notion, we comment on the definition.

(i) We note that the factor $\frac{1}{N}$ in (39) is just for notational convenience, as will be seen later. In fact, the $\frac{1}{N}$ can be absorbed by redefining the functions $g_n(\cdot)$. Also, it is implicitly assumed that the random vectors $g_n(\mathbf{z}_n(i))$ are integrable w.r.t. the measures $\mathbb{P}_\theta$ for $\theta \in \mathcal{U}$.

(ii) Let $h(\mathcal{U}) \subset \mathbb{R}^{\overline{M}}$ denote the range of $h(\cdot)$. The continuity of $h(\cdot)$ implies that $h(\mathcal{U})$ is open. Let $h^{-1} : h(\mathcal{U}) \longmapsto \mathbb{R}^M$ denote the inverse of $h(\cdot)$ (which is necessarily continuous on $h(\mathcal{U})$.) It then follows that $h^{-1}(\cdot)$ has a measurable extension defined over all of $\mathbb{R}^{\overline{M}}$. In the following we will assume that $h^{-1}(\cdot)$ has been measurably extended and, by abusing notation, denote this extension by h^{-1}.

(iii) We will show that the notion of separably estimable models introduced above is, in fact, necessary and sufficient to guarantee the existence of consistent distributed estimation procedures for a wide range of practical scenarios. This condition may also be viewed as a natural generalization of the observability constraint of Assumption **(A.5)** in the linear model. Indeed, if, assuming the linear model, we define $g_n(\mathbf{z}_n(i)) = \overline{H}_n^T\mathbf{z}_n(i)$, $\forall 1 \leq n \leq N$ in (39), we have $h(\theta) = G\theta$, where G is defined in (5). Then, invertibility of (39) is equivalent to Assumption **(A.5)**, i.e., to invertibility of G; hence, the linear model is an example of a separably estimable problem. Note that, if an observation model is separably estimable, then the choice of functions $g_n(\cdot)$ is not unique. Indeed, given a separably estimable model, it is important to figure out an appropriate decomposition, as in (39), because the convergence properties of the algorithms (Algorithm $\mathcal{NU}$, Section III-C) to be studied are intimately related to the behavior of these functions. Finally, we note that, in general, $\overline{M} \neq M$, and the dimension $\overline{M}$ of the range space of $h(\cdot)$ is very much linked to the memory and transmission requirements of the distributed algorithm $\mathcal{NLU}$ to be studied in Section IV. In this sense, the function $h(\cdot)$ plays the role of a complete sufficient statistic as used in classical (centralized) estimation, the major difference being the *distributed computability* (to be made precise later) of $h(\cdot)$ in the current setting.

In Sections III-C and IV, respectively, we will present algorithms $\mathcal{NU}$ and $\mathcal{NLU}$ for distributed parameter estimation in separably estimable models. While the $\mathcal{NLU}$ provides consistent parameter estimates for all separably estimable models, the $\mathcal{NU}$ requires further (mainly of the Lipschitz type) conditions on the functions $g_n(\cdot)$ and $h(\cdot)$. However, in cases where the $\mathcal{NU}$ is applicable, it automatically leads to convergence rate guarantees in the context of asymptotic normality. These differences are further clarified in Section V. Before discussing these algorithms in detail, we provide examples of separably estimably models in the following.

Examples: Signal in additive noise models

We now demonstrate an important and large class of distributed observation models possessing the separably estimable property, thus justifying the generality and applicability of the notion.

A wide range of observation models are of the signal in additive noise type. In particular, for each n, denote by $\{\zeta_n(i)\}$ the zero mean i.i.d. observation noise at the n-th sensor of arbitrary distribution (the distribution may vary from sensor to sensor.) The sensor observation model is said to be of *signal in additive noise type*, if the observation sequence $\{\mathbf{z}_n(i)\}$ at the n-th sensor is of the form:

$$\mathbf{z}_n(i) = f_n(\theta^*) + \zeta_n(i) \tag{40}$$

Here $f_n : \mathcal{U} \longmapsto \mathbb{R}^{M_n}$ denotes the transformed (nonlinearly) signal (or parameter) observed at sensor n, further corrupted by additive noise. The following simple proposition characterizes the subclass of signal in additive noise observable models with the separably estimable property:

Proposition 13 Let $f : \mathcal{U} \longmapsto \mathbb{R}^{\sum_{n=1}^{N} M_n}$ be defined by, $f(\theta) = [f_1^T(\theta) \cdots f_N^T(\theta)]^T$. Then, the above signal in additive noise observation model (see (40)) is separably estimable if $f(\cdot)$ is continuous and invertible on $\mathcal{U}$.

Before providing the rather straightforward proof, we note the consequences of Proposition 13. Consider a hypothetical centralized estimator having access to all the sensor observations at all times. Clearly, the $\sum_{n=1}^{N} M_n$ dimensional i.i.d. observation sequence $\{\mathbf{z}(i)\}$ at such a center is given by:

$$\mathbf{z}(i) = f(\theta^*) + \zeta(i)$$

In general, for arbitrary statistics of the noise sequence $\{\zeta(i)\}$, it is necessary that the function f be invertible, for the center to yield a consistent estimate of the parameter. In fact, for consistent centralized estimates, the invertibility of $f(\cdot)$ is required, even when the observation noise is identically zero. On the other hand, Proposition 13 asserts that the invertibility of $f(\cdot)$ (and its continuity) is sufficient to guarantee that the model is separably estimable and hence the existence of consistent distributed estimation schemes. Hence, at least in the class of widely adopted signal in additive noise models, centralized observability is equivalent to distributed observability (formulated here in terms of separable estimability.) This further justifies the notion of separable estimability as a reasonable generalization of the concept of centralized observability to distributed nonlinear settings. In Section IV-D, we will show that the $\mathcal{NLU}$ algorithm provides a completely distributed approach to the static phase estimation problem in power grids of generators and loads based on line flow measurements, an important practical example of a distributed nonlinear signal in additive noise model.

Proof: The proof follows in a straightforward manner from the definition. For each n, define the function $g_n : \mathbb{R}^{M_n} \longmapsto \mathbb{R}^{\sum_{n=1}^{N} M_n}$ by

$$g_n(\mathbf{y}) = [\mathbf{0}_{M_1}^T \mathbf{0}_{M_2}^T \cdots \mathbf{y}^T \cdots \mathbf{0}_{M_N}^T]^T, \quad \forall \mathbf{y} \in \mathbb{R}^{M_n} \tag{41}$$

Recall $\mathbf{0}_{M_1} \in \mathbb{R}^{M_1}$ denotes the column vector of M_1 zeros and so on. By the independence of the noise sequence $\{\zeta(i)\}$, it the follows that

$$\sum_{n=1}^{N} \mathbb{E}_{\theta^*}\left[g_n(\mathbf{z}_n(i))\right] = f(\theta^*)$$

The continuity and invertibility of $f(\cdot)$ then establishes the separable estimability of the model (Definition 11) by the correspondence $h(\cdot) = \frac{1}{N} f(\cdot)$. ■

By using the same arguments we demonstrate a larger class of separably estimable models as follows:

Proposition 14 Let the observation sequence $\{\mathbf{z}_n(i)\}$ at the n-th sensor be of the form:

$$\mathbf{z}_n(i) = f_n(\theta^*, \zeta_n^1(i)) + \zeta_n^2(i) \tag{42}$$

where $f_n : \mathcal{U} \times \mathbb{R}^{M_n^1} \longmapsto \mathbb{R}^{M_n}$, $\{\zeta_n^1(i), \zeta_n^2(i)\} \in \mathbb{R}^{M_n^1} \times \mathbb{R}^{M_n}$ is a temporally i.i.d. sequence and $\{\zeta_n^2(i)\}$ is zero mean. Assuming that the moments exist, define the function $\overline{f}_n : \mathcal{U} \longmapsto \mathbb{R}^{M_n}$, for each n, by

$$\overline{f}_n(\theta) = \mathbb{E}_\theta\left[f_n(\theta^*, \zeta_n^1(i))\right]$$

Further, let $\overline{f}:\mathcal{U}\longmapsto\mathbb{R}^{\sum_{n=1}^{N}M_n}$ be defined by, $\overline{f}(\theta)=[\overline{f}_1^T(\theta)\cdots\overline{f}_N^T(\theta)]^T$. Then, the observation model in (42) is separably estimable if $\overline{f}(\cdot)$ is continuous and invertible on $\mathcal{U}$.

Remark 15 The generic model considered in Proposition 14 subsumes the class of signals with multiplicative noise models, by suitably defining the functions $f_n(\theta^*,\zeta_n^1(i))$ and setting the additive noise component $\zeta_n^2(i)$ to zero. We also note that a general guideline for choosing the functions $g_n(\cdot)$ for the signal in additive noise type models based on problem data is given in (41). From a similar line of reasoning, it follows that the same choice of $g_n(\cdot)$ works for the larger class of separably estimable models considered in Proposition 14.

In the following subsection, we present the algorithm $\mathcal{NU}$ for distributed parameter estimation in nonlinear separably estimable observation models.

B. Algorithm $\mathcal{NU}$ and Assumptions

Before introducing the algorithm, we formally state the generic observation and communication assumptions required by the $\mathcal{NU}$.

(D.1)Separably Estimable Model: The nonlinear observation model (38) is separably estimable (Definition 11). In particular, at iteration i, the observations across different sensors need not be independent. In other words, we allow spatial correlation, but require temporal independence. Also, other than the structural assumption of *separable estimability*, no assumptions are required on the noise statistics, in particular, its distribution.

(D.2)Random Link Failure, Quantized Communication: The random link failure model is the model given in Section I-B; similarly, we assume quantized inter-sensor communication with subtractive dithering.

(D.3)Independence and Moment Assumptions: The sequences $\{L(i)\}_{i\geq 0}$,$\{\mathbf{z}_n(i)\}_{1\leq n\leq N,\ i\geq 0}$,$\{\nu_{nl}^m(i)\}$ (dither sequence, as in (II-A)) are mutually independent. Let $\overline{M}=\sum_{n=1}^{N}M_n$ and define $h_n:\mathbb{R}^M\longmapsto\mathbb{R}^{\overline{M}}$, by

$$h_n(\theta)=\mathbb{E}_\theta\left[g_n(\mathbf{z}_n(i))\right],\ \forall 1\leq n\leq N$$

We make the assumption $\forall\theta\in\mathcal{U}$:

$$\mathbb{E}_\theta\left[\left\|\frac{1}{N}\sum_{n=1}^{N}g_n(\mathbf{z}_n(i))-h(\theta)\right\|^2\right]=\eta(\theta)<\infty, \tag{43}$$

We thus assume the existence of quadratic moments of the (transformed) random variables $g_n(\mathbf{z}_n(i))$. For example, under the reasonable hypotheses of Propositions 13-14, the functions $g_n(\cdot)$ may be taken to be linear, and Assumption **(D.3)** then coincides with the existence of quadratic moment of the observations $\mathbf{z}_n(i)$. In general, since the choice of the functions $g_n(\cdot)$ for a separably estimable model is not unique, the moment Assumption **(D.3)** may enter as a selection criterion of the transformations $g_n(\cdot)$.

In Section III-C and Section IV, we give two algorithms, $\mathcal{NU}$ and $\mathcal{NLU}$, respectively, for the distributed estimation problem **(D.1)-(D.3)** and provide conditions for consistency and other properties of the estimates.

C. Algorithm $\mathcal{NU}$

In this subsection, we present the algorithm $\mathcal{NU}$ for distributed parameter estimation in separably estimable models under Assumptions **(D.1)-(D.3)**.

Algorithm $\mathcal{NU}$: Each sensor n performs the following estimate update:

$$\mathbf{x}_n(i+1)=\mathbf{x}_n(i)-\alpha(i)\left[\sum_{l\in\Omega_n(i)}\beta\left(\mathbf{x}_n(i)-\right.\right.$$
$$\left.\left.\mathbf{q}(\mathbf{x}_l(i)+\nu_{nl}(i))\right)+\mathcal{K}_n\left(h_n(\mathbf{x}_n(i))-g_n(\mathbf{z}_n(i))\right)\right]$$

based on $\mathbf{x}_n(i)$, $\{\mathbf{q}(\mathbf{x}_l(i)+\nu_{nl}(i))\}_{l\in\Omega_n(i)}$, and $\mathbf{z}_n(i)$, which are all available to it at time i. The sequence, $\left\{\mathbf{x}_n(i)\in\mathbb{R}^M\right\}_{i\geq 0}$, is the estimate (state) sequence generated at sensor n. The weight sequence $\{\alpha(i)\}_{i\geq 0}$ satisfies the persistence condition of Assumption **(B.5)** and $\beta>0$ is chosen to be an appropriate constant. Finally,

$\mathcal{K}_n \in \mathbb{R}^{M\times\overline{M}}$ is an appropriately chosen matrix gain, possibly varying from sensor to sensor. Similar to (7) the above update can be written in compact form as

$$\begin{aligned}\mathbf{x}(i+1) = \mathbf{x}(i) - \alpha(i)\left[\beta(L(i)\otimes I_M)\mathbf{x}(i) + \right.\\ \left. + \mathcal{K}\left(M(\mathbf{x}(i)) - J(\mathbf{z}(i))\right) + \mathbf{\Upsilon}(i) + \mathbf{\Psi}(i)\right]\end{aligned} \tag{44}$$

where $\mathbf{\Upsilon}(i), \mathbf{\Psi}(i)$ are as in (8)-(9) and $\mathbf{x}(i) = [\mathbf{x}_1^T(i)\cdots\mathbf{x}_N^T(i)]^T$ is the vector of sensor states (estimates). The functions $M(\mathbf{x}(i))$ and $J(\mathbf{z}(i))$ are given by

$$\begin{aligned}M(\mathbf{x}(i)) &= \left[h_1^T(\mathbf{x}_1(i))\cdots h_N^T(\mathbf{x}_N(i))\right]^T,\\ J(\mathbf{z}(i)) &= \left[g_1^T(\mathbf{z}_1(i))\cdots g_N^T(\mathbf{z}_N(i))\right]^T\end{aligned} \tag{45}$$

and $\mathcal{K} = \operatorname{diag}(\mathcal{K}_1, \cdots, \mathcal{K}_N)$ is the block diagonal matrix of gains.

As an example, for the linear observation model, by defining $g_n(\mathbf{z}_n(i))$ to be $\overline{H}_n^T\mathbf{z}_n(i)$ (and choosing the matrix gains $\mathcal{K}_n$ to be I_M), the $\mathcal{NU}$ reduces to the $\mathcal{LU}$ updates (11).

We note that the update scheme in (44) is nonlinear and hence convergence properties can, in general, be characterized through the existence of appropriate stochastic Lyapunov functions. In particular, if we can show that the iterative scheme in (44) falls under the purview of a general result like Theorem 29, we can establish properties like consistency, normality etc. To this end, we note, that (44) can be written as

$$\begin{aligned}\mathbf{x}(i+1) = \mathbf{x}(i) - \alpha(i)\left[\beta\left(\overline{L}\otimes I_M\right)\left(\mathbf{x}(i) - \mathbf{1}_N\otimes\theta^*\right)\right.\\ + \beta\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x}(i) + \mathcal{K}\left(M(\mathbf{x}(i)) - M(\mathbf{1}_N\otimes\theta^*)\right)\\ \left.-\mathcal{K}\left(J(\mathbf{z}(i)) - M(\mathbf{1}_N\otimes\theta^*)\right) + \mathbf{\Upsilon}(i) + \mathbf{\Psi}(i)\right]\end{aligned}$$

which becomes in the notation of Theorem 29

$$\begin{aligned}&\mathbf{x}(i+1)=\mathbf{x}(i)+\alpha(i)[R(\mathbf{x}(i)) + \Gamma\left(i+1, \mathbf{x}(i), \omega\right)]\\ &R\left(\mathbf{x}\right) = -\left[\beta\left(\overline{L}\otimes I_M\right)\left(\mathbf{x} - \mathbf{1}_N\otimes\theta^*\right)\right.\\ &\qquad\qquad\left. + \mathcal{K}\left(M\left(\mathbf{x}\right) - M(\mathbf{1}_N\otimes\theta^*)\right)\right]\end{aligned} \tag{46}$$

$$\begin{aligned}\Gamma\left(i+1, \mathbf{x}, \omega\right) = -\left[\beta\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{x} - \right.\\ \left. - \mathcal{K}\left(J(\mathbf{z}(i)) - M(\mathbf{1}_N\otimes\theta^*)\right) + \mathbf{\Upsilon}(i) + \mathbf{\Psi}(i)\right]\end{aligned} \tag{47}$$

Consider the filtration, $\{\mathcal{F}_i\}_{i\geq 0}$,

$$\mathcal{F}_i=\sigma\Big(\mathbf{x}(0), \Big\{L(j), \{\mathbf{z}_n(j)\}_{1\leq N}, \mathbf{\Upsilon}(j), \mathbf{\Psi}(j)\Big\}_{0\leq j<i}\Big) \tag{48}$$

Clearly, under **(D.1)-(D.3)**, the $\{\mathbf{x}(i)\}_{i\geq 0}$ generated by $\mathcal{NU}$ is Markov w.r.t. $\{\mathcal{F}_i\}_{i\geq 0}$, and the definition in (47) renders the random family, $\{\Gamma\left(i+1, \mathbf{x}, \omega\right)\}_{\mathbf{x}\in\mathbb{R}^{NM}}$, $\mathcal{F}_{i+1}$ measurable, zero-mean, and independent of $\mathcal{F}_i$ for fixed $i+1$. Thus **(B.1)-(B.2)** of Theorem 29 are satisfied, and we have the following.

Proposition 16 ($\mathcal{NU}$:Consistency/ asymp. normality) Let the sequence $\{\mathbf{x}(i)\}_{i\geq 0}$ be generated by $\mathcal{NU}$. Let $R\left(\mathbf{x}\right), \Gamma\left(i+1, \mathbf{x}, \omega\right), \mathcal{F}_i$ be as in (46)-(47). Then, if there exists a function $V\left(\mathbf{x}\right)$ satisfying **(B.3)-(B.4)** at $\mathbf{x}^* = \mathbf{1}_N\otimes\theta^*$, the estimate sequence $\{\mathbf{x}_n(i)\}_{i\geq 0}$ at any sensor n is consistent. In other words,

$$\mathbb{P}_{\theta^*}[\lim_{i\to\infty}\mathbf{x}_n(i) = \theta^*,\ \forall n] = 1$$

If, in addition, **(C.1)-(C.4)** are satisfied, the sequence $\{\mathbf{x}_n(i)\}_{i\geq 0}$ at any sensor n is asymptotically normal.

