NONLINEAR CONSENSUS+INNOVATIONS UNDER CORRELATED HEAVY-TAILED NOISES: MEAN SQUARE CONVERGENCE RATE AND ASYMPTOTICS

MANOJLO VUKOVIC*, DUSAN JAKOVETIC†, DRAGANA BAJOVIC‡, AND SOUMMYA KAR§

Abstract. We consider distributed recursive estimation of consensus+innovations type in the presence of heavy-tailed sensing and communication noises. We allow that the sensing and communication noises are mutually correlated while independent identically distributed (i.i.d.) in time, and that they may both have infinite moments of order higher than one (hence having infinite variances). Such heavy-tailed, infinite-variance noises are highly relevant in practice and are shown to occur, e.g., in dense internet of things (IoT) deployments. We develop a consensus+innovations distributed estimator that employs a general nonlinearity in both consensus and innovations steps to combat the noise. We establish the estimator's almost sure convergence, asymptotic normality, and mean squared error (MSE) convergence. Moreover, we establish and explicitly quantify for the estimator a sublinear MSE convergence rate. We then quantify through analytical examples the effects of the nonlinearity choices and the noises correlation on the system performance. Finally, numerical examples corroborate our findings and verify that the proposed method works in the simultaneous heavy-tail communication-sensing noise setting, while existing methods fail under the same noise conditions.

Key words. nonlinear mappings, consensus+innovations, distributed estimation, heavy-tailed noise, mean square convergence rate, correlated noises

AMS subject classifications. 93E10, 93E35, 60G35, 94A13, 62M05

1. Introduction.

We consider a distributed estimation problem where a network of agents cooperates to estimate an unknown static vector parameter $\boldsymbol{\theta}^* \in \mathbb{R}^M$. Specifically, we are interested in *consensus+innovations* distributed estimation, e.g., [18, 16, 17]. With consensus+innovations, each agent iteratively updates its unknown parameter's estimate by 1) exchanging its estimate with immediate neighbors in the network; and 2) assimilating a newly acquired observation (measurement).

Consensus+innovations distributed estimators have been extensively studied, e.g., [18, 16, 17]; see also [20, 22, 23, 27, 30, 24, 38] for related diffusion-type and other methods. Typically, such distributed estimators exhibit strong convergence guarantees under various imperfection models (noises) in 1) sensing (observations) and/or 2) inter-agent communications. For example, reference [18] establishes almost sure (a.s.) convergence and asymptotic normality of the estimators developed therein. The authors of [18] allow for an observation noise with finite variance and a network model that accounts for random link failures and dithered quantization (effectively an additive noise with finite variance). Reference [16] considers consensus+innovations distributed estimation in the presence of random link failures without quantization or additive noise, and it develops estimators that are asymptotically efficient, i.e., that achieve the minimal possible asymptotic variance. The authors of [17] propose adap-

*University of Novi Sad, Faculty of Technical Sciences, Department of Fundamental Sciences (manojlo.vukovic@uns.ac.rs).

†University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics (dusan.jakovetic@dmi.uns.ac.rs).

‡University of Novi Sad, Faculty of Technical Sciences, Department of Power, Electronic and Communication Engineering (dbajovic@uns.ac.rs).

§Department of Electrical and Computer Engineering, Carnegie Mellon University (soummyak@andrew.cmu.edu).

tive asymptotically efficient estimators, wherein the innovation gains are adaptively learned during the algorithm progress. Consensus+innovations distributed detection and related distributed detection methods have also been considered, e.g., [25, 3, 2, 14]. The above distributed estimation and distributed detection-related works typically assume that the noises have finite moments of a certain order greater than two, and hence they have finite variance.