Proposition 16 states that, a.s. asymptotically, the network reaches consensus, and the estimates at each sensor converge to the true value of the parameter vector $\theta^\star$. The Proposition relates these convergence properties of $\mathcal{NU}$ to the existence of suitable Lyapunov functions. For a particular observation model characterized by the corresponding functions $h_n(\cdot), g_n(\cdot)$, if one can come up with an appropriate Lyapunov function satisfying the assumptions of Proposition 16, then consistency and asymptotic normality are guaranteed. Existence of a suitable Lyapunov condition is sufficient for consistency, but may not be necessary. In particular, there may be observation models for which the $\mathcal{NU}$ algorithm is consistent, but there exists no Lyapunov function satisfying the assumptions of Proposition 16.[9] Also, even if a suitable Lyapunov function exists, it may be difficult to guess its form, because there is no systematic (constructive) way of coming up with Lyapunov functions for generic models.

However, for our problem of interest, some additional weak assumptions on the observation model, for example, Lipschitz continuity of the functions $h_n(\cdot)$, will guarantee the existence of suitable Lyapunov functions, thus establishing convergence properties of the $\mathcal{NU}$ algorithm. The rest of this subsection studies this issue and presents different sufficient conditions on the observation model, which guarantee that the assumptions of Proposition 16 are satisfied, leading to the a.s. convergence of the $\mathcal{NU}$ algorithm. For the development in the rest of the subsection, we assume that $\overline{M} = M$ in the decomposition (39) and $\mathcal{K}_n = I_M$ for all n. The extensions of Theorems 18-19 to $\overline{M} \neq M$ and arbitrary gains $\mathcal{K}_n$ are immediate. We start with the following definition:

Definition 17 (Consensus Subspace) We define the consensus subspace, $\mathcal{C} \subset \mathbb{R}^{MN}$ as

$$\mathcal{C} = \left\{ \mathbf{y} \in \mathbb{R}^{NM} \,\middle|\, \mathbf{y} = \mathbf{1}_N \otimes \widetilde{\mathbf{y}},\, \widetilde{\mathbf{y}} \in \mathbb{R}^M \right\} \tag{49}$$

For $\mathbf{y} \in \mathbb{R}^{NM}$, we denote its component in $\mathcal{C}$ by $\mathbf{y}_{\mathcal{C}}$ and its orthogonal component by $\mathbf{y}_{\mathcal{C}}^{\perp}$.

Theorem 18 ($\mathcal{NU}$:Consistency under Lipschitz on h_n) Let $\{\mathbf{x}(i)\}_{i \geq 0}$ be the state sequence generated by the $\mathcal{NU}$ algorithm (Assumptions **(D.1)-(D.3)**). Further, $\forall \theta \neq \widetilde{\theta} \in \mathbb{R}^M, 1 \leq n \leq N$, let the functions $h_n(\cdot)$ be:

1) Lipschitz continuous with constants $k_n > 0$, i.e.,

$$\| h_n(\theta) - h_n(\widetilde{\theta}) \| \leq k_n \| \theta - \widetilde{\theta} \| \tag{50}$$

2)

$$\left(\theta - \widetilde{\theta} \right)^T \left(h_n(\theta) - h_n(\widetilde{\theta}) \right) \geq 0. \tag{51}$$

Define K as

$$K = \max(k_1, \cdots, k_N) \tag{52}$$

Then, for every $\beta > 0$, the estimate sequence is consistent. In other words,

$$\mathbb{P}_{\theta^*} \left[\lim_{i \to \infty} \mathbf{x}_n(i) = \theta^*,\ \forall n \right] = 1$$

The proof is in Appendix B. The conditions in (50)-(51) are much easier to verify than guessing a Lyapunov function. Also, as will be shown in the proof, the conditions in Theorem 18 determine a Lyapunov function explicitly, which may be used to analyze properties like convergence rate. The Lipschitz assumption is quite common in the stochastic approximation literature, while the assumption in (51) holds for a large class of functions. As a matter of fact, in the one-dimensional case ($M = 1$), it is satisfied if the functions $h_n(\cdot)$ are non-decreasing. Also, in general, it can be shown from the proof (Appendix B) that the Lipschitz continuity assumption in Theorem 18 may be replaced by continuity of the functions $h_n(\cdot),\ 1 \leq n \leq N$, and linear growth conditions, i.e., for constants $c_{n,1}, c_{n,2} > 0$,

$$\| h_n(\theta) \|^2 \leq c_{n,1} + c_{n,2} \| \theta \|^2, \forall \theta \in \mathbb{R}^M, 1 \leq n \leq N.$$

We now present another set of sufficient conditions that guarantee consistency of $\mathcal{NU}$. If the observation model is separably estimable, in some cases even if the underlying model is nonlinear, it may be possible to choose the functions, $g_n(\cdot)$, such that the function $h(\cdot)$ possesses nice properties. This is the next result.

Theorem 19 ($\mathcal{NU}$:Consistency–h strict monotonicity) Consider the $\mathcal{NU}$ algorithm (Assumptions **(D.1)-(D.3)**). Suppose that the functions $g_n(\cdot)$ can be chosen, such that the functions $h_n(\cdot)$ are Lipschitz continuous with constants $k_n > 0$ and the function $h(\cdot)$ satisfies

$$\left(\theta - \widetilde{\theta} \right)^T \left(h(\theta) - h(\widetilde{\theta}) \right) \geq \gamma \| \theta - \widetilde{\theta} \|^2,\ \forall \theta, \widetilde{\theta} \in \mathbb{R}^M \tag{53}$$

[9] This is because converse theorems in stability theory do not hold in general, see, [52].

for some constant $\gamma > 0$. Then, for

$$K = \max(k_1, \cdots, k_N),$$

if

$$\beta > \frac{K^2 + K\gamma}{\gamma\lambda_2\overline{L}},$$

the algorithm $\mathcal{NU}$ is consistent, i.e.,

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\mathbf{x}_n(i) = \theta^*,\ \forall n\right] = 1$$

The proof is provided in Appendix B. We comment that, in comparison to Theorem 18, strengthening the assumptions on $h(\cdot)$, see (53), considerably weakens the assumptions on the functions $h_n(\cdot)$. Eqn. (53) is an analog of strict monotonicity. For example, if $h(\cdot)$ is linear, the left hand side of (53) becomes a quadratic and the condition says that this quadratic is strictly away from zero, i.e., monotonically increasing with rate γ.

IV. Nonlinear Observation Models: Algorithm $\mathcal{NLU}$

In this Section, we present the algorithm $\mathcal{NLU}$ for distributed estimation in separably estimable observation models. As explained later, this is a mixed time-scale algorithm, where the consensus time-scale dominates the observation update time-scale as time progresses. The $\mathcal{NLU}$ algorithm is based on the fact that, for separably estimable models, it suffices to know $h(\theta^*)$, because θ^* can be unambiguously determined from the invertible function $h(\theta^*)$. To be precise, if the function $h(\cdot)$ has a continuous inverse, then any iterative scheme converging to $h(\theta^*)$ will lead to consistent estimates, obtained by inverting the sequence of iterates. The algorithm $\mathcal{NLU}$ is shown to yield consistent and unbiased estimators at each sensor for any separably observable model, under the assumption that the function $h(\cdot)$ has a continuous inverse. Thus, the algorithm $\mathcal{NLU}$ presents a more reliable alternative than the algorithm $\mathcal{NU}$, because, as shown in Section III-C, the convergence properties of the latter can be guaranteed only under certain assumptions on the observation model. We briefly comment on the organization of this section. The $\mathcal{NLU}$ algorithm for separably estimable observation models is presented in Section IV-A. Section IV-B offers interpretations of the $\mathcal{NLU}$ algorithm and presents the main results regarding consistency, mean-square convergence, asymptotic unbiasedness proved in the paper. In Section IV-C we prove the main results about the $\mathcal{NLU}$ algorithm and provide insights behind the analysis (in particular, why standard stochastic approximation results cannot be used directly to give its convergence properties.) Finally, Section V presents discussions on the $\mathcal{NLU}$ algorithm and suggests future research directions.

A. Algorithm $\mathcal{NLU}$

Algorithm $\mathcal{NLU}$: Let $\mathbf{x}(0) = [\mathbf{x}_1^T \cdots \mathbf{x}_N^T]^T$ be the initial set of states (estimates) at the sensors. The $\mathcal{NLU}$ generates the sequence $\{\mathbf{x}_n(i)\} \in \mathbb{R}^M$ at the n-th sensor according to the distributed recursive scheme:

$$\begin{aligned}\mathbf{x}_n(i+1) = h^{-1}\Bigg(& h(\mathbf{x}_n(i)) - \\ & - \beta(i)\left(\sum_{l\in\Omega_n(i)}\left[h(\mathbf{x}_n(i)) - \mathbf{q}\left(h(\mathbf{x}_l(i)) + \nu_{nl}(i)\right)\right]\right) \\ & - \alpha(i)\left[h(\mathbf{x}_n(i)) - g_n(\mathbf{z}_n(i))\right]\Bigg)\end{aligned} \tag{54}$$

based on the information,

$$\mathbf{x}_n(i), \left\{\mathbf{q}\left(h(\mathbf{x}_l(i)) + \nu_{nl}(i)\right)\right\}_{l\in\Omega_n(i)}, \mathbf{z}_n(i),$$

available to it at time i (we assume that at time i sensor l sends a quantized version of $h(\mathbf{x}_l(i)) + \nu_{nl}(i)$ to sensor n.) Here $h^{-1}(\cdot)$ denotes the inverse of the function $h(\cdot)$ and $\{\beta(i)\}_{i\geq 0}, \{\alpha(i)\}_{i\geq 0}$ are appropriately chosen weight sequences. In the sequel, we analyze the $\mathcal{NLU}$ algorithm under the model Assumptions **(D.1)-(D.3)**, and in addition we assume:

(D.4): There exists $\epsilon_1 > 0$, such that, $\forall \theta \in \mathcal{U}$, the following moment exists:

$$\mathbb{E}_\theta \left[\left\| J(\mathbf{z}(i)) - PJ(\mathbf{z}(i)) \right\|^{2+\epsilon_1} \right] = \kappa(\theta) < \infty,$$

where $J(\mathbf{z}(i))$ is defined in (45) and the matrix P is given by

$$\begin{aligned} P &= \frac{1}{N} \left(\mathbf{1}_N \otimes I_{\overline{M}} \right) \left(\mathbf{1}_N \otimes I_{\overline{M}} \right)^T \\ &= \frac{1}{N} \left(\mathbf{1}_N \mathbf{1}_N^T \right) \otimes I_{\overline{M}} \\ &= P_N \otimes I_{\overline{M}} \end{aligned}$$

The above moment condition is slightly stronger than the moment assumption required by the $\mathcal{NU}$ algorithm in (43), where only existence of the quadratic moment of the random variables $g_n(\mathbf{z}_n(i))$ was assumed. For example, for the models considered in Propositions 13-14, the above condition coincides with the existence of slightly higher than quadratic moments of the observations $\mathbf{z}_n(i)$. The latter is clearly justified for any reasonable observation noise distribution.

We also define, $\forall \theta \in \mathcal{U}$:

$$\begin{aligned} \mathbb{E}_\theta \left[\left\| J(\mathbf{z}(i)) - PJ(\mathbf{z}(i)) \right\| \right] &= \kappa_1(\theta) < \infty \\ \mathbb{E}_\theta \left[\left\| J(\mathbf{z}(i)) - PJ(\mathbf{z}(i)) \right\|^2 \right] &= \kappa_2(\theta) < \infty. \end{aligned} \tag{55}$$

(D.5): The weight sequences $\{\alpha(i)\}_{i \geq 0}$, and $\{\beta(i)\}_{i \geq 0}$ are given by

$$\alpha(i) = \frac{a}{(i+1)^{\tau_1}}, \quad \beta(i) = \frac{b}{(i+1)^{\tau_2}}$$

where $a, b > 0$ are constants. We assume the following:

$$.5 < \tau_1, \tau_2 \leq 1, \quad \tau_1 > \frac{1}{2+\epsilon_1} + \tau_2, \quad 2\tau_2 > \tau_1 \tag{56}$$

We note that, under Assumption **(D.4)**, $\epsilon_1 > 0$, such weight sequences always exist. As an example, if $\frac{1}{2+\epsilon_1} = .49$, then the choice $\tau_1 = 1$ and $\tau_2 = .505$ satisfies the inequalities in (56).

To write the $\mathcal{NLU}$ in a more compact form, introduce the *transformed* state sequence, $\{\widetilde{\mathbf{x}}(i)\}_{i \geq 0}$, where $\widetilde{\mathbf{x}}(i) = [\widetilde{\mathbf{x}}_1^T(i) \cdots \widetilde{\mathbf{x}}_N^T(i)]^T \in \mathbb{R}^{N\overline{M}}$ and the iterations are

$$\begin{aligned} \widetilde{\mathbf{x}}(i+1) = &\widetilde{\mathbf{x}}(i) - \beta(i) \left(L(i) \otimes I_M \right) \widetilde{\mathbf{x}}(i) - \\ &- \alpha(i) \left[\widetilde{\mathbf{x}}(i) - J(\mathbf{z}(i)) \right] - \beta(i) \left(\mathbf{\Upsilon}(i) + \mathbf{\Psi}(i) \right) \end{aligned} \tag{57}$$

$$\mathbf{x}(i) = \left[\left(h^{-1}(\widetilde{\mathbf{x}}_1(i)) \right)^T \cdots \left(h^{-1}(\widetilde{\mathbf{x}}_N(i)) \right)^T \right]^T \tag{58}$$

Here $\mathbf{\Upsilon}(i), \mathbf{\Psi}(i) \in \mathbb{R}^{N\overline{M}}$ model the dithered quantization error effects as in algorithm $\mathcal{NU}$ resulting from the quantized transmissions in (54). The update model in (57) is a mixed time-scale procedure, where the consensus time-scale is determined by the weight sequence $\{\beta(i)\}_{i \geq 0}$. On the other hand, the observation update time-scale is governed by the weight sequence $\{\alpha(i)\}_{i \geq 0}$. It follows from Assumption **(D.5)** that $\tau_1 > \tau_2$, which in turn implies, $\frac{\beta(i)}{\alpha(i)} \to \infty$ as $i \to \infty$. Thus, the consensus time-scale dominates the observation update time-scale as the algorithm progresses making it a mixed time-scale algorithm that does not directly fall under the purview of stochastic approximation results like Theorem 29. Also, the presence of the random link failures and quantization noise (which operate at the same time-scale as the consensus update) precludes standard approaches like time-scale separation for the limiting system.

Remark 20 We comment on the distributed implementation of the $\mathcal{NLU}$. Based on (54) and (57)-(58), the $\mathcal{NLU}$ may be implemented either in the estimate domain with $\{\mathbf{x}(i)\}$ as the algorithm state sequence or in the transformed domain with $\{\widetilde{\mathbf{x}}(i)\}$ as the state sequence. The implementation in the estimate domain, (54), would require the sensors to store and transmit the instantaneous estimate $\mathbf{x}_n(i)$, however, implementation of the update involves computation of the functions $h(\mathbf{x}_n(i))$ followed by an inverse $(h^{-1}(\cdot))$ at every step. On the other hand, the implementation in the transformed domain, (57)-(58), requires the sensors to store and transmit the states $\widetilde{\mathbf{x}}_n(i)$. The advantage in the latter implementation form is that the transformed state update rule, (57), is linear in the state $\widetilde{\mathbf{x}}(i)$ and, in particular, does not require function computations and inverses at each step. (We note that the inverse in (58) may not be implemented at all iterations and does not affect the propagation of the transformed state sequence. In fact, (58) may be implemented only once to obtain the actual estimates from the transformed state sequence when the latter converge based on a suitable stopping criterion.) Hence, in practice, to simplify computations, the $\mathcal{NLU}$ may be implemented in the transformed domain with $\widetilde{\mathbf{x}}_n(i)$ being the state at a sensor n.

B. Algorithm $\mathcal{NLU}$: Discussions and Main Results

We comment on the $\mathcal{NLU}$ algorithm. As is clear from (57)-(58), the $\mathcal{NLU}$ algorithm operates in a *transformed* domain. As a matter of fact, the function $h(\cdot)$ (c.f. definition 11) can be viewed as an invertible transformation on the parameter space $\mathcal{U}$. The transformed state sequence, $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$, is then a transformation of the estimate sequence $\{\mathbf{x}(i)\}_{i\geq 0}$, and, as seen from (57), the evolution of the sequence $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ is linear. This is an important feature of the $\mathcal{NLU}$ algorithm, which is linear in the transformed domain, although the underlying observation model is nonlinear. Intuitively, this approach can be thought of as a distributed stochastic version of homomorphic filtering (see [53]), where, by suitably transforming the state space, linear filtering is performed on a certain non-linear problem of filtering. In our case, for models of the separably estimable type, the function $h(\cdot)$ then plays the role of the analogous transformation in homomorphic filtering, and, in this transformed space, one can design linear estimation algorithms with desirable properties. This makes the $\mathcal{NLU}$ algorithm significantly different from algorithm $\mathcal{NU}$, with the latter operating on the untransformed space and is non-linear. This linear property of the $\mathcal{NLU}$ algorithm in the transformed domain leads to nice statistical properties (for example, consistency asymptotic unbiasedness) under much weaker assumptions on the observation model than required by the nonlinear $\mathcal{NU}$ algorithm, but not asymptotic normality.

We now state the main results about the $\mathcal{NLU}$ algorithm developed in the paper. We show that, if the observation model is separably estimable, then, in the transformed domain, the $\mathcal{NLU}$ algorithm is consistent. More specifically, if θ^* is the true (but unknown) parameter value, then the transformed sequence $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ converges a.s. and in mean-squared sense to $h(\theta^*)$. We note that, unlike the $\mathcal{NU}$ algorithm, this only requires the observation model to be separably estimable and no other conditions on the functions $h_n(\cdot), h(\cdot)$. We summarize these in the following theorem.

Theorem 21 Consider the $\mathcal{NLU}$ algorithm under the Assumptions **(D.1)-(D.5)**, and the sequence $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ generated according to (57). We then have

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\widetilde{\mathbf{x}}_n(i)=h(\theta^*),\ \forall 1\leq n\leq N\right]=1 \tag{59}$$

$$\lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\|\widetilde{\mathbf{x}}_n(i)-h(\theta^*)\|^2\right]=0,\ \forall 1\leq n\leq N \tag{60}$$

In particular,

$$\lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\widetilde{\mathbf{x}}_n(i)\right]=h(\theta^*),\quad \forall 1\leq n\leq N$$

In other words, in the transformed domain, the estimate sequence $\{\widetilde{\mathbf{x}}_n(i)\}_{i\geq 0}$ at sensor n, is consistent, asymptotically unbiased and converges in mean-squared sense to $h(\theta^*)$.

As an immediate consequence of Theorem 21, we have the following result, which characterizes the statistical properties of the untransformed state sequence $\{\mathbf{x}(i)\}_{i\geq 0}$.

Theorem 22 Consider the $\mathcal{NLU}$ algorithm under the Assumptions **(D.1)-(D.5)**. Let $\{\mathbf{x}(i)\}_{i\geq 0}$ be the state sequence generated, as given by (57)-(58). We then have

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\mathbf{x}_n(i)=\theta^*,\ \forall\ 1\leq n\leq N\right]=1$$

In other words, the $\mathcal{NLU}$ algorithm is consistent.

If in addition, the function $h^{-1}(\cdot)$ is Lipschitz continuous, the $\mathcal{NLU}$ algorithm is asymptotically unbiased, i.e.,

$$\lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\mathbf{x}_n(i)\right]=\theta^*,\ \forall\ 1\leq n\leq N$$

The next subsection is concerned with the proofs of Theorems 21, 22.

C. Consistency and Asymptotic Unbiasedness of $\mathcal{NLU}$: Proofs of Theorems 21,22

The present subsection is devoted to proving the consistency and unbiasedness of the $\mathcal{NLU}$ algorithm under the stated Assumptions. The proof is lengthy and we start by explaining why standard stochastic approximation results like Theorem 29 do not apply directly. A careful inspection shows that there are essentially two different time-scales embedded in (57). The consensus time-scale is determined by the weight sequence $\{\beta(i)\}_{i\geq 0}$, whereas the observation update time-scale is governed by the weight sequence $\{\alpha(i)\}_{i\geq 0}$. It follows from Assumption **(D.5)** that $\tau_1 > \tau_2$, which, in turn, implies $\frac{\beta(i)}{\alpha(i)} \to \infty$ as $i \to \infty$. Thus, the consensus time-scale dominates the observation update time-scale as the algorithm progresses making it a mixed time-scale algorithm that does not directly fall under the purview of stochastic approximation results like Theorem 29. Also, the presence of the random link failures and quantization noise (which operate at the same time-scale as the consensus update) precludes standard approaches like time-scale separation for the limiting system.

Finally, we note that standard stochastic approximation assumes that the state evolution follows a stable deterministic system perturbed by *zero-mean* stochastic noise. More specifically, if $\{\mathbf{y}(i)\}_{i\geq 0}$ is the sequence of interest, Theorem 29 assumes that $\{\mathbf{y}(i)\}_{i\geq 0}$ evolves as

$$\mathbf{y}(i+1)=\mathbf{y}(i)+\gamma(i)[R(\mathbf{y}(i))+\Gamma(i+1,\omega,\mathbf{y}(i))] \tag{61}$$

where $\{\gamma(i)\}_{i\geq 0}$ is the weight sequence, $\Gamma(i+1,\omega,\mathbf{y}(i))$ is the *zero-mean* noise. If the sequence $\{\mathbf{y}(i)\}_{i\geq 0}$ is supposed to converge to $\mathbf{y}_0$, it further assumes that $R(\mathbf{y}_0)=\mathbf{0}$ and $\mathbf{y}_0$ is a stable equilibrium of the deterministic system

$$\mathbf{y}_d(i+1)=\mathbf{y}_d(i)+\gamma(i)R(\mathbf{y}_d(i))$$

The $\mathcal{NU}$ algorithm (and its linear version, $\mathcal{LU}$) falls under the purview of this, and we can establish convergence properties using standard stochastic approximation (see Sections II,III-B.) However, the $\mathcal{NLU}$ algorithm cannot be represented in the form of (61), even ignoring the presence of multiple time-scales. Indeed, as established by Theorem 21, the sequence $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ is supposed to converge to $\mathbf{1}_N\otimes h(\theta^*)$ a.s. and hence writing (57) as a stochastically perturbed system around $\mathbf{1}_N\otimes h(\theta^*)$ we have

$$\widetilde{\mathbf{x}}(i+1)=\widetilde{\mathbf{x}}(i)+\gamma(i)\left[R(\widetilde{\mathbf{x}}(i))+\Gamma(i+1,\omega,\widetilde{\mathbf{x}}(i))\right]$$

where,

$$\begin{aligned}R(\widetilde{\mathbf{x}}(i)) &= -\beta(i)\left(\overline{L}\otimes I_M\right)\left(\widetilde{\mathbf{x}}(i)-\mathbf{1}_N\otimes h(\theta^*)\right)-\\ &\quad -\alpha(i)\left(\widetilde{\mathbf{x}}(i)-\mathbf{1}_N\otimes h(\theta^*)\right)\\ \Gamma(i+1,\omega,\widetilde{\mathbf{x}}(i)) &= -\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\left(\widetilde{\mathbf{x}}(i)-\right.\\ &\quad \left.-\mathbf{1}_N\otimes h(\theta^*)\right)-\beta(i)\left(\boldsymbol{\Upsilon}(i)+\boldsymbol{\Psi}(i)\right)+\\ &\quad +\alpha(i)\left(J(\mathbf{z}(i))-\mathbf{1}_N\otimes h(\theta^*)\right)\end{aligned}$$

Although, $R(\mathbf{1}_N\otimes h(\theta^*))=\mathbf{0}$ in the above decomposition, the noise $\Gamma(i+1,\omega,\widetilde{\mathbf{x}}(i))$ is not unbiased as the term $(J(\mathbf{z}(i))-\mathbf{1}_N\otimes h(\theta^*))$ is *not* zero-mean.

With the above discussion in mind, we proceed to the proof of Theorems 21,22, which we develop in stages. The detailed proofs of the intermediate results are provided in Appendix F.

In parallel to the evolution of the state sequence $\{\mathbf{x}(i)\}_{i\geq 0}$, we consider the following update of the auxiliary sequence, $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$:

$$\begin{aligned}\widetilde{\mathbf{x}}^{\circ}(i+1) &= \widetilde{\mathbf{x}}^{\circ}(i)-\beta(i)\left(\overline{L}\otimes I_M\right)\widetilde{\mathbf{x}}^{\circ}(i)-\\ &\quad -\alpha(i)\left[\widetilde{\mathbf{x}}^{\circ}(i)-J(\mathbf{z}(i))\right]\end{aligned} \tag{62}$$

with $\widetilde{\mathbf{x}}^{\circ}(0)=\widetilde{\mathbf{x}}(0)$. Note that in (62) the random Laplacian L is replaced by the average Laplacian $\overline{L}$ and the quantization noises $\boldsymbol{\Upsilon}(i)$ and $\boldsymbol{\Psi}(i)$ are not included. In other words, in the absence of link failures and quantization, the recursion (57) reduces to (62), i.e., the sequences $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ and $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$ are the same.

Now consider the sequence whose recursion adds as input to the recursion in (62) the quantization noises $\mathbf{\Upsilon}(i)$ and $\mathbf{\Psi}(i)$. In other words, in the absence of link failures, but with quantization included, define similarly the sequence $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$ given by

$$\begin{aligned}\widehat{\mathbf{x}}(i+1) = \widehat{\mathbf{x}}(i) - \beta(i)\left(\overline{L}\otimes I_M\right)\widehat{\mathbf{x}}(i) - \\ - \alpha(i)\left[\widehat{\mathbf{x}}(i) - J(\mathbf{z}(i))\right] - \beta(i)\left(\mathbf{\Upsilon}(i) + \mathbf{\Psi}(i)\right)\end{aligned} \tag{63}$$

with $\widehat{\mathbf{x}}(0) = \widetilde{\mathbf{x}}(0)$. Like before, the recursions (57,58) will reduce to (63) when there are no link failures. However, notice that in (63) the quantization noise sequences $\mathbf{\Upsilon}(i)$ and $\mathbf{\Psi}(i)$ are the sequences resulting from quantizing $\widetilde{\mathbf{x}}(i)$ in (57) and not from quantizing $\widehat{\mathbf{x}}(i)$ in (63).

Define the instantaneous averages over the network as

$$\begin{aligned}\mathbf{x}_{\text{avg}}(i) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n(i) = \frac{1}{N}\left(\mathbf{1}_N\otimes I_M\right)^T\mathbf{x}(i)\\ \widetilde{\mathbf{x}}_{\text{avg}}(i) &= \frac{1}{N}\sum_{n=1}^{N}\widetilde{\mathbf{x}}_n(i) = \frac{1}{N}\left(\mathbf{1}_N\otimes I_M\right)^T\widetilde{\mathbf{x}}(i)\\ \mathbf{x}^{\circ}_{\text{avg}}(i) &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{\circ}_n(i) = \frac{1}{N}\left(\mathbf{1}_N\otimes I_M\right)^T\mathbf{x}^{\circ}(i)\\ \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) &= \frac{1}{N}\sum_{n=1}^{N}\widetilde{\mathbf{x}}^{\circ}_n(i) = \frac{1}{N}\left(\mathbf{1}_N\otimes I_M\right)^T\widetilde{\mathbf{x}}^{\circ}(i)\end{aligned} \tag{64}$$

We sketch the main steps of the proof here. While proving consistency and mean-squared sense convergence, we first show that the average sequence, $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$, converges a.s. to $h(\theta^*)$. This can be done by invoking standard stochastic approximation arguments. Then we show that the sequence $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$ reaches consensus a.s., and clearly the limiting consensus value must be $h(\theta^*)$. Intuitively, the a.s. consensus comes from the fact that, after a sufficiently large number of iterations, the consensus effect dominates over the observation update effect, thus asymptotically leading to consensus. The final step in the proof uses a series of comparison arguments to show that the sequence $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ also reaches consensus a.s. with $h(\theta^*)$ as the limiting consensus value.

We now detail the proofs of Theorems 21,22 in the following steps.

(I): The first step consists of studying the convergence properties of the sequence $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$, see (62), for which we establish the following result.

Lemma 23 Consider the sequence, $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$, given by (62), under the Assumptions **(D.1)-(D.5)**. Then,

$$\begin{aligned}\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\widetilde{\mathbf{x}}^{\circ}(i) = \mathbf{1}_N\otimes h(\theta^*)\right] &= 1\\ \lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\|\widetilde{\mathbf{x}}^{\circ}(i) - \mathbf{1}_N\otimes h(\theta^*)\|^2\right] &= 0\end{aligned}$$

Lemma 23 says that the sequence $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$ converges a.s. and in $\mathcal{L}_2$ to $\mathbf{1}_N\otimes h(\theta^*)$. For proving Lemma 23 we first consider the corresponding average sequence $\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\}_{i\geq 0}$, see (64). For the sequence $\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\}_{i\geq 0}$, we can invoke stochastic approximation algorithms to prove that it converges a.s. and in $\mathcal{L}_2$ to $h(\theta^*)$. This is carried out in Lemma 24, which we state now.

Lemma 24 Consider the sequence, $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$, given by (64), under the Assumptions **(D.1)-(D.5)**. Then,

$$\begin{aligned}\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) = h(\theta^*)\right] &= 1\\ \lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - h(\theta^*)\right\|^2\right] &= 0\end{aligned}$$

The arguments in Lemmas 24,23 and subsequent results require the following property of real number sequences, which we state here (see Appendix C for proof.)

Lemma 25 Let the sequences $\{r_1(i)\}_{i\geq 0}$ and $\{r_2(i)\}_{i\geq 0}$ be given by

$$r_1(i)=\frac{a_1}{(i+1)^{\delta_1}},\quad r_2(i)=\frac{a_2}{(i+1)^{\delta_2}} \tag{69}$$

where $a_1, a_2, \delta_2 \geq 0$ and $0 \leq \delta_1 \leq 1$. Then, if $\delta_1 = \delta_2$, there exists $B > 0$, such that, for sufficiently large non-negative integers, $j < i$,

$$0 \leq \sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-r_1(l))\right)r_2(k)\right] \leq B \tag{70}$$

Moreover, the constant B can be chosen independently of i, j. Also, if $\delta_1 < \delta_2$, then, for arbitrary fixed j,

$$\lim_{i\to\infty}\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-r_1(l))\right)r_2(k)\right]=0$$

(We use the convention that, $\prod_{l=k+1}^{i-1}(1-r_l)=1$, for $k = i-1$.)

We note that Lemma 25 essentially studies stability of time-varying deterministic scalar recursions of the form:

$$y(i+1) = r_1(i)y(i) + r_2(i) \tag{71}$$

where $\{y(i)\}_{i\geq 0}$ is a scalar sequence evolving according to (71) with $y(0) = 0$, and the sequences $\{r_1(i)\}_{i\geq 0}$ and $\{r_2(i)\}_{i\geq 0}$ are given by (69).

(II): In this step, we study the convergence properties of the sequence $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$, see (63), for which we establish the following result.

Lemma 26 Consider the sequence $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$ given by (63) under the Assumptions **(D.1)-(D.5)**. We have

$$\begin{aligned}\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\widehat{\mathbf{x}}(i)=\mathbf{1}_N\otimes h(\theta^*)\right] &= 1\\ \lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\left\|\widehat{\mathbf{x}}(i)-\mathbf{1}_N\otimes h(\theta^*)\right\|^2\right] &= 0\end{aligned}$$

The proof of Lemma 26 is given in Appendix E, and mainly consists of a comparison argument involving the sequences $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$ and $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$.

(III): This is the final step in the proofs of Theorems 21,22. The proof of Theorem 21 consists of a comparison argument between the sequences $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$ and $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$, which is detailed in Appendix F. The proof of Theorem 22, also detailed in Appendix F, is a consequence of Theorem 21 and the Assumptions.

D. Application: Distributed Static Phase Estimation in Smart Grids

In this subsection we show that our development of the $\mathcal{NLU}$ for separably estimable observation models leads to a completely distributed solution of the static phase estimation problem in smart grids. We briefly review the application scenario in the following, for a more complete treatment of the classical problem of static phase estimation in power grids the reader is referred to one of the many existing excellent textbooks, for example, [54]. For our purpose, we may assume the power grid to be a physical network of N generators and loads (hereafter called nodes), interconnected through transmission lines. The physical grid may then be modeled as a network $G_p = (V, E_p)$, where[10] E_p denotes the set of transmission lines or interconnections. The physical state of a node n consists of the pair $(\mathcal{V}_n, \theta_n)$, denoting the voltage magnitude and the phase angle respectively. The real power flowing through the transmission line connecting nodes n and l is then given by (see [55])

$$\mathcal{P}_{nl}=\mathcal{V}_n^2 a_{nl}-\mathcal{V}_n\mathcal{V}_l a_{nl}\cos(\theta_{nl})+\mathcal{V}_n\mathcal{V}_l b_{nl}\sin(\theta_{nl}) \tag{74}$$

[10] Note that the physical connections E_p and the inter-node communication links E (to be used by the distributed information processing algorithms) are, in general, different. This is a common feature of cyberphysical architectures, where a sensor network is instrumented on top of an existing physical infrastructure. In the following, we will assume that each physical node is equipped with a sensor for information processing, although the inter-sensor communication topology may be different from the physical inter-node connections. Also, the terms nodes and sensors will be used synonymously in the sequel.

where $a_{nl} + jb_{nl}$ is the complex line admittance and $\theta_{nl} = \theta_n - \theta_l$. In view of the physical network structure, the following assumptions on the physical grid are supposed to hold:

(P.1) The physical grid, represented by the graph (V, E_p), is connected. This is reasonable, as the physical power grid is often studied with aggregated models that have dense interconnections, see, for example, the benchmark IEEE 30 bus and IEEE 118 bus systems ([55]).

(P.2) The real and imaginary parts, a_{nl}, b_{nl}, of the line admittance are taken to be zero, if no direct physical connection (link) exists between the nodes n and l. Similarly, if nodes n and l are connected by a transmission line, we assume both the components a_{nl}, b_{nl}, of the line admittance to be non-zero. In particular, from (74), the real power flow $\mathcal{P}_{nl}$ between nodes n and l is non-zero *iff* there exists a physical transmission line connecting the nodes. Also, $\mathcal{V}_n \neq 0$ for all n.

In the following, we will assume that the node voltages are known constants and the unknown parameter of interest is the vector of node phases $\theta = [\theta_1, \cdots, \theta_N]^T$. This is justified by the common assumption of phase-voltage decoupling in power grids, where the voltage magnitude is generally seen to fluctuate at a much slower time-scale than the phase (see [54]). Also, to keep the exposition simple, we assume that the node phase differences are small, i.e., $\cos(\theta_{nl}) \approx 1$, commonly used in the steady state grid operating regime ([54]). With these simplifications, the real power flow in a transmission line connecting nodes n and l is approximately given by

$$\mathcal{P}_{nl}(\theta) = \mathcal{V}_n \mathcal{V}_l b_{nl} \sin(\theta_{nl})$$

The goal of centralized static phase estimation is to estimate the unknown vector θ of phases by using line flow data. Since, only relative quantities (phase differences) are involved in this problem, it is customary to assume (see [55]) that one of the nodes is a slack (or reference) bus, whose phase is a known constant. W.l.o.g. we assume that node N is the slack bus in our system, whose phase angle θ_N is a known constant. Hence, the effective parameter vector is $\theta = [\theta_1, \cdots, \theta_{N-1}]^T \in \mathbb{R}^{N-1}$. We now provide conditions for the distributed observation model (77) to be separably estimable. We show that our $\mathcal{NLU}$ algorithm can be used to obtain a distributed solution to this problem, leading to a consistent estimate of θ at each sensor. To setup the distributed observation model, let $E_m \subset E_p$ denote the set of physical transmission lines equipped with power flow measuring devices (usually some form of relays, see [55].) The successive power flow measurements, $\{z_{nl}(i)\}$ at the physical line (n, l) (assuming $(n, l) \in E_m$) are then noisy versions of the power flow $\mathcal{P}_{nl}$, i.e.,

$$\begin{aligned} z_{nl}(i) &= \mathcal{P}_{nl}(\theta) + \zeta_{nl}(i) \\ &= \mathcal{V}_n \mathcal{V}_l b_{nl} \sin(\theta_{nl}) + \zeta_{nl}(i) \end{aligned} \tag{75}$$

where, $\{\zeta_{nl}(i)\}$ is the zero mean i.i.d. measurement noise. For distributed information processing, we assume that the measurement sequence $\{z_{nl}(i)\}$ is forwarded to one of the adjacent nodes n or l. As will be seen, the particular choice of n or l is not important, as long as it stays constant for all i. In general, denoting by Ω_n^m to be the set of physical neighbors of node n w.r.t. the physical graph (V, E_m), let $\mathcal{M}_n \subset \Omega_n^m$ denote the set such that $l \in \mathcal{M}_n$ if the line flow data $\{z_{nl}(i)\}$ is forwarded to node n. The observation sequence, $\{\mathbf{z}_n(i)\}$, is then

$$\mathbf{z}_n(i) = \{z_{nl}(i), \ \forall \ l \in \mathcal{M}_n\} \tag{76}$$

By (75)-(76) and the development in Section III-A, the above distributed observation model corresponds to the signal in additive noise type. Following the notation in Section III-A, the observation process $\{\mathbf{z}_n(i)\}$ may be written as

$$\mathbf{z}_n(i) = f_n(\theta) + \zeta_n(i) \tag{77}$$

where $f_n : \mathcal{U} \longmapsto \mathbb{R}^{|\mathcal{M}_n|}$ and $\zeta_n(i)$ are given by

$$\begin{aligned} f_n(\theta) &= [\mathcal{V}_n \mathcal{V}_l b_{nl} \sin(\theta_{nl}), \ l \in \mathcal{M}_n]^T \\ \zeta_n(i) &= [\zeta_{nl}, \ l \in \mathcal{M}_n]^T. \end{aligned}$$

We now provide conditions for the distributed observation model (77) to be separably estimable.