It is highly relevant to investigate distributed estimators in the presence of heavy-tailed *communication* and *sensing* noises, as they arise in many application scenarios. For example, edge devices in Internet of Things (IoT) systems or sensor networks can be subject to noise distributions that may not have finite moments of order higher than one, e.g., [6, 31, 12, 37, 11, 7], like, e.g., symmetric α-stable noise distributions. This effect may occur due to interference, e.g., when wireless sensor network is relatively densely deployed. In this case, the signals of neighboring nodes interfere with each other and corrupt the signal to be received. References [10, 36] analyze the probability distribution of the interference and demonstrate that it has heavy-tails. More precisely, [10, 36] show that the interference power has an alpha-stable distribution in a network with infinite radius and no guard zone when the interferers are placed according to a Poisson point process, where alpha depends on the path loss coefficient between the interferers and the receiver (see [10, 36] for details). Empirical evidence for the emergence of heavy-tail interference noise in certain IoT systems has been provided in [6].

Moreover, observation and communication noises may be mutually correlated due to the common interference processes in the environment that the sensing and communication devices are exposed to.

Several recent works [19, 21, 35, 33, 5, 1, 4, 26] consider distributed estimation methods in the presence of *impulsive observations noise*,[1] but still assuming a *finite noise variance* and *no communication noise*. For example, reference [19] introduces a method based on Wilcoxon-norm; [21] utilizes a Huber-loss function; and [35] adopts a mean error minimization approach. Robust distributed estimation methods based on adaptive subgradient projections are considered in [33, 5]. To cope with the impulsive observation noise, several references employ a certain *nonlinearity* in the innovation step. Reference [1] develops a method that adaptively learns an optimized nonlinearity at the innovation step for each agent in the network. Reference [4] employs a saturation nonlinearity in the innovation step to cope with measurement attacks. Further results on distributed estimation under impulsive observations noise can be found in a recent survey [26]. Very recently, we have developed a consensus+innovations distributed estimator [15] that provably works under a heavy-tailed communications noise and a light-tailed observations noise. Specifically, under the assumed setting, [15] establishes almost sure convergence and asymptotic normality of the method therein. However, [15] is not concerned with mean squared error (MSE) rate analysis of the method. While asymptotic normality is a useful result that provides the algorithm's rate of convergence (in the weak convergence sense) *asymptotically*, it does not capture the (MSE) algorithm behavior in non-asymptotic regimes.

In summary, we identify for the current literature the following major gaps with respect to design and analysis of distributed estimation methods under heavy-tailed

[1] As explained in, e.g., [1], an impulsive noise may be described as one whose realizations contain sparse, random samples of amplitude much higher than nominally accounted for. Impulsive noise may have a finite or infinite variance. Existing works on distributed estimation in impulsive noises assume a *finite noise variance*.

noises. 1) All existing works assume a finite observations noise variance. That is, even when impulsive observation noise is assumed, existing works still require the variance of the noise to be finite. This assumption can be restrictive and is violated for several commonly used heavy-tail noise models like α-stable distributions [11]. 2) No existing work simultaneously handles heavy-tailed (infinite-variance) sensing and heavy-tailed (infinite-variance) observation noises. 3) MSE convergence rate analysis has not been developed for distributed estimation in the presence of either infinite-variance sensing and/or infinite-variance communication noises. 4) Existing works on distributed estimation in the presence of infinite-variance (either sensing and/or communication noises) assume mutually independent sensing and communication noises.

Contributions. In this paper, we close the gaps identified above by developing a nonlinear consensus+innovations distributed estimator that provably works under the simultaneous presence of correlated heavy-tailed (infinite variance) observation and communication noises. We allow for a very general model of the sensing and communication noises, only assuming that they exhibit symmetric zero-mean distributions with finite first moments. Hence, the variances of both sensing and communication noises may be infinite. Moreover, we allow that, for a fixed time instant t, the additive sensing and communication noises may be mutually dependent, while they are both independent identically distributed (i.i.d.) in time. The proposed estimator employs a generic nonlinearity both at the innovations and the consensus terms. The encompassed nonlinearities are very general and include a broad class of (possibly discontinuous) odd functions, such as the component-wise sign and clipping functions. We establish for the proposed estimator almost sure convergence, asymptotic normality, and we explicitly evaluate the corresponding asymptotic variance. Furthermore, we establish for the proposed method, under a carefully designed step size sequence, a MSE convergence rate $O(1/t^{\kappa})$, and we quantify the rate $\kappa \in (0,1)$ in terms of the system parameters. In addition, we quantify through analytical examples the effects of correlation between sensing and observation noises, and we demonstrate how the derived asymptotic covariance results may be used as a guideline to optimize the employed nonlinearities for a problem at hand. Finally, we compare the proposed method with existing works in [1] and [15], both through analytical examples and by simulation. Most notably, we show that the existing methods fail to converge under the simultaneous presence of heavy-tailed (infinite-variance) observation and communication noises, while the proposed method provably works in the heavy-tailed setting.