Proposition 27 Depending on the phase θ_N of the reference bus N (as to whether $\theta_N \in [0, \pi/2)$ or $\theta_N \in [-\pi/4, \pi/4)$), let the parameter domain $\mathcal{U} \subset \mathbb{R}^{N-1}$ be either $\times_{n=1}^{N-1}[0, \pi/2)$ or $\times_{n=1}^{N-1}[-\pi/4, \pi/4)$. Define, $\forall \theta \in \mathcal{U}$, $f : \mathcal{U} \longmapsto \mathbb{R}^{\sum_{n=1}^{N} |\mathcal{M}_n|}$ by

$$f(\theta) = [f_1^T(\theta) \cdots f_N^T(\theta)]^T = [\mathcal{P}_{nl}(\theta), (n, l) \in E_m]^T. \tag{78}$$

Then, if the graph (V, E_m) (with E_m as the edge set) is connected and assumption **(P.2)** holds, $f(\cdot)$ is invertible on $\mathcal{U}$ and the observation model (77) is separably estimable.

Before proceeding to the proof we comment on Proposition 27. The observation model in (77) is of the signal in additive noise type and hence, by the development in Section III-A, the invertibility of $f(\cdot)$ is equivalent to separable estimability and necessary for the consistency or observability of the centralized estimator also (see the text following Proposition 13.) Proposition 27 shows that the invertibility of $f(\cdot)$ holds if the graph formed by the physical transmission links equipped with power flow measuring devices is connected.

Proof: The proof is based on a simple inductive argument. First, we note that the continuity of $f(\cdot)$ on $\mathcal{U}$ holds trivially. To establish the invertibility of $f(\cdot)$ on $\mathcal{U}$, it suffices to show that we can uniquely recover the value of $\theta \in \mathcal{U}$ given the value $f(\theta)$. To this end, assume that $f(\theta)$ is given. Recall the form of $f(\theta)$, as in (78). Since, θ_N is known, given $f(\theta)$, the components θ_n, $n \in \Omega_N^m$ may be uniquely determined. Indeed, knowing $f(\theta)$ amounts to knowing the values of the quantities $\mathcal{P}_{n,N}$, $n \in \Omega_N^m$. Hence, under assumption **(P.2)**, and the fact that $\theta \in \mathcal{U}$, we can uniquely determine θ_n, $n \in \Omega_N^m$. To continue the induction, define $\mathcal{J}_1 \subset V$ by $\mathcal{J}_1 = \Omega_N^m$. Once the components θ_n, $n \in \mathcal{J}_1$ are known, by using similar reasoning, the components θ_l, $l \in \Omega_n^m$ for each $n \in \mathcal{J}_1$ may be uniquely determined. Hence, in the second step, the set of known components $\mathcal{J}_2 \subset V$ is

$$\mathcal{J}_2 = \{l \in \Omega_n^m \mid n \in \mathcal{J}_1\}$$

Continuing the same recursion, the set of components known at the k-th step, $\mathcal{J}_k$ is given by

$$\mathcal{J}_k = \{l \in \Omega_n^m \mid n \in \mathcal{J}_{k-1}\}$$

Note that $\mathcal{J}_1 \subset \mathcal{J}_2 \subset \cdots \subset \mathcal{J}_k \subset \cdots$. However, the number of nodes is finite; hence, the sets $\mathcal{J}_k$ cannot increase forever. Due to the connectivity of the graph (V, E_m) in a finite number of steps k_0 (at most equal to the diameter of the graph), the process will converge with $\mathcal{J}_{k_0} = V$. Hence, all components of θ will be uniquely determined, establishing the invertibility of $f(\cdot)$. ■

The following result demonstrates the applicability of the $\mathcal{NLU}$ to distributed consistent phase estimation in power grids. It follows from Theorem 22.

Theorem 28 Consider the power grid described above and let the observation process $\{\mathbf{z}_n(i)\}$ at the n-th node (sensor) be given by (77). Let the measurement noise process $\{\zeta(i)\}$ satisfy the moment conditions **(D.4)** and the physical grid conditions in the hypothesis of Proposition 27 hold. Suppose the nodes are instrumented with a communication architecture satisfying assumptions **(D.2)-(D.3)**. Let $\{\mathbf{x}_n(i)\}$ be the estimate sequence of the vector θ^* of phase angles generated at node n by an instantiation of the $\mathcal{NLU}$ distributed parameter estimation algorithm under assumption **(D.5)**. Then,

$$\mathbb{P}_{\theta^*}\left(\mathbf{x}_n(i) = \theta^*, \quad \forall n\right) = 1, \quad \forall\ \theta^* \in \mathcal{U}$$

V. Conclusion

This paper studies linear and nonlinear *distributed* (vector) parameter estimation problems as may arise in constrained sensor networks. Our problem statement is quite general, including communication among sensors that is quantized, noisy, and with channels that fail at random times. These are characteristic of packet communication in wireless sensor networks. We introduce a generic observability condition, the separable estimability condition, that generalizes to distributed estimation the general observability condition of centralized parameter estimation. We study three recursive distributed estimators, $\mathcal{ALU}$, $\mathcal{NU}$, and $\mathcal{NLU}$. We study their asymptotic properties, namely: consistency, asymptotic unbiasedness, and for the $\mathcal{ALU}$ and $\mathcal{NU}$ algorithms their asymptotic normality. The $\mathcal{NLU}$ works in a transformed domain where the recursion is actually linear, and a final nonlinear transformation, justified by the separable estimability condition, recovers the parameter estimate (a stochastic generalization of homeomorphic filtering.) For example, Theorem 21 shows that, in the transformed domain, the $\mathcal{NLU}$ leads to consistent and asymptotically unbiased estimators at every sensor for all separably estimable observation models satisfying **(D.4)**[11]. Since, the function $h(\cdot)$ is invertible, for practical purposes, a knowledge of $h(\theta^*)$ is sufficient for knowing θ^*. In that respect, the algorithm $\mathcal{NLU}$ is much more applicable than the algorithm $\mathcal{NU}$, which requires

[11] The $\mathcal{NLU}$ requires a slightly stronger moment condition, Assumption **(D.4)**. However, for reasonable observation noise statistics arising in practice, this assumption is justified.

further assumptions on the observation model for the existence of consistent and asymptotically unbiased estimators. However, in case, the algorithm $\mathcal{NU}$ is applicable, it provides convergence rate guarantees (for example, asymptotic normality) that follow from standard stochastic approximation theory. On the other hand, the algorithm $\mathcal{NLU}$ does not fall under the purview of standard stochastic approximation theory (see Section IV-C) and hence does not inherit these convergence rate properties. In this paper, we presented a convergence theory of the three algorithms under broad conditions. An interesting future research direction is to establish a convergence rate theory for the $\mathcal{NLU}$ algorithm (and in general, distributed stochastic algorithms of this form, which involve mixed time-scale behavior and biased perturbations.)

Appendix A
Some Results on Stochastic Approximation

We present some classical results on stochastic approximation from [56] regarding the convergence properties of generic stochastic recursive procedures, which will be used to characterize the convergence properties (consistency, convergence rate) of the $\mathcal{LU}$ algorithm.

Theorem 29 Let $\left\{\mathbf{x}(i)\in\mathbb{R}^l\right\}_{i\geq 0}$ be a random sequence:

$$\mathbf{x}(i+1)=\mathbf{x}(i)+\alpha(i)\left[R(\mathbf{x}(i))+\Gamma\left(i+1,\mathbf{x}(i),\omega\right)\right] \tag{79}$$

where, $R(\cdot):\mathbb{R}^l\longmapsto\mathbb{R}^l$ is Borel measurable and $\left\{\Gamma(i,\mathbf{x},\omega)\right\}_{i\geq 0,\ \mathbf{x}\in\mathbb{R}^l}$ is a family of random vectors in $\mathbb{R}^l$, defined on a probability space $(\Omega,\mathcal{F},\mathcal{P})$, and $\omega\in\Omega$ is a canonical element. Let the following sets of assumptions hold:

(B.1): The function $\Gamma(i,\cdot,\cdot):\mathbb{R}^l\times\Omega\longrightarrow\mathbb{R}^l$ is $\mathcal{B}^l\otimes\mathcal{F}$ measurable for every i; $\mathcal{B}^l$ is the Borel algebra of $\mathbb{R}^l$.

(B.2): There exists a filtration $\left\{\mathcal{F}_i\right\}_{i\geq 0}$ of $\mathcal{F}$, such that, for each i, the family of random vectors $\left\{\Gamma\left(i,\mathbf{x},\omega\right)\right\}_{\mathbf{x}\in\mathbb{R}^l}$ is $\mathcal{F}_i$ measurable, zero-mean and independent of $\mathcal{F}_{i-1}$.

(If Assumptions **(B.1)-(B.2)** hold, $\left\{\mathbf{x}(i)\right\}_{i\geq 0}$, is Markov.)

(B.3): There exists a function $V\left(\mathbf{x}\right)\in\mathbb{C}_2$ with bounded second order partial derivatives and a point $\mathbf{x}^*\in\mathbb{R}^l$ satisfying:

$$V\left(\mathbf{x}^*\right)=0,\ \ V\left(\mathbf{x}\right)>0,\,\mathbf{x}\neq\mathbf{x}^*,\ \ \lim_{\|\mathbf{x}\|\rightarrow\infty}V\left(\mathbf{x}\right)=\infty,$$
$$\sup_{\epsilon<\|\mathbf{x}-\mathbf{x}^*\|<\frac{1}{\epsilon}}\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)<0,\ \forall\epsilon>0$$

where $V_{\mathbf{x}}\left(\mathbf{x}\right)$ denotes the gradient (vector) of $V(\cdot)$ at $\mathbf{x}$.

(B.4): There exist constants $k_1,k_2>0$, such that,

$$\left\|R\left(\mathbf{x}\right)\right\|^2+\mathbb{E}\left[\left\|\Gamma\left(i+1,\mathbf{x},\omega\right)\right\|^2\right]\leq k_1\left(1+V\left(\mathbf{x}\right)\right)- \\ -k_2\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)$$

(B.5): The weight sequence $\left\{\alpha(i)\right\}_{i\geq 0}$ satisfies

$$\alpha(i)>0,\ \sum_{i\geq 0}\alpha_i=\infty,\ \sum_{i\geq 0}\alpha^2(i)<\infty$$

(C.1): The function $R\left(\mathbf{x}\right)$ admits the representation

$$R\left(\mathbf{x}\right)=B\left(\mathbf{x}-\mathbf{x}^*\right)+\delta\left(\mathbf{x}\right) \tag{80}$$

where

$$\lim_{\mathbf{x}\rightarrow\mathbf{x}^*}\frac{\left\|\delta\left(\mathbf{x}\right)\right\|}{\left\|\mathbf{x}-\mathbf{x}^*\right\|}=0 \tag{81}$$

(Note, in particular, if $\delta\left(\mathbf{x}\right)\equiv 0$, then (81) is satisfied.)

(C.2): The weight sequence, $\left\{\alpha(i)\right\}_{i\geq 0}$ is of the form,

$$\alpha(i)=\frac{a}{i+1},\ \ \forall i\geq 0 \tag{82}$$

where $a > 0$ is a constant (note that **(C.2)** implies **(B.5)**).

(C.3): Let I be the $l \times l$ identity matrix and a, B as in (82) and (80), respectively. Then, the matrix $\Sigma = aB + \frac{1}{2}I$ is stable.

(C.4): The entries of the matrices, $\forall i \geq 0, x \in \mathbf{R}^l$,

$$A\left(i, \mathbf{x}\right) = \mathbb{E}\left[\Gamma\left(i+1, \mathbf{x}, \omega\right)\Gamma^T\left(i+1, \mathbf{x}, \omega\right)\right],$$

are finite, and the following limit exists:

$$\lim_{i \to \infty,\ \mathbf{x} \to \mathbf{x}^*} A\left(i, \mathbf{x}\right) = S_0$$

(C.5): There exists $\epsilon > 0$, such that

$$\lim_{R \to \infty} \sup_{\|\mathbf{x} - \mathbf{x}^*\| < \epsilon} \sup_{i \geq 0} \int_{\|\Gamma(i+1, \mathbf{x}, \omega)\| > R} \left\|\Gamma\left(i+1, \mathbf{x}, \omega\right)\right\|^2 dP = 0 \tag{83}$$

Then we have the following:

Let Assumptions **(B.1)-(B.5)** hold for $\{\mathbf{x}(i)\}_{i \geq 0}$ in (79). Then, starting from an arbitrary initial state, the Markov process, $\{\mathbf{x}(i)\}_{i \geq 0}$, converges a.s. to $\mathbf{x}^*$. In other words,

$$\mathbb{P}\left[\lim_{i \to \infty} \mathbf{x}(i) = \mathbf{x}^*\right] = 1$$

The normalized process, $\left\{\sqrt{i}\left(\mathbf{x}(i) - \mathbf{x}^*\right)\right\}_{i \geq 0}$, is asymptotically normal if, besides Assumptions **(B.1)-(B.5)**, Assumptions **(C.1)-(C.5)** are also satisfied. In particular, as $i \to \infty$

$$\sqrt{i}\left(\mathbf{x}(i) - \mathbf{x}^*\right) \Longrightarrow \mathcal{N}(\mathbf{0}, S) \tag{84}$$

where $\Longrightarrow$ denotes convergence in distribution (weak convergence.) Also, the asymptotic variance, S, in (84) is

$$S = a^2 \int_0^\infty e^{\Sigma v} S_0 e^{\Sigma^T v} dv$$

Proof: For a proof see [56] (c.f. Theorems 4.4.4, 6.6.1). ■

APPENDIX B
PROOFS OF THEOREMS 18,19

Proof of Theorem 18

Proof: We noted the recursive scheme (44) satisfies Assumptions **(B.1)-(B.2)** of Theorem 29. To prove consistency, we verify Assumptions **(B.3)-(B.4)**. Let the Lyapunov function

$$V\left(\mathbf{x}\right) = \left\|\mathbf{x} - \mathbf{1}_N \otimes \theta^*\right\|^2$$

Clearly,

$$V(\mathbf{1}_N \otimes \theta^*) = 0, V\left(\mathbf{x}\right) > 0, \mathbf{x} \neq \mathbf{1}_N \otimes \theta^*, \lim_{\|\mathbf{x}\| \to \infty} V\left(\mathbf{x}\right) = \infty$$

By Assumptions (50)-(51), $h(\cdot)$ is Lipschitz continuous and

$$\left(\theta - \widetilde{\theta}\right)^T \left(h(\theta) - h(\widetilde{\theta})\right) > 0, \quad \forall\ \theta \neq \widetilde{\theta} \in \mathbb{R}^M \tag{85}$$

where (85) follows from the invertibility of $h(\cdot)$ and

$$h\left(\theta\right) = \frac{1}{N}\sum_{n=1}^{N} h_n\left(\theta\right), \quad \forall\ \theta \in \mathbb{R}^M$$

Recall $R\left(\mathbf{x}\right), \Gamma\left(i+1, \mathbf{x}, \omega\right)$ in (46)-(47). Then

$$\begin{aligned}
\left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right) = & \\
& -2\beta\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left(\overline{L}\otimes I_M\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right) \\
& -2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M\left(\mathbf{x}\right)-M(\mathbf{1}_N\otimes\theta^*)\right] \\
= -2\beta\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T & \left(\overline{L}\otimes I_M\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)- \\
& -2\sum_{n=1}^{N}\left[\left(\mathbf{x}_n-\theta^*\right)^T\left(h_n(\mathbf{x}_n)-h_n(\theta^*)\right)\right]\leq 0
\end{aligned} \tag{86}$$

where the last step follows from the positive-semidefiniteness of $\overline{L}\otimes I_M$ and (51). To verify Assumption **(B.3)**, show

$$\sup_{\epsilon<\|\mathbf{x}-\mathbf{1}_N\theta^*\|<\frac{1}{\epsilon}}\left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right)<0, \quad \forall\epsilon>0 \tag{87}$$

Assume, on the contrary, (87) not satisfied. Then from (86)

$$\sup_{\epsilon<\|\mathbf{x}-\mathbf{1}_N\theta^*\|<\frac{1}{\epsilon}}\left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right)=0, \quad \forall\epsilon>0$$

Then, there exists a sequence, $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ in $\left\{\mathbf{x}\in\mathbb{R}^{NM}\;\middle|\;\epsilon<\|\mathbf{x}-\mathbf{1}_N\theta^*\|<\frac{1}{\epsilon}\right\}$, such that

$$\lim_{k\to\infty}\left(R(\mathbf{x}^k), V_{\mathbf{x}}(\mathbf{x}^k)\right)=0$$

Since $\left\{\mathbf{x}\in\mathbb{R}^{NM}\mid\epsilon<\|\mathbf{x}-\mathbf{1}_N\theta^*\|<\frac{1}{\epsilon}\right\}$ is relatively compact, $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ has a limit point, $\widehat{\mathbf{x}}$, such that $\epsilon\leq\|\widetilde{\mathbf{x}}-\mathbf{1}_N\theta^*\|\leq\frac{1}{\epsilon}$, and, by continuity of $\left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right)$:

$$(R(\widetilde{\mathbf{x}}), V_{\mathbf{x}}(\widetilde{\mathbf{x}}))=0$$

From (51) and (86), we then have

$$\left(\widetilde{\mathbf{x}}-\mathbf{1}_N\otimes\theta^*\right)^T\left(\overline{L}\otimes I_M\right)\left(\widetilde{\mathbf{x}}-\mathbf{1}_N\otimes\theta^*\right)=0, \tag{88}$$

$$\left(\widetilde{\mathbf{x}}_n-\theta^*\right)^T\left(h_n(\widetilde{\mathbf{x}}_n)-h_n(\theta^*)\right)=0,\ \forall n \tag{89}$$

The equality in (88) and the properties of the Laplacian imply that $\widetilde{\mathbf{x}}\in\mathcal{C}$ and hence there exists $\mathbf{a}\in\mathbb{R}^M$, such that,

$$\widetilde{\mathbf{x}}_n=\mathbf{a}, \quad \forall n$$

The inequalities in (89) then imply

$$\left(\mathbf{a}-\theta^*\right)^T\left(h(\mathbf{a})-h(\theta^*)\right)=0$$

which is a contradiction by (85) since $\mathbf{a}\neq\theta^*$. Thus, we have (87) that verifies Assumption **(B.3)**. Finally, we note that,

$$\begin{aligned}
\|R\left(\mathbf{x}\right)\|^2 &= \left\|\beta\left(\overline{L}\otimes I_M\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)+\right. \\
&\qquad\qquad \left.+\left(M\left(\mathbf{x}\right)-M(\mathbf{1}_N\otimes\theta^*)\right)\right\|^2 \\
&\leq 4\beta^2\left\|\left(\overline{L}\otimes I_M\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)\right\|^2+ \\
&\qquad\qquad +4\left\|M\left(\mathbf{x}\right)-M(\mathbf{1}_N\otimes\theta^*)\right\|^2 \\
&\leq 4\beta^2\lambda_N(\overline{L})\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|^2+4K^2\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|^2
\end{aligned}$$

where the second step follows from the Lipschitz continuity of $h_n(\cdot)$ and K is defined in (52). To verify Assumption **(B.4)**, we have then along similar lines as in Theorem 7

$$\begin{aligned}\|R\left(\mathbf{x}\right)\|^2+\mathbb{E}\left[\|\Gamma\left(i+1,\mathbf{x},\omega\right)\|^2\right] &\leq k_1(1+V\left(\mathbf{x}\right))\\ &\leq k_1(1+V\left(\mathbf{x}\right))-\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)\end{aligned}$$

for a constant $k_1>0$ (the last step follows from (86).) Hence, the assumptions are satisfied and the claim follows. ■

Proof of Theorem 19

Proof: The recursive scheme in (44) satisfies Assumptions **(B.1)-(B.2)** of Theorem 29. To prove consistency, we verify Assumptions **(B.3)-(B.4)**. Consider the Lyapunov function

$$V\left(\mathbf{x}\right)=\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|^2$$

Clearly,

$$V(\mathbf{1}_N\otimes\theta^*)=0, V\left(\mathbf{x}\right)>0, \mathbf{x}\neq\mathbf{1}_N\otimes\theta^*, \lim_{\|\mathbf{x}\|\rightarrow\infty}V\left(\mathbf{x}\right)=\infty$$

Recall the definitions of $R\left(\mathbf{x}\right),\Gamma\left(i+1,\mathbf{x},\omega\right)$ in (46)-(47), and the consensus subspace in (49). We then have

$$\begin{aligned}&\left(R\left(\mathbf{x}\right),V_{\mathbf{x}}\left(\mathbf{x}\right)\right)= \qquad (90)\\ &-2\beta\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left(\overline{L}\otimes I_M\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)-\\ &\qquad-2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M\left(\mathbf{x}\right)-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &\leq-2\beta\lambda_2(\overline{L})\|\mathbf{x}_{\mathcal{C}^\perp}\|^2-\\ &\qquad-2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M\left(\mathbf{x}\right)-M(\mathbf{x}_{\mathcal{C}})\right]\\ &\qquad-2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &\leq-2\beta\lambda_2(\overline{L})\|\mathbf{x}_{\mathcal{C}^\perp}\|^2+\\ &\qquad+2\left\|\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M\left(\mathbf{x}\right)-M(\mathbf{x}_{\mathcal{C}})\right]\right\|\\ &\qquad-2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &\leq-2\beta\lambda_2(\overline{L})\|\mathbf{x}_{\mathcal{C}^\perp}\|^2+2K\|\mathbf{x}_{\mathcal{C}^\perp}\|\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|\\ &\qquad-2\left(\mathbf{x}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &=-2\beta\lambda_2(\overline{L})\|\mathbf{x}_{\mathcal{C}^\perp}\|^2+2K\|\mathbf{x}_{\mathcal{C}^\perp}\|\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|-\\ &-2\mathbf{x}_{\mathcal{C}^\perp}^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &\qquad-2\left(\mathbf{x}_{\mathcal{C}}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &\leq-2\beta\lambda_2(\overline{L})\|\mathbf{x}_{\mathcal{C}^\perp}\|^2+2K\|\mathbf{x}_{\mathcal{C}^\perp}\|\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|+\\ &\qquad2\left\|\mathbf{x}_{\mathcal{C}^\perp}^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\right\|\\ &\qquad-2\left(\mathbf{x}_{\mathcal{C}}-\mathbf{1}_N\otimes\theta^*\right)^T\left[M(\mathbf{x}_{\mathcal{C}})-M(\mathbf{1}_N\otimes\theta^*)\right]\\ &\leq-2\beta\lambda_2(\overline{L})\|\mathbf{x}_{\mathcal{C}^\perp}\|^2+2K\|\mathbf{x}_{\mathcal{C}^\perp}\|\|\mathbf{x}-\mathbf{1}_N\otimes\theta^*\|+\\ &\quad+2K\|\mathbf{x}_{\mathcal{C}^\perp}\|\|\mathbf{x}_{\mathcal{C}}-\mathbf{1}_N\otimes\theta^*\|-2\gamma\|\mathbf{x}_{\mathcal{C}}-\mathbf{1}_N\otimes\theta^*\|^2\\ &=\left(-2\beta\lambda_2(\overline{L})+2K\right)\|\mathbf{x}_{\mathcal{C}^\perp}\|^2+\\ &\quad+4K\|\mathbf{x}_{\mathcal{C}^\perp}\|\|\mathbf{x}_{\mathcal{C}}-\mathbf{1}_N\otimes\theta^*\|-2\gamma\|\mathbf{x}_{\mathcal{C}}-\mathbf{1}_N\otimes\theta^*\|^2\end{aligned}$$

where the second to last step is justified because $\mathbf{x}_{\mathcal{C}} = \mathbf{1}_N \otimes \widetilde{\mathbf{y}}$ for some $\widetilde{\mathbf{y}} \in \mathbb{R}^M$ and

$$\begin{aligned}
&\left(\mathbf{x}_{\mathcal{C}} - \mathbf{1}_N \otimes \theta^*\right)^T \left[M(\mathbf{x}_{\mathcal{C}}) - M(\mathbf{1}_N \otimes \theta^*)\right] = \\
&\qquad\qquad \sum_{n=1}^{N} \left(\widetilde{\mathbf{y}} - \theta^*\right)^T \left[h_n(\widetilde{\mathbf{y}}) - h_n(\theta^*)\right] \\
&= \left(\widetilde{\mathbf{y}} - \theta^*\right)^T \sum_{n=1}^{N} \left[h_n(\widetilde{\mathbf{y}}) - h_n(\theta^*)\right] \\
&= N\left(\widetilde{\mathbf{y}} - \theta^*\right)^T \left[h(\widetilde{\mathbf{y}}) - h(\theta^*)\right] \geq N\gamma \left\|\widetilde{\mathbf{y}} - \theta^*\right\|^2 \\
&= \gamma \left\|\mathbf{x}_{\mathcal{C}} - \mathbf{1}_N \otimes \theta^*\right\|^2
\end{aligned}$$

It can be shown that, if $\beta > \frac{K^2 + K\gamma}{\gamma\lambda_2\bar{L}}$, the term on the R.H.S. of (90) is always non-positive. We thus have

$$\left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right) \leq 0, \quad \forall \mathbf{x} \in \mathbb{R}^{MN}$$

By the continuity of $\left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right)$ and the relative compactness of $\left\{\mathbf{x} \in \mathbb{R}^{NM} \;\middle|\; \epsilon < \|\mathbf{x} - \mathbf{1}_N\theta^*\| < \frac{1}{\epsilon}\right\}$, we can show along similar lines as in Theorem 18 that

$$\sup_{\epsilon < \|\mathbf{x} - \mathbf{1}_N\theta^*\| < \frac{1}{\epsilon}} \left(R\left(\mathbf{x}\right), V_{\mathbf{x}}\left(\mathbf{x}\right)\right) < 0, \quad \forall \epsilon > 0$$

verifying Assumption **(B.3)**. Assumption **(B.4)** is verified in similar manner to Theorem 18 and the result follows. ■

APPENDIX C
PROOF OF LEMMA 25

Proof of Lemma 25: We prove for the case $\delta_1 < 1$ first. Consider j sufficiently large, such that,

$$r_1(i) \leq 1, \quad \forall i \geq j$$

Then, for $k \geq j$, using $1 - a \leq e^{-a}$, $0 \leq a \leq 1$:

$$\prod_{l=k+1}^{i-1} \left(1 - r_1(l)\right) \leq e^{-\sum_{l=k+1}^{i-1} r_1(l)} \tag{91}$$

It follows from the properties of the Riemann integral that

$$\begin{aligned}
\sum_{l=k+1}^{i-1} r_1(l) &= \sum_{l=k+1}^{i-1} \frac{a_1}{(l+1)^{\delta_1}} \\
&\geq a_1 \int_{k+2}^{i+1} \frac{1}{t^{\delta_1}} dt \\
&= \frac{a_1}{1-\delta_1}\left[(i+1)^{1-\delta_1} - (k+2)^{1-\delta_1}\right]
\end{aligned}$$

We thus have from (91)-(92)

$$\begin{aligned}
&\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(1 - r_1(l)\right)\right) r_2(l)\right] \leq \\
&\sum_{k=j}^{i-1}\left[e^{-\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}} e^{\frac{a_1}{1-\delta_1}(k+2)^{1-\delta_1}}\right] \frac{a_2}{(k+1)^{\delta_2}} = \\
&a_2 e^{-\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}} \sum_{k=j}^{i-1}\left[e^{\frac{a_1}{1-\delta_1}(k+2)^{1-\delta_1}} \frac{1}{(k+1)^{\delta_2}}\right]
\end{aligned}$$

From properties of Riemann integration, for j large enough:

$$\begin{aligned}
&\sum_{k=j}^{i-1}\left[e^{\frac{a_1}{1-\delta_1}(k+2)^{1-\delta_1}}\frac{1}{(k+1)^{\delta_2}}\right] \leq \\
&\qquad\qquad \sum_{k=j}^{i-1}\left[e^{\frac{a_1}{1-\delta_1}(k+2)^{1-\delta_1}}\frac{1}{(\frac{k}{2}+1)^{\delta_2}}\right] \\
&= 2^{\delta_2}\sum_{k=j}^{i-1}\left[e^{\frac{a_1}{1-\delta_1}(k+2)^{1-\delta_1}}\frac{1}{(k+2)^{\delta_2}}\right] \\
&= 2^{\delta_2}\sum_{k=j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}k^{1-\delta_1}}\frac{1}{k^{\delta_2}}\right] \\
&= 2^{\delta_2}e^{\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}}\frac{1}{(i+1)^{\delta_2}}+ \\
&\qquad\qquad + 2^{\delta_2}\sum_{k=j+2}^{i}\left[e^{\frac{a_1}{1-\delta_1}k^{1-\delta_1}}\frac{1}{k^{\delta_2}}\right] \\
&\leq 2^{\delta_2}e^{\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}}\frac{1}{(i+1)^{\delta_2}}+ \\
&\qquad\qquad + 2^{\delta_2}\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}\right]dt
\end{aligned} \tag{93}$$

Again by the fundamental theorem of calculus,

$$\begin{aligned}
e^{\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}} &= a_1\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_1}}\right]dt + C_1 \\
&= a_1\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}t^{\delta_2-\delta_1}\right]dt + C_1
\end{aligned} \tag{94}$$

where $C_1 = C_1(j) > 0$ for sufficiently large j. From (93)-(94) we have

$$\begin{aligned}
&\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-r_1(l))\right)r_2(i)\right] = \\
&\quad a_2 e^{-\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}}\sum_{k=j}^{i-1}\left[e^{\frac{a_1}{1-\delta_1}(k+2)^{1-\delta_1}}\frac{1}{(k+1)^{\delta_2}}\right] \leq \\
&\leq \frac{2^{\delta_2}a_2 e^{\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}}\frac{1}{(i+1)^{\delta_2}}+2^{\delta_2}a_2\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}\right]dt}{e^{\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}}} \\
&= \frac{2^{\delta_2}a_2}{(i+1)^{\delta_2}} + \frac{2^{\delta_2}a_2\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}\right]dt}{e^{\frac{a_1}{1-\delta_1}(i+1)^{1-\delta_1}}} \\
&\leq \frac{2^{\delta_2}a_2}{(i+1)^{\delta_2}} + \frac{2^{\delta_2}a_2\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}\right]dt}{a_1\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}t^{\delta_2-\delta_1}\right]dt + C_1}
\end{aligned} \tag{95}$$

The second term stays bounded if $\delta_1 = \delta_2$ and goes to zero as $i \to \infty$ if $\delta_1 < \delta_2$, thus establishing the Lemma for the case $\delta_1 < 1$. Also, in the case $\delta_1 = \delta_2$, we have from (95):

$$\begin{aligned}\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-r_1(l))\right)r_2(i)\right] &\leq \frac{2^{\delta_2}a_2}{(i+1)^{\delta_2}}+\\ &+\frac{2^{\delta_2}a_2}{a_1+C_1\left[\int_{j+2}^{i+1}\left[e^{\frac{a_1}{1-\delta_1}t^{1-\delta_1}}\frac{1}{t^{\delta_2}}\right]dt\right]^{-1}}\\ &\leq 2^{\delta_2}a_2+\frac{2^{\delta_2}a_2}{a_1}\end{aligned}$$

thus making the choice of B in (70) independent of i, j.

Now consider the case $\delta_1 = 1$. Consider j sufficiently large, such that,

$$r_1(i) \leq 1, \quad \forall i \geq j$$

Using a similar set of manipulations for $k \geq j$, we have

$$\begin{aligned}\prod_{l=k+1}^{i-1}(1-r_1(l)) &\leq e^{-a_1\sum_{l=k+1}^{i-1}\frac{1}{l+1}}\\ &\leq e^{-a_1\int_{k+2}^{i+1}\frac{1}{t}dt}\\ &= e^{-a_1\ln\left(\frac{i+1}{k+2}\right)}\\ &= \frac{(k+2)^{a_1}}{(i+1)^{a_1}}\end{aligned}$$

We thus have

$$\begin{aligned}\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-r_1(l))\right)r_2(i)\right] &\leq \frac{a_2}{(i+1)^{a_1}}\sum_{k=j}^{i-1}\frac{(k+2)^{a_1}}{(k+1)^{\delta_2}}\\ &\leq \frac{2^{\delta_2}a_2}{(i+1)^{a_1}}\sum_{k=j}^{i-1}\frac{(k+2)^{a_1}}{(k+2)^{\delta_2}}\\ &= \frac{2^{\delta_2}a_2}{(i+1)^{a_1}}\sum_{k=j+2}^{i+1}\frac{k^{a_1}}{k^{\delta_2}}\end{aligned}$$

Now, if $a_1 \geq \delta_2$, then

$$\begin{aligned}&\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-r_1(l))\right)r_2(i)\right] \leq\\ &\leq \frac{2^{\delta_2}a_2}{(i+1)^{a_1}}\sum_{k=j+2}^{i+1}k^{a_1-\delta_2}\\ &= \frac{2^{\delta_2}a_2}{(i+1)^{a_1}}\left[(i+1)^{a_1-\delta_2}+\sum_{k=j+2}^{i}k^{a_1-\delta_2}\right]\\ &\leq \frac{2^{\delta_2}a_2}{(i+1)^{a_1}}\left[(i+1)^{a_1-\delta_2}+\int_{j+2}^{i+1}t^{a_1-\delta_2}dt\right]\\ &= \frac{2^{\delta_2}a_2}{(i+1)^{\delta_2}}+\frac{2^{\delta_2}a_2}{a-\delta_2+1}\frac{(i+1)^{a-\delta_2+1}-(j+2)^{a-\delta_2+1}}{(i+1)^{a_1}}\end{aligned}$$

The second term is bounded if $\delta_2 = 1$ and vanishes if $\delta_2 > 1$. If $a_1 < \delta_2$ is resolved similarly. ■

APPENDIX D
PROOFS OF LEMMAS 24,23

Proof of Lemma 24

Proof: It follows from (62) and (64), and the fact that

$$(\mathbf{1}_N \otimes I_M)^T \left(\overline{L} \otimes I_M\right) = \mathbf{0}$$

that the evolution of the sequence, $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$ is given by

$$\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1) = \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - \alpha(i)\left[\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - \frac{1}{N}\sum_{n=1}^{N} g_n(\mathbf{z}_n(i))\right] \tag{96}$$

We note that (96) can be written as

$$\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1) = \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) + \alpha(i)\left[R(\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)) + \Gamma(i+1, \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i), \omega)\right]$$

where

$$R(\mathbf{y}) = -\left(\mathbf{y} - h(\theta^*)\right), \tag{97}$$

$$\Gamma(i+1, \mathbf{y}, \omega) = \frac{1}{N}\sum_{n=1}^{N} g_n(\mathbf{z}_n(i)) - h(\theta^*), \mathbf{y} \in \mathbb{R}^M \tag{98}$$

Such a definition of $R(\cdot), \Gamma(\cdot)$ clearly satisfies Assumptions **(B.1)-(B.2)** of Theorem 29. Now, defining

$$V(\mathbf{y}) = \|\mathbf{y} - h(\theta^*)\|^2$$

we have

$$V\left(h(\theta^*)\right) = 0, V(\mathbf{y}) > 0, \mathbf{y} \neq h(\theta^*), \lim_{\|\mathbf{y}\|\to\infty} V(\mathbf{y}) = \infty$$

Also, we have for $\epsilon > 0$

$$\begin{aligned}\sup_{\epsilon < \|\mathbf{y} - h(\theta^*)\| < \frac{1}{\epsilon}} & (R(\mathbf{y}), V_{\mathbf{y}}(\mathbf{y})) = \\ & = \sup_{\epsilon < \|\mathbf{y} - h(\theta^*)\| < \frac{1}{\epsilon}} \left(-2\|\mathbf{y} - h(\theta^*)\|^2\right) \\ & \leq -2\epsilon^2 \\ & < 0\end{aligned}$$

thus verifying Assumption **(B.3)**. Finally from (43) and (97)-(98), we have

$$\begin{aligned}\|R(\mathbf{y})\|^2 + \mathbb{E}_{\theta^*}\left[\|\Gamma(i+1, \mathbf{y}, \omega)\|^2\right] = & \\ & = \|\mathbf{y} - h(\theta^*)\|^2 + \eta(\theta^*) \\ & \leq k_1(1 + V(\mathbf{y})) \\ & \leq k_1(1 + V(\mathbf{y})) - (R(\mathbf{y}), V_{\mathbf{y}}(\mathbf{y}))\end{aligned}$$

for $k_1 = \max(1, \eta(\theta^*))$. Thus the Assumptions **(B.1)-(B.4)** are satisfied, and we have the claim in (67).

To establish (68), we note that, for sufficiently large i,

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - h(\theta^*)\right\|^2\right] = & \\ & = (1 - \alpha(i-1))^2 \mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i-1) - h(\theta^*)\right\|^2\right] + \\ & \qquad + \alpha^2(i-1)\eta(\theta^*) \\ & \leq (1 - \alpha(i-1))\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i-1) - h(\theta^*)\right\|^2\right] + \\ & \qquad + \alpha^2(i-1)\eta(\theta^*)\end{aligned} \tag{99}$$

where the last step follows from the fact that $0 \leq (1-\alpha(i)) \leq 1$ for sufficiently large i. Continuing the recursion in (99), we have for sufficiently large $j \leq i$

$$\begin{aligned}\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - h(\theta^*)\right\|^2\right] \leq & \qquad (100)\\ \leq \left(\prod_{k=j}^{i-1}(1-\alpha(k))\right)\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(0) - h(\theta^*)\right\|^2 + &\\ + \eta(\theta^*)\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\alpha(l))\right)\alpha^2(k)\right] &\\ \leq \left(e^{-\sum_{k=j}^{i-1}\alpha(k)}\right)\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(0) - h(\theta^*)\right\|^2 + &\\ + \eta(\theta^*)\sum_{k=0}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\alpha(l))\right)\alpha^2(k)\right] &\end{aligned}$$

From Assumption **(D.5)**, we note that $\sum_{k=j}^{i-1}\alpha(k) \to \infty$ as $i \to \infty$ because $0.5 < \tau_1 \leq 1$. Thus, the first term in (100) goes to zero as $i \to \infty$. The second term in (100) falls under the purview of Lemma 25 with $\delta_1 = \tau_1$ and $\delta_2 = 2\tau_1$ and hence goes to zero as $i \to \infty$. We thus have

$$\lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - h(\theta^*)\right\|^2\right] = 0$$

■

Proof of Lemma 23

Proof: Recall from (62) and (96) that the evolution of the sequences $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$ and $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$ are given by

$$\begin{aligned}\widetilde{\mathbf{x}}^{\circ}(i+1) = \widetilde{\mathbf{x}}^{\circ}(i) - \beta(i)\left(\overline{L}\otimes I_M\right)\widetilde{\mathbf{x}}^{\circ}(i) - & \qquad (101)\\ - \alpha(i)\left[\widetilde{\mathbf{x}}(i) - J(\mathbf{z}(i))\right] &\\ \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1) = \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - & \qquad (102)\\ - \alpha(i)\left[\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) - \frac{1}{N}\sum_{n=1}^{N} g_n(\mathbf{z}_n(i))\right] &\end{aligned}$$

To establish the claim (65), Lemma 23, we prove

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\left\|\widetilde{\mathbf{x}}^{\circ}(i) - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\| = 0\right] = 1$$

Recall the matrix

$$P = \frac{1}{N}\left(\mathbf{1}_N \otimes I_M\right)\left(\mathbf{1}_N \otimes I_M\right)^T \qquad (103)$$

and note that

$$P\widetilde{\mathbf{x}}^{\circ}(i) = \mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i), \quad P\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i) = \mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i), \forall i$$

From (101)-(102), we then have

$$\begin{aligned}\widetilde{\mathbf{x}}^{\circ}(i+1) - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1)\right) = & \qquad (104)\\ = \left[I_{NM} - \beta(i)\left(\overline{L}\otimes I_M\right) - \alpha(i) I_{NM} - P\right]\left[\widetilde{\mathbf{x}}^{\circ}(i) - \right. &\\ \left. - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right] &\\ + \alpha(i)\left[J(\mathbf{z}(i)) - PJ(\mathbf{z}(i))\right] &\end{aligned}$$

Choose δ satisfying

$$0 < \delta < \tau_1 - \frac{1}{2+\epsilon_1} - \tau_2. \qquad (105)$$

Such a choice exists by Assumption **(D.5)**. Now claim:

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\frac{1}{(i+1)^{\frac{1}{2+\epsilon_1}+\delta}}\left\|J(\mathbf{z}(i))-PJ(\mathbf{z}(i))\right\|=0\right]=1 \tag{106}$$

Indeed, consider any $\epsilon>0$. We then have from Assumption **(D.4)** and Chebyshev's inequality

$$\begin{aligned}\sum_{i\geq 0}\mathbb{P}_{\theta^*}\left[\frac{1}{(i+1)^{\frac{1}{2+\epsilon_1}+\delta}}\left\|J(\mathbf{z}(i))-PJ(\mathbf{z}(i))\right\|>\epsilon\right]\leq&\\ \leq&\sum_{i\geq 0}\frac{1}{(i+1)^{1+\delta(2+\epsilon_1)}\epsilon^{2+\epsilon_1}}\\ \mathbb{E}_{\theta}\left[\left\|J(\mathbf{z}(i))-PJ(\mathbf{z}(i))\right\|^{2+\epsilon_1}\right]=&\frac{\kappa(\theta^*)}{\epsilon^{2+\epsilon_1}}\sum_{i\geq 0}\frac{1}{(i+1)^{1+\delta(2+\epsilon_1)}}\\ <&\infty\end{aligned}$$

It then follows from the Borel-Cantelli Lemma (see [50]) that for arbitrary $\epsilon>0$

$$\mathbb{P}_{\theta^*}\left[\frac{1}{(i+1)^{\frac{1}{2+\epsilon_1}+\delta}}\left\|J(\mathbf{z}(i))-PJ(\mathbf{z}(i))\right\|>\epsilon\text{ i.o.}\right]=0 \tag{107}$$

where i.o. stands for infinitely often. Since the above holds for ϵ arbitrarily small, we have (see [50]) the a.s. claim in (106).

Consider the set $\Omega_1\subset\Omega$ with $\mathbb{P}_{\theta^*}[\Omega_1]=1$, where the a.s. property in (106) holds. Also, consider the set $\Omega_2\subset\Omega$ with $\mathbb{P}_{\theta^*}[\Omega_2]=1$, where the sequence $\left\{\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right\}_{i\geq 0}$ converges to $h(\theta^*)$. Let $\Omega_3=\Omega_1\cap\Omega_2$. It is clear that $\mathbb{P}_{\theta^*}[\Omega_3]=1$. We will now show that, on Ω_3, the sample paths of the sequence $\{\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$ converge to $(\mathbf{1}_N\otimes h(\theta^*))$, thus proving the Lemma. In the following we index the sample paths by ω to emphasize the fact that we are establishing properties pathwise.

From (104), we have on $\omega\in\Omega_3$

$$\begin{aligned}&\left\|\widetilde{\mathbf{x}}^{\circ}(i+1,\omega)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1,\omega)\right)\right\|\leq\\ &\quad\leq\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\|\\ &\qquad\qquad\left\|\widetilde{\mathbf{x}}^{\circ}(i,\omega)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i,\omega)\right)\right\|\\ &\quad+\frac{a}{(i+1)^{\tau_1-\frac{1}{2+\epsilon_1}-\delta}}\\ &\qquad\qquad\left\|\frac{1}{(i+1)^{\frac{1}{2+\epsilon_1}+\delta}}\left[J(\mathbf{z}(i,\omega))-PJ(\mathbf{z}(i,\omega))\right]\right\|\end{aligned}$$

For sufficiently large i, we have

$$\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\|\leq 1-\beta(i)\lambda_2(\overline{L}) \tag{108}$$

From (107) for $\omega\in\Omega_3$ we can choose $\epsilon>0$ and $j(\omega)$ such that $\forall i\geq j(\omega)$

$$\left\|\frac{1}{(i+1)^{\frac{1}{2+\epsilon_1}+\delta}}\left[J(\mathbf{z}(i,\omega))-PJ(\mathbf{z}(i,\omega))\right]\right\|\leq\epsilon. \tag{109}$$

Let $j(\omega)$ be sufficiently large such that (108) is also satisfied in addition to (109). We then have for $\omega \in \Omega_3,\ i \geq j(\omega)$

$$\begin{aligned}&\left\|\widetilde{\mathbf{x}}^{\circ}(i,\omega)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i,\omega)\right)\right\| \leq \\ &\leq \left(\prod_{k=j(\omega)}^{i-1}\left(1-\beta(k)\lambda_2(\overline{L})\right)\right)\|\widetilde{\mathbf{x}}^{\circ}(j(\omega),\omega)- \\ &\qquad -\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(j(\omega),\omega)\right)\| + \\ &+a\epsilon\sum_{k=j(\omega)}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(1-\beta(l)\lambda_2(\overline{L})\right)\right)\frac{1}{(k+1)^{\tau_1-\frac{1}{2+\epsilon_1}-\delta}}\right]\end{aligned} \tag{110}$$

For the first term on the R.H.S. of (110) we note that

$$\begin{aligned}\prod_{k=j(\omega)}^{i-1}\left(1-\beta(k)\lambda_2(\overline{L})\right) &\leq e^{-\lambda_2(\overline{L})\sum_{k=j(\omega)}^{i-1}\beta(k)} \\ &= e^{-b\lambda_2(\overline{L})\sum_{k=j(\omega)}^{i-1}\frac{1}{(k+1)^{\tau_2}}}\end{aligned}$$

which goes to zero as $i\rightarrow\infty$ since $\tau_2<1$ by Assumption **(D.5)**. Hence the first term on the R.H.S. of (110) goes to zero as $i\rightarrow\infty$. The summation in the second term on the R.H.S. of (110) falls under the purview of Lemma 25 with $\delta_1=\tau_2$ and $\delta_2=\tau_1-\frac{1}{2+\epsilon_1}-\delta$. It follows from the choice of δ in (105) and Assumption **(D.5)** that $\delta_1<\delta_2$ and hence the term $\sum_{k=j(\omega)}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(1-\beta(l)\lambda_2(\overline{L})\right)\right)\frac{1}{(k+1)^{\tau_1-\frac{1}{2+\epsilon_1}-\delta}}\right]\rightarrow 0$ as $i\rightarrow\infty$. We then conclude from (110) that, for $\omega\in\Omega_3$

$$\lim_{i\rightarrow\infty}\left\|\widetilde{\mathbf{x}}^{\circ}(i,\omega)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i,\omega)\right)\right\|=0$$

The Lemma then follows from the fact that $\mathbb{P}_{\theta^*}\left[\Omega_3\right]=1$.

To establish (66), we have from (104)

$$\begin{aligned}&\left\|\widetilde{\mathbf{x}}^{\circ}(i+1)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1)\right)\right\|^2 \leq \\ &\leq\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\|^2 \\ &\qquad\left\|\widetilde{\mathbf{x}}^{\circ}(i)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|^2+ \\ &+2\alpha(i)\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\| \\ &\left\|\widetilde{\mathbf{x}}^{\circ}(i)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|\left\|J(\mathbf{z}(i))-PJ(\mathbf{z}(i))\right\|+ \\ &+\alpha^2(i)\left\|J(\mathbf{z}(i))-PJ(\mathbf{z}(i))\right\|^2\end{aligned}$$

Taking expectations on both sides and from (55)

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(i+1)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1)\right)\right\|^2\right]\leq \\ &\leq\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\|^2 \\ &\qquad\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(i)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|^2\right]+ \\ &+2\alpha(i)\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\|\kappa_1\left(\theta^*\right) \\ &\qquad\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(i)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|^2\right]+ \\ &+2\alpha(i)\left\|I-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}-P\right\|\kappa_1\left(\theta^*\right)+ \\ &\qquad+\alpha^2(i)\kappa_2(\theta^*)\end{aligned}$$

where we used the inequality that $\forall i$

$$\left\|\widetilde{\mathbf{x}}^{\circ}(i)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|\leq\left\|\widetilde{\mathbf{x}}^{\circ}(i)-\left(\mathbf{1}_N\otimes\widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|^2+1.$$

Choose j sufficiently large such that $\forall i \geq j$

$$\left\|I - \beta(i)\left(\overline{L} \otimes I_M\right) - \alpha(i) I_{NM} - P\right\| 1 - \beta(i)\lambda_2(\overline{L}).$$

For $i \geq j$, it can be shown that, for $c_1 > 0$ a constant:

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(i+1) - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i+1)\right)\right\|^2\right] \leq \\
&\leq \left[1 - \beta(i)\lambda_2(\overline{L}) + 2\alpha(i)\kappa_1(\theta^*)\right] \\
&\qquad \mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(i) - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|^2\right] + \alpha(i)c_1.
\end{aligned} \tag{112}$$

Now choose $j_1 \geq j$ and $0 < c_2 < \lambda_2(\overline{L})$[12] such that,

$$1 - \beta(i)\lambda_2(\overline{L}) + 2\alpha(i)\kappa_1(\theta^*) \leq 1 - \beta(i)c_2, \quad \forall i \geq j_1$$

Then the claim in (66) follows because for $i \geq j_1$

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(i) - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(i)\right)\right\|^2\right] \leq \\
&\left(\prod_{k=j_1}^{i-1}\left(1 - \beta(k)c_2\right)\right)\mathbb{E}_{\theta^*}\left[\left\|\widetilde{\mathbf{x}}^{\circ}(j_1) - \left(\mathbf{1}_N \otimes \widetilde{\mathbf{x}}^{\circ}_{\text{avg}}(j)\right)\right\|^2\right] + \\
&\qquad + c_1 \sum_{k=j_1}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(1 - \beta(l)c_2\right)\right)\alpha(k)\right]
\end{aligned}$$

and the first and second terms on the R.H.S. of (112) vanish as $i \to \infty$ by the argument in (111) and Lemma 25, respectively. ■

APPENDIX E
PROOF OF LEMMA 26

Proof of Lemma 26: From (62) and (63) we have

$$\begin{aligned}
\widehat{\mathbf{x}}(i+1) - \widetilde{\mathbf{x}}^{\circ}(i+1) = &\left[I_{NM} - \beta(i)\left(\overline{L} \otimes I_M\right) - \right. \\
&\left. - \alpha(i) I_{NM}\right]\left[\widehat{\mathbf{x}}(i) - \widetilde{\mathbf{x}}^{\circ}(i)\right] - \beta(i)\left(\boldsymbol{\Upsilon}(i) + \boldsymbol{\Psi}(i)\right)
\end{aligned} \tag{113}$$

For sufficiently large j, we have

$$\left\|I - \beta(i)\left(\overline{L} \otimes I_M\right) - \alpha(i) I_{NM}\right\| \leq 1 - \alpha(i), \forall i \geq j \tag{114}$$

We then have from (113), for $i \geq j$,

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left[\left\|\widehat{\mathbf{x}}(i+1) - \widetilde{\mathbf{x}}^{\circ}(i+1)\right\|^2\right] \leq \\
&\leq \left(1 - \alpha(i)\right)^2 \mathbb{E}_{\theta^*}\left[\left\|\widehat{\mathbf{x}}(i) - \widetilde{\mathbf{x}}^{\circ}(i)\right\|^2\right] + \\
&\qquad + \beta^2(i)\mathbb{E}_{\theta^*}\left[\left\|\boldsymbol{\Upsilon}(i) + \boldsymbol{\Psi}(i)\right\|^2\right] \\
&\leq \left(1 - \alpha(i)\right)\mathbb{E}_{\theta^*}\left[\left\|\widehat{\mathbf{x}}(i) - \widetilde{\mathbf{x}}^{\circ}(i)\right\|^2\right] + \eta_q\beta^2(i)
\end{aligned}$$

[12] Such a choice exists because $\tau_1 > \tau_2$.

where the last step follows from the fact that $0 \leq (1-\alpha(i)) \leq 1$ for $i \geq j$ and (10). Continuing the recursion, we have

$$\mathbb{E}_{\theta^*}\left[\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|^2\right] \leq \left(\prod_{k=j}^{i-1}(1-\alpha(k))\right)\|\widehat{\mathbf{x}}(j)-\widetilde{\mathbf{x}}^{\circ}(j)\|^2 + \\ + \eta_q \sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\alpha(l))\right)\beta^2(k)\right] \tag{115}$$

The first and second terms on the R.H.S. of (115) vanish as $i \to \infty$, respectively because 1) of an argument similar to the proof of Lemma 24, and 2) by Lemma 25, with $\delta_1 = \tau_1, \delta_2 = 2\tau_2$, since by Assumption **(D.5)**, $2\tau_2 > \tau_1$. Thus:

$$\lim_{i\to\infty} \mathbb{E}_{\theta^*}\left[\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|^2\right] = 0$$

which shows that the sequence $\left\{\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|\right\}_{i\geq 0}$ converges to 0 in $\mathcal{L}_2$ (mean-squared sense). We then have from Lemma 23

$$\lim_{i\to\infty} \mathbb{E}_{\theta^*}\left[\|\widehat{\mathbf{x}}(i)-\mathbf{1}_N\otimes h(\theta^*)\|^2\right] \leq \\ \leq 2\lim_{i\to\infty} \mathbb{E}_{\theta^*}\left[\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|^2\right] + \\ + 2\lim_{i\to\infty} \mathbb{E}_{\theta^*}\left[\|\widetilde{\mathbf{x}}^{\circ}(i)-\mathbf{1}_N\otimes h(\theta^*)\|^2\right] = 0$$

thus establishing the claim in (73).

We now show that the sequence $\left\{\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|\right\}_{i\geq 0}$ also converges a.s. to a finite random variable. Choose j sufficiently large as in (114). We then have from (113)

$$\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i) = \tag{116} \\ \left(\prod_{k=j}^{i-1}\left(I_{NM}-\beta(k)\left(\overline{L}\otimes I_M\right)-\alpha(k)I\right)\right)\left(\widehat{\mathbf{x}}(j)-\widetilde{\mathbf{x}}^{\circ}(j)\right) - \\ -\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(I_{NM}-\beta(l)\left(\overline{L}\otimes I_M\right)-\alpha(l)I\right)\right) \\ \beta(k)\mathbf{\Upsilon}(k)\right] - \\ -\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(I_{NM}-\beta(l)\left(\overline{L}\otimes I_M\right)-\alpha(l)I\right)\right) \\ \beta(k)\mathbf{\Psi}(k)\right]$$

The first term on the R.H.S. of (116) converges a.s. to zero as $i \to \infty$ by a similar argument as in the proof of Lemma 24. Since the sequence $\{\mathbf{\Upsilon}(i)\}_{i\geq 0}$ is i.i.d., the second term is a weighted summation of independent random vectors. Define the triangular array of weight matrices, $\{A_{i,k},\ j \leq k \leq i-1\}_{i>j}$, by

$$A_{i,k} = \prod_{l=k+1}^{i-1}\left(I_{NM}-\beta(l)\left(\overline{L}\otimes I_M\right)-\alpha(l)I\right)\beta(k)$$

We then have

$$\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}\left(I_{NM}-\beta(l)\left(\overline{L}\otimes I_M\right)-\alpha(l)I\right)\right) \\ \beta(k)\mathbf{\Upsilon}(k)\right] = \sum_{k=j}^{i-1}A_{i,k}\mathbf{\Upsilon}(k)$$

By Lemma 25 and Assumption **(D.5)** we note that

$$\limsup_{i\to\infty}\sum_{k=j}^{i-1}\|A_{i,k}\|^2 \leq$$

$$\leq \limsup_{i\to\infty}\sum_{k=j}^{i-1}\left[\left(\prod_{l=k+1}^{i-1}(1-\alpha(l))\right)\beta^2(k)\right]$$

$$=0$$

It then follows that

$$\sup_{i>j}\sum_{k=j}^{i-1}\|A_{i,k}\|^2 = C_3 < \infty$$

The sequence $\left\{\sum_{k=j}^{i-1}A_{i,k}\mathbf{\Upsilon}(k)\right\}_{i>j}$ then converges a.s. to a finite random vector by standard results from the limit theory of weighted summations of independent random vectors (see [57], [58], [59]).

In a similar way, the last term on the R.H.S of (116) converges a.s. to a finite random vector since by the properties of dither the sequence $\{\mathbf{\Psi}(i)\}_{i\geq 0}$ is i.i.d. It then follows from (116) that the sequence $\{\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\}_{i\geq 0}$ converges a.s. to a finite random vector, which in turn implies that the sequence $\{\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|\}_{i\geq 0}$ converges a.s. to a finite random variable. However, we have already shown that the sequence $\{\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|\}_{i\geq 0}$ converges in mean-squared sense to 0. It then follows from the uniqueness of the mean-squared and a.s. limit, that the sequence $\{\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|\}_{i\geq 0}$ converges a.s. to 0. In other words,

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\|\widehat{\mathbf{x}}(i)-\widetilde{\mathbf{x}}^{\circ}(i)\|=0\right]=1 \tag{117}$$

The claim in (72) then follows from (117) and Lemma 23.

■

Appendix F
Proofs of Theorems 21,22

Proof of Theorem 21

Proof: Recall the evolution of the sequences $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$, $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$ in (57) and (63).

Then writing $L(i)=\overline{L}+\widetilde{L}(i)$ and using the fact that

$$\left(\widetilde{L}(i)\otimes I_M\right)\widehat{\mathbf{x}}(i)=\left(\widetilde{L}(i)\otimes I_M\right)\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i),\ \forall i$$

we have from (57) and (63)

$$\widetilde{\mathbf{x}}(i+1)-\widehat{\mathbf{x}}(i+1)=[I_{NM}-\beta(i)\left(L(i)\otimes I_M\right)- \\ -\alpha(i)I_{NM}]\left(\widetilde{\mathbf{x}}(i)-\widehat{\mathbf{x}}(i)\right)-\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i) \tag{118}$$

For ease of notation, introduce the sequence $\{\mathbf{y}(i)\}_{i\geq 0}$, given by

$$\mathbf{y}(i)=\widetilde{\mathbf{x}}(i)-\widehat{\mathbf{x}}(i)$$

To prove (59), it clearly suffices (from Lemma 26) to prove

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\mathbf{y}(i)=\mathbf{0}\right]=1$$

From (118), the evolution of the sequence $\{\mathbf{y}(i)\}_{i\geq 0}$ is:

$$\mathbf{y}(i+1)=\left[I_{NM}-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}\right]\mathbf{y}(i)- \\ -\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}(i)-\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i) \tag{119}$$

The sequence $\{\mathbf{y}(i)\}_{i\geq 0}$ is not uniformly bounded, in general, because of $\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)$. However, from Lemma 26:

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)=\mathbf{0}\right]=1$$

and, hence, asymptotically, its effect diminishes. However, $\{\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\}_{i\geq 0}$ is not uniformly bounded over sample paths and, hence, we use truncation arguments (see, e.g., [56]). For a scalar a, define its truncation $(a)^R$ at level $R>0$ by

$$(a)^R=\begin{cases}\frac{a}{|a|}\min(|a|,R) & \text{if } a\neq 0\\ 0 & \text{if } a=0\end{cases}$$

For a vector, the truncation operation applies componentwise. For $R>0$, we also consider the sequences, $\{\mathbf{y}_R(i)\}_{i\geq 0}$:

$$\begin{aligned}\mathbf{y}_R(i+1)=&\left[I_{NM}-\beta(i)\left(\overline{L}\otimes I_M\right)-\alpha(i)I_{NM}\right]\mathbf{y}_R(i)-\\&-\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}_R(i)-\\&-\beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^R\end{aligned}\tag{120}$$

We will show that for every $R>0$

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\mathbf{y}_R(i)=\mathbf{0}\right]=1\tag{121}$$

Now, the sequence $\{\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\}_{i\geq 0}$ converges a.s. to zero, and, hence, for every $\epsilon>0$, there exists $R(\epsilon)>0$ (see [50]), such that

$$\mathbb{P}_{\theta^*}\left[\sup_{i\geq 0}\left\|\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)-\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^{R(\epsilon)}\right\|=0\right]>1-\epsilon$$

and, hence, from (119)-(120)

$$\mathbb{P}_{\theta^*}\left[\sup_{i\geq 0}\left\|\mathbf{y}(i)-\mathbf{y}^{R(\epsilon)}(i)\right\|=0\right]>1-\epsilon$$

This, together with (121), will then imply

$$\mathbb{P}_{\theta^*}\left[\lim_{i\to\infty}\mathbf{y}(i)=\mathbf{0}\right]>1-\epsilon\tag{122}$$

Since $\epsilon>0$ is arbitrary in (122), we will be able to conclude (59). Thus, the proof reduces to establishing (121) for every $R>0$, which is carried out in the following.

For a given $R>0$ consider the recursion given in (120). Choose $\varepsilon_1>0$ and $\varepsilon_2<0$ such that

$$1-\varepsilon_2<2\tau_2-\varepsilon_1.$$

Because $\tau_2>.5$ in Assumption **(D.5)** permits such choice of $\varepsilon_1,\varepsilon_2$. Let $\rho>0$ be constant and define $V:\mathbb{N}\times\mathbb{R}^{NM}\longmapsto\mathbb{R}^+$

$$V\left(i,\mathbf{x}\right)=i^{\varepsilon_1}\mathbf{x}^T\left(\overline{L}\otimes I_M\right)\mathbf{x}+\rho i^{\varepsilon_2}.$$

Recall the filtration $\{\mathcal{F}_i\}_{i\geq 0}$ in (48)

$$\mathcal{F}_i=\sigma\left(\mathbf{x}(0),\left\{L(j),\{\mathbf{z}_n(j)\}_{1\leq N},\ \mathbf{\Upsilon}(j),\mathbf{\Psi}(j)\right\}_{0\leq j<i}\right)$$

to which all the processes of interest are adapted. We now show that there exists an integer $i_R>0$ sufficiently large, such that the process $\{V(i,\mathbf{y}_R(i))\}_{i\geq i_R}$ is a non-negative supermartingale w.r.t. the filtration $\{\mathcal{F}_i\}_{i\geq i_R}$. To this end, we note that, using the recursion (120):

$$
\begin{aligned}
&\mathbb{E}_{\theta^*}\left[V(i+1,\mathbf{y}_R(i+1))\,|\,\mathcal{F}_i\right]-V(i,\mathbf{y}_R(i))= \\
&\quad(i+1)^{\varepsilon_1}\mathbf{y}_R^T(i+1)\left(\overline{L}\otimes I_M\right)\mathbf{y}_R(i+1)+\rho(i+1)^{\varepsilon_2} \\
&\qquad\qquad-i^{\varepsilon_1}\mathbf{y}_R^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_R(i)-\rho i^{\varepsilon_2} \\
&=(i+1)^{\varepsilon_1}\Big[\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)- \\
&\qquad\qquad-2\beta(i)\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)^2\mathbf{y}_{R,\mathcal{C}^\perp}(i)- \\
&-2\alpha(i)\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)+ \\
&\qquad\qquad+2\beta(i)\alpha(i)\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)^2\mathbf{y}_{R,\mathcal{C}^\perp}(i)+ \\
&+\beta^2(i)\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)^3\mathbf{y}_{R,\mathcal{C}^\perp}(i)+ \\
&\qquad\qquad+\alpha^2(i)\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)+ \\
&+\beta^2(i)\mathbb{E}_{\theta^*}\Big[\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\widetilde{L}(i)\otimes I_M\right)\left(\overline{L}\otimes I_M\right) \\
&\qquad\qquad\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)\,\Big|\,\mathcal{F}_i\Big]+ \\
&+2\beta^2(i)\mathbb{E}_{\theta^*}\Big[\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\widetilde{L}(i)\otimes I_M\right)\left(\overline{L}\otimes I_M\right) \\
&\qquad\qquad\left(\widetilde{L}(i)\otimes I_M\right)\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^R\,\Big|\,\mathcal{F}_i\Big]+ \\
&+\beta^2(i)\mathbb{E}_{\theta^*}\Big[\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}^T(i)\right)^R\left(\widetilde{L}(i)\otimes I_M\right)\left(\overline{L}\otimes I_M\right) \\
&\qquad\qquad\left(\widetilde{L}(i)\otimes I_M\right)\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^R|\mathcal{F}_i\Big]\Big]+ \\
&+(i+1)^{\varepsilon_2}-i^{\varepsilon_1}\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)-\rho i^{\varepsilon_2}
\end{aligned}
\tag{123}
$$

where we repeatedly used the fact that

$$
\begin{aligned}
\left(\overline{L}\otimes I_M\right)\mathbf{y}_R(i)&=\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i) \\
\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}_R(i)&=\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)
\end{aligned}
$$

and $\widetilde{L}(i)$ is independent of $\mathcal{F}_i$.

In going to the next step we use the following inequalities, where $c_1 > 0$ is a constant:

$$
\begin{aligned}
&\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)^2\mathbf{y}_{R,\mathcal{C}^\perp}(i)\geq\lambda_2^2(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\quad=\frac{\lambda_2^2(\overline{L})}{\lambda_N(\overline{L})}\lambda_N(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\quad\geq\frac{\lambda_2^2(\overline{L})}{\lambda_N(\overline{L})}\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i) \\
&\quad\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)^2\mathbf{y}_{R,\mathcal{C}^\perp}(i)\leq\lambda_N^2(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\quad=\frac{\lambda_N^2(\overline{L})}{\lambda_2(\overline{L})}\lambda_2(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\quad\leq\frac{\lambda_N^2(\overline{L})}{\lambda_2(\overline{L})}\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i) \\
&\quad\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)^3\mathbf{y}_{R,\mathcal{C}^\perp}(i)\leq\lambda_N^3(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\quad=\frac{\lambda_N^3(\overline{L})}{\lambda_2(\overline{L})}\lambda_2(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\quad\leq\frac{\lambda_N^3(\overline{L})}{\lambda_2(\overline{L})}\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)
\end{aligned}
\tag{124}
$$

$$
\begin{aligned}
&\mathbb{E}_{\theta^*}\left[\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\widetilde{L}(i)\otimes I_M\right)\left(\overline{L}\otimes I_M\right)\right. \\
&\qquad\qquad\left.\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)\;\middle|\;\mathcal{F}_i\right]\leq \\
&\leq\lambda_N(\overline{L})\mathbb{E}_{\theta^*}\left[\left\|\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2\mid\mathcal{F}_i\right] \\
&\leq c_1\lambda_N(\overline{L})\mathbb{E}_{\theta^*}\left[\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2\;\middle|\;\mathcal{F}_i\right] \\
&=c_1\lambda_N(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\leq\frac{c_1\lambda_N(\overline{L})}{\lambda_2}\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i) \\
&\mathbb{E}_{\theta^*}\left[\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\widetilde{L}(i)\otimes I_M\right)\left(\overline{L}\otimes I_M\right)\right. \\
&\qquad\qquad\left.\left(\widetilde{L}(i)\otimes I_M\right)\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^R\;\middle|\;\mathcal{F}_i\right] \\
&\leq\mathbb{E}_{\theta^*}\left[\left\|\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\right\|\left\|\left(\widetilde{L}(i)\otimes I_M\right)\right\|\right. \\
&\left.\left\|\left(\overline{L}\otimes I_M\right)\right\|\left\|\left(\widetilde{L}(i)\otimes I_M\right)\right\|\left\|\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^R\right\|\;\middle|\;\mathcal{F}_i\right]
\end{aligned}
\tag{125}
$$

$$
\leq Rc_1\lambda_N(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\| \tag{126}
$$

$$
\begin{aligned}
&\leq Rc_1\lambda_N(\overline{L})+Rc_1\lambda_N(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^\perp}(i)\right\|^2 \\
&\leq Rc_1\lambda_N(\overline{L})+\frac{Rc_1\lambda_N(\overline{L})}{\lambda_2(\overline{L})}\mathbf{y}_{R,\mathcal{C}^\perp}^T(i)\left(\overline{L}\otimes I_M\right)\mathbf{y}_{R,\mathcal{C}^\perp}(i) \\
&\mathbb{E}_{\theta^*}\left[\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}^T(i)\right)^R\left(\widetilde{L}(i)\otimes I_M\right)\left(\overline{L}\otimes I_M\right)\right. \\
&\qquad\qquad\left.\left(\widetilde{L}(i)\otimes I_M\right)\left(\widehat{\mathbf{x}}_{\mathcal{C}^\perp}(i)\right)^R\;\middle|\;\mathcal{F}_i\right] \\
&\leq R^2c_1\lambda_N(\overline{L}) \\
&(i+1)^{\varepsilon_1}-i^{\varepsilon_1}\leq\varepsilon_1(i+1)^{\varepsilon_1-1} \\
&\rho(i+1)^{\varepsilon_2}-\rho i^{\varepsilon_2}\leq\rho\varepsilon_2 i^{\varepsilon_2-1}.
\end{aligned}
\tag{127}
$$

We go from (125) to (126) because $\left\|\left(\widehat{\mathbf{x}}_{\mathcal{C}^{\perp}}(i)\right)^{R}\right\| \leq R$. Using inequalities (124)-(127), we have from (123)

$$\begin{aligned}
&\mathbb{E}_{\theta^{*}}\left[V(i+1,\mathbf{y}_{R}(i+1))\,\middle|\,\mathcal{F}_{i}\right]-V(i,\mathbf{y}_{R}(i))\leq \\
&(i+1)^{\varepsilon_{1}}\left[\frac{\varepsilon_{1}}{(i+1)^{1}}-2\beta(i)\frac{\lambda_{2}^{2}(\overline{L})}{\lambda_{N}(\overline{L})}-2\alpha(i)+\right. \\
&+2\beta(i)\alpha(i)\frac{\lambda_{N}^{2}(\overline{L})}{\lambda_{2}(\overline{L})}+\beta^{2}(i)\frac{\lambda_{N}^{3}(\overline{L})}{\lambda_{2}(\overline{L})}+\alpha^{2}(i)+\beta^{2}(i)\frac{c_{1}\lambda_{N}(\overline{L})}{\lambda_{2}}+ \\
&\left.+2\beta^{2}(i)\frac{Rc_{1}\lambda_{N}(\overline{L})}{\lambda_{2}(\overline{L})}\right]\mathbf{y}_{R,\mathcal{C}^{\perp}}^{T}(i)\left(\overline{L}\otimes I_{M}\right)\mathbf{y}_{R,\mathcal{C}^{\perp}}(i)+ \\
&+\left[\frac{1}{2\tau_{2}-\varepsilon_{1}}\left(2Rc_{1}\lambda_{N}(\overline{L})+R^{2}c_{1}\lambda_{N}(\overline{L})\right)+\rho\varepsilon_{2}i^{\varepsilon_{2}-1}\right]
\end{aligned} \tag{128}$$

For the first term on the R.H.S. of (128) involving $\mathbf{y}_{R,\mathcal{C}^{\perp}}^{T}(i)\left(\overline{L}\otimes I_{M}\right)\mathbf{y}_{R,\mathcal{C}^{\perp}}(i)$, the coefficient $-2\beta(i)(i+1)^{\varepsilon_{1}}$ dominates all other coefficients eventually ($\tau_{2}<1$ by Assumption **(D.5)**); hence, the first term on the R.H.S. of (128) becomes negative eventually (for sufficiently large i). The second term on the R.H.S. of (128) becomes negative eventually because $\rho\varepsilon_{2}<0$ and $1-\varepsilon_{2}<2\tau_{2}-\varepsilon_{1}$ by assumption. Hence there exists sufficiently large i, say i_{R}, such that,

$$\mathbb{E}_{\theta^{*}}\left[V(i+1,\mathbf{y}_{R}(i+1))\,\middle|\,\mathcal{F}_{i}\right]-V(i,\mathbf{y}_{R}(i))\leq 0,\forall i\geq i_{R}.$$

This shows $\{V(i,\mathbf{y}_{R}(i))\}_{i\geq i_{R}}$ is a non-negative supermartingale w.r.t. the filtration $\{\mathcal{F}_{i}\}_{i\geq i_{R}}$. Thus, $\{V(i,\mathbf{y}_{R}(i))\}_{i\geq i_{R}}$ converges a.s. to a finite random variable (see [50]). Clearly, the sequence $\rho i^{\varepsilon_{2}}$ goes to zero as $\varepsilon_{2}<0$. Then:

$$\mathbb{P}_{\theta^{*}}\left[\lim_{i\to\infty}i^{\varepsilon_{1}}\mathbf{y}_{R}^{T}(i)\left(\overline{L}\otimes I_{M}\right)\mathbf{y}_{R}(i)\text{ exists and is finite}\right]=1$$

Since $i^{\varepsilon_{1}}\to\infty$ as $i\to\infty$, it follows

$$\mathbb{P}_{\theta^{*}}\left[\lim_{i\to\infty}\mathbf{y}_{R}^{T}(i)\left(\overline{L}\otimes I_{M}\right)\mathbf{y}_{R}(i)=0\right]=1 \tag{129}$$

Since $\mathbf{y}_{R}^{T}(i)\left(\overline{L}\otimes I_{M}\right)\mathbf{y}_{R}(i)\geq\lambda_{2}(\overline{L})\left\|\mathbf{y}_{R,\mathcal{C}^{\perp}}(i)\right\|^{2}$, from (129) we have

$$\mathbb{P}_{\theta^{*}}\left[\lim_{i\to\infty}\mathbf{y}_{R,\mathcal{C}^{\perp}}(i)=0\right]=1 \tag{130}$$

To establish (121) we note that

$$\mathbf{y}_{R,\mathcal{C}}(i)=\mathbf{1}_{N}\otimes\mathbf{y}_{R,\text{avg}}(i) \tag{131}$$

where

$$\mathbf{y}_{R,\text{avg}}(i+1)=(1-\alpha(i))\,\mathbf{y}_{R,\text{avg}}(i)$$

Since $\sum_{i\geq 0}\alpha(i)=\infty$, it follows from standard arguments that $\mathbf{y}_{R,\text{avg}}(i)\to 0$ as $i\to\infty$. We then have from (131)

$$\mathbb{P}_{\theta^{*}}\left[\lim_{i\to\infty}\mathbf{y}_{R,\mathcal{C}}(i)=0\right]=1$$

which together with (130) establishes (121). The claim in (59) then follows from the arguments above.

We now prove the claim in (60). Recall the matrix P in (103). Using the fact,

$$P\left(L(i)\otimes I_{M}\right)=P\left(\overline{L}\otimes I_{M}\right)=\mathbf{0},\quad\forall i$$

we have

$$\begin{aligned}
P\widetilde{\mathbf{x}}(i+1)=P\widetilde{\mathbf{x}}(i)&-\alpha(i)\left[P\widetilde{\mathbf{x}}(i)-PJ(\mathbf{z}(i))\right] \\
&-\beta(i)P\left(\mathbf{\Upsilon}(i)+\mathbf{\Psi}(i)\right)
\end{aligned}$$

and similarly

$$\begin{aligned} P\widehat{\mathbf{x}}(i+1) &= P\widehat{\mathbf{x}}(i) - \alpha(i)\left[P\widehat{\mathbf{x}}(i) - PJ(\mathbf{z}(i))\right] \\ &\quad - \beta(i)P\left(\boldsymbol{\Upsilon}(i) + \boldsymbol{\Psi}(i)\right) \end{aligned}$$

Since the sequences $\{P\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ and $\{P\widehat{\mathbf{x}}(i)\}_{i\geq 0}$ follow the same recursion and start with the same initial state $P\widetilde{\mathbf{x}}(0)$, they are equal, and we have $\forall i$

$$\begin{aligned} P\mathbf{y}(i) &= P\left(\widetilde{\mathbf{x}}(i) - \widehat{\mathbf{x}}(i)\right) \\ &= 0 \end{aligned}$$

From (119) we then have

$$\begin{aligned} \mathbf{y}(i+1) &= \left[I_{NM} - \beta(i)\left(\overline{L}\otimes I_M\right) - \alpha(i)I_{NM} - P\right]\mathbf{y}(i) - \\ &\quad - \beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\mathbf{y}(i) - \beta(i)\left(\widetilde{L}(i)\otimes I_M\right)\widehat{\mathbf{x}}(i) \end{aligned}$$

By Lemma 26, to prove the claim in (59), it suffices to prove

$$\lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right] = 0$$

From Lemma 26, we note that the sequence $\{\widehat{\mathbf{x}}(i)\}_{i\geq 0}$ converges in $\mathcal{L}_2$ to $\mathbf{1}_N\otimes h(\theta^*)$ and hence $\mathcal{L}_2$ bounded, i.e., there exists constant $c_3 > 0$, such that,

$$\sup_{i\geq 0}\mathbb{E}_{\theta^*}\left[\|\widehat{\mathbf{x}}(i)\|^2\right] \leq c_3 < \infty$$

Choose j large enough, such that, for $i \geq j$

$$\left\|I_{NM} - \beta(i)\left(\overline{L}\otimes I_M\right) - \alpha(i)I_{NM} - P\right\| \leq 1 - \beta(i)\lambda_2(\overline{L})$$

Noting that $\widetilde{L}(i)$ is independent of $\mathcal{F}_i$ and $\left\|\widetilde{L}(i)\right\| \leq c_2$ for some constant $c_2 > 0$, we have for $i \geq j$,

$$\begin{aligned} &\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i+1)\|^2\right] = \\ &\mathbb{E}_{\theta^*}\Big[\mathbf{y}^T(i)\left(I_{NM} - \beta(i)\left(\overline{L}\otimes I_M\right) - \alpha(i)I_{NM} - P\right)^2\mathbf{y}(i) \\ &+\beta^2(i)\mathbf{y}^T(i)\left(\widetilde{L}(i)\right)^2\mathbf{y}(i) + \beta^2(i)\widehat{\mathbf{x}}^T(i)\left(\widetilde{L}(i)\right)^2\widehat{\mathbf{x}}(i) \\ &+\beta^2(i)\mathbf{y}^T(i)\left(\widetilde{L}(i)\right)^2\widehat{\mathbf{x}}(i)\Big] \\ &\leq \left(1 - \beta(i)\lambda_2(\overline{L})\right)\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right] + c_2^2\beta^2(i)\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right] \\ &+ c_2^2c_3\beta^2(i) + \left(2\beta^2(i)c_2^2c_3^{\frac{1}{2}}\right)\mathbb{E}_{\theta^*}^{\frac{1}{2}}\left[\|\mathbf{y}(i)\|^2\right] \\ &\leq \left(1 - \beta(i)\lambda_2(\overline{L}) + c_2^2\beta^2(i) + 2\beta^2(i)c_2^2c_3^{\frac{1}{2}}\right)\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right] \\ &+ \beta^2(i)\left(c_2^2c_3 + 2c_2^2c_3^{\frac{1}{2}}\right) \end{aligned} \tag{133}$$

where in the last step we used the inequality

$$\mathbb{E}_{\theta^*}^{\frac{1}{2}}\left[\|\mathbf{y}(i)\|^2\right] \leq \mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right] + 1$$

Now similar to Lemma 23, choose $j_1 \geq j$ and $0 < c_4 < \lambda_2(\overline{L})$, such that,

$$1 - \beta(i)\lambda_2(\overline{L}) + c_2^2\beta^2(i) + 2\beta^2(i)c_2^2c_3^{\frac{1}{2}} \leq 1 - \beta(i)c_4, \quad \forall i \geq j_1$$

Then, for $i \geq j_1$, from (133)

$$\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i+1)\|^2\right] \leq (1-\beta(i)c_4)\,\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right] + \\ + \beta^2(i)\left(c_2^2 c_3 + 2c_2^2 c_3^{\frac{1}{2}}\right)$$

from which we conclude that $\lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\|\mathbf{y}(i)\|^2\right]=0$ by Lemma 25 (see also Lemma 23.)

■

Proof of Theorem 22

Proof: Consistency follows from the fact that by Theorem 21 the sequence $\{\widetilde{\mathbf{x}}(i)\}_{i\geq 0}$ converges a.s. to $\mathbf{1}_N \otimes h(\theta^*)$, and the function $h^{-1}(\cdot)$ exists and is continuous on the open set $\mathcal{U}$.

To establish the second claim, we note that, if $h^{-1}(\cdot)$ is Lipschitz continuous, there exists constant $k>0$, such that

$$\left\|h^{-1}(\widetilde{\mathbf{y}}_1) - h^{-1}(\widetilde{\mathbf{y}}_2)\right\| \leq k\left\|\widetilde{\mathbf{y}}_1 - \widetilde{\mathbf{y}}_2\right\|, \quad \forall\, \widetilde{\mathbf{y}}_1, \widetilde{\mathbf{y}}_2 \in \mathbb{R}^M$$

Since $\mathcal{L}_2$ convergence implies $\mathcal{L}_1$, we then have from Theorem 21 for $1 \leq n \leq N$

$$\begin{aligned}\lim_{i\to\infty}\|\mathbb{E}_{\theta^*}[\mathbf{x}_n(i)-\theta^*]\| &\leq \lim_{i\to\infty}\mathbb{E}_{\theta^*}[\|\mathbf{x}_n(i)-\theta^*\|] \\ &= \lim_{i\to\infty}\mathbb{E}_{\theta^*}\left[\left\|h^{-1}(\widetilde{\mathbf{x}}_n(i)) - h^{-1}(h(\theta^*))\right\|\right] \\ &\leq k\lim_{i\to\infty}\mathbb{E}_{\theta^*}[\|\widetilde{\mathbf{x}}_n(i)-h(\theta^*)\|] \\ &= 0\end{aligned}$$

which establishes the theorem. ■

References

[1] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1520–1533, Sept. 2004.

[2] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. AC-48, no. 6, pp. 988–1001, June 2003.

[3] L. Xiao and S. Boyd, "Fast linear iterations for distributed averaging," *Syst. Contr. Lett.*, vol. 53, pp. 65–78, 2004.

[4] S. Kar and J. M. F. Moura, "Sensor networks with random links: Topology design for distributed consensus," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3315–3326, July 2008.

[5] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks with communication channel noise and random link failures," in *41st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2007.

[6] S. Kar and J. M. F. Moura, "Distributed average consensus in sensor networks with quantized inter-sensor communication," in *Proceedings of the 33rd International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, April 1-4 2008.

[7] Y. Hatano and M. Mesbahi, "Agreement over random networks," in *43rd IEEE Conference on Decision and Control*, vol. 2, Dec. 2004, pp. 2010–2015.

[8] T. C. Aysal, M. Coates, and M. Rabbat, "Distributed average consensus using probabilistic quantization," in *IEEE/SP 14th Workshop on Statistical Signal Processing Workshop*, Maddison, Wisconsin, USA, August 2007, pp. 640–644.

[9] M. E. Yildiz and A. Scaglione, "Differential nested lattice encoding for consensus problems," in *ACM/IEEE Information Processing in Sensor Networks*, Cambridge, MA, April 2007.

[10] A. Kashyap, T. Basar, and R. Srikant, "Quantized consensus," *Automatica*, vol. 43, pp. 1192–1203, July 2007.

[11] P. Frasca, R. Carli, F. Fagnani, and S. Zampieri, "Average consensus on networks with quantized communication," *Submitted to the Int. J. Robust and Nonlinear Control*, 2008.

[12] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *Technical Report 2778, LIDS-MIT*, Nov. 2007.

[13] 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.

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

[15] A. Das and M. Mesbahi, "Distributed linear parameter estimation in sensor networks based on Laplacian dynamics consensus algorithm," in *3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks*, vol. 2, Reston, VA, USA, 28-28 Sept. 2006, pp. 440–449.

[16] R. Olfati-Saber, "Distributed Kalman filter with embedded consensus filters," in *CDC-ECC'05, 44th IEEE Conference on Decision and Control and 2005 European Control Conference*, 2005, pp. 8179–8184.

[17] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in Ad Hoc WSNs with noisy links - part I: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, January 2008.

[18] U. A. Khan and J. M. F. Moura, "Distributing the Kalman filter for large-scale systems," *IEEE Transactions on Signal Processing*, vol. 56, no. 10, p. 49194935, October 2008.

[19] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering using consensus strategies," *IEEE Journal on Selected Areas of Communications*, vol. 26, no. 4, pp. 622–633, September 2008.

[20] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. M. F. Moura, "Distributed detection via Gaussian running consensus: Large deviations asymptotic analysis," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, May 2011.

[21] D. Jakovetic, J. M. F. Moura, and J. Xavier, "Distributed detection over noisy networks: Large deviations analysis," August 2011, accepted for publication, *IEEE Transactions on Signal Processing*, vol. 60, 2012.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE Journal on Selected Topics on Signal Processing*, vol. 5, 2011.

[27] S. Kar and J. M. F. Moura, "Gossip and distributed Kalman filtering: Weak consensus under weak detectability," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1766-1784, April 2011.

[28] D. Jakovetic, J. ao Xavier, and J. M. F. Moura, "Weight optimization for consenus algorithms with correlated switching topology," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3788–3801, July 2010.

[29] J. N. Tsitsiklis, "Problems in decentralized decision making and computation," Ph.D., Massachusetts Institute of Technology, Cambridge, MA, 1984.

[30] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. AC-31, no. 9, pp. 803–812, September 1986.

[31] 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.

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

[33] S. Kar, S. A. Aldosari, and J. M. F. Moura, "Topology for distributed inference on graphs," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2609–2613, June 2008.

[34] A. H. Sayed, *Fundamentals of Adaptive Filtering*, ser. Wiley-Interscience. Hoboken, NJ: John Wiley & Sons, 2003.

[35] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms: Stability and Performance*. Prentice Hall, 1995.

[36] L. Ljung, *System Identification: Theory for the User*. Prentice Hall, 1999.

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

[38] 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.

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

[40] B. Mohar, "The Laplacian spectrum of graphs," in *Graph Theory, Combinatorics, and Applications*, Y. Alavi, G. Chartrand, O. R. Oellermann, and A. J. Schwenk, Eds. New York: J. Wiley & Sons, 1991, vol. 2, pp. 871–898.

[41] B. Bollobas, *Modern Graph Theory*. New York, NY: Springer Verlag, 1998.

[42] L. Schuchman, "Dither signals and their effect on quantization noise," *IEEE Trans. Commun. Technol.*, vol. COMM-12, pp. 162–165, December 1964.

[43] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: A theoretical survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375, May 1992.

[44] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 442–448, October 1977.

[45] R. M. Gray and T. G. Stockham, "Dithered quantizers," *IEEE Trans. Information Theory*, vol. 39, pp. 805–811, May 1993.

[46] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1383–1400, March 2010, initial post: Dec. 2007. [Online]. Available: http://arxiv.org/abs/0712.1609

[47] E. Lehmann, *Theory of Point Estimation*. John Wiley and Sons, Inc., 1983.

[48] H.-F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*. Boston: Birkhauser, 1991.

[49] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2508–2530, 2006.

[50] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. Springer Series in Statistics., 2002.

[51] S. Kar, "Large scale networked dynamical systems: Distributed inference," Ph.D., Carnegie Mellon University, Pittsburgh, PA, 2010.

[52] N. Krasovskii, *Stability of Motion*. Stanford University Press, 1963.

[53] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Prentice-Hall, 1975.

[54] M. D. Ilic' and J. Zaborszky, *Dynamics and Control of Large Electric Power Systems*. Wiley, 2000.

[55] J. S. Thorp, A. G. Phadke, and K. J. Karimi, "Real time voltage-phasor measurements for static state estimation," *IEEE Transactions on Power Apparatus and Systems*, vol. 104, no. 11, pp. 3098–3106, November 1985.

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

[57] Y. Chow, "Some convergence theorems for independent random variables," *Ann. Math. Statist.*, vol. 37, pp. 1482–1493, 1966.

[58] Y. Chow and T. Lai, "Limiting behavior of weighted sums of independent random variables," *Ann. Prob.*, vol. 1, pp. 810–824, 1973.

[59] W. Stout, "Some results on the complete and almost sure convergence of linear combinations of independent random variables and martingale differences," *Ann. Math. Statist.*, vol. 39, pp. 1549–1562, 1968.

PLACE
PHOTO
HERE

Soummya Kar (S'05–M'10) received the B.Tech. degree in Electronics and Electrical Communication Engineering from the Indian Institute of Technology, Kharagpur, India, in May 2005 and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 2010. From June 2010 to May 2011 he was with the EE Department at Princeton University as a Postdoctoral Research Associate. He is currently an Assistant Research Professor of ECE at Carnegie Mellon University. His research interests include performance analysis and inference in large-scale networked systems, adaptive stochastic systems, stochastic approximation, and large deviations.

PLACE
PHOTO
HERE

José M. F. Moura (S'71–M'75–SM'90–F'94) received degrees from Instituto Superior Técnico (IST), Portugal, and from the Massachusetts Institute of Technology (MIT), Cambridge, MA. He is University Professor at Carnegie Mellon University (CMU), was on the faculty at IST and has been visiting Professor at MIT. His interests include statistical and algebraic signal and image processing.

Dr. Moura is Division IX Director of the IEEE, was President of the IEEE Signal Processing Society (SPS), and Editor in Chief of the IEEE Transactions on Signal Processing. Dr. Moura is a Fellow from IEEE and the AAAS and an Academy of Sciences of Portugal corresponding member. He has received several awards including the SPS Technical Achievement Award. In 2010, he was elected University Professor at CMU.

PLACE
PHOTO
HERE

Kavita Ramanan received her Ph.D. from the Division of Applied Mathematics at Brown University in 1998. After a post-doctoral position at the Technion in Haifa, Israel, Dr. Ramanan joined the Mathematical Sciences Center of Bell Laboratories, Lucent, as a Member of Technical Staff. In 2003, Dr. Ramanan returned to academia as an Associate Professor at the Department of Mathematical Sciences at Carnegie Mellon University (CMU), Pittsburgh. In 2010, she moved to the Division of Applied Mathematics at Brown University where she is a Professor of Applied Mathematics. Dr. Ramanan works on probability theory, stochastic processes and their applications. Since 2008, she has also been an adjunct professor at the Chennai Mathematical Institute, India.