Paper organization. Section 2 provides a description of the distributed estimation model that is considered and also gives all basic assumptions. In Section 3, we present the proposed nonlinear consensus+innovations estimator. Section 4 establishes almost sure convergence, asymptotic normality and the MSE rate of the proposed distributed estimator. Section 5 presents analytical and numerical examples. The conclusion is given in Section 6. Some auxiliary supporting arguments are provided in [34].

Notation. We denote by $\mathbb{R}$ the set of real numbers and by $\mathbb{R}^m$ the m-dimensional Euclidean real coordinate space. We use normal lower-case letters for scalars, lower case boldface letters for vectors, and upper case boldface letters for matrices. Further, to represent a vector $\mathbf{a} \in \mathbb{R}^m$ through its component, we write $\mathbf{a} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m]^\top$ and we denote by: $\mathbf{a}_i$ or $[\mathbf{a}_i]$, as appropriate, the i-th element of vector $\mathbf{a}$; $\mathbf{A}_{ij}$ or $[\mathbf{A}_{ij}]$, as appropriate, the entry in the i-th row and j-th column of a matrix $\mathbf{A}$; $\mathbf{A}^\top$ the transpose of a matrix $\mathbf{A}$; $\otimes$ the Kronecker product of matrices. Further, we use either $\mathbf{a}^\top \mathbf{b}$ or $\langle \mathbf{a}, \mathbf{b} \rangle$ for the inner products of vectors $\mathbf{a}$ and $\mathbf{b}$. Next, we let $\mathbf{I}$, $\mathbf{0}$, and

$\mathbf{1}$ be, respectively, the identity matrix, the zero vector, and the column vector with unit entries; Diag($\mathbf{a}$) the diagonal matrix whose diagonal entries are the elements of vector $\mathbf{a}$; $\mathbf{J}$ the $N \times N$ matrix $\mathbf{J} := (1/N)\mathbf{1}\mathbf{1}^\top$. When appropriate, we indicate the matrix or vector dimension through a subscript. Next, $\mathbf{A} \succ 0\,(\mathbf{A} \succeq 0)$ means that the symmetric matrix A is positive definite (respectively, positive semi-definite). We further denote by: $\|\cdot\| = \|\cdot\|_2$ the Euclidean (respectively, spectral) norm of its vector (respectively, matrix) argument; $\lambda_i(\cdot)$ the i-th smallest eigenvalue; $g'(v)$ the derivative evaluated at v of a function $g : \mathbb{R} \to \mathbb{R}$; $\nabla h(\mathbf{w})$ and $\nabla^2 h(\mathbf{w})$ the gradient and Hessian, respectively, evaluated at w of a function $h : \mathbb{R}^m \to \mathbb{R}$, $m > 1$; $\mathbb{P}(\mathcal{A})$ and $\mathbb{E}[u]$ the probability of an event $\mathcal{A}$ and expectation of a random variable u, respectively; and by sign(a) the sign function, i.e., sign(a) = 1, for $a > 0$, sign(a) = −1, for $a < 0$, and sign(0) = 0. Finally, for two positive sequences η_n and χ_n, we have: $\eta_n = O(\chi_n)$ if $\limsup_{n\to\infty} \frac{\eta_n}{\chi_n} < \infty$.

2. Problem model and basic assumptions. We consider a network of N agents (sensors), through which the parameter of interest $\boldsymbol{\theta}^* \in \mathbb{R}^M$ is to be estimated. At each time $t = 0, 1, ...,$ each agent $i = 1, 2, ..., N$ observes parameter $\boldsymbol{\theta}^*$ following the linear regression model:

$$z_i^t = \mathbf{h}_i^\top \boldsymbol{\theta}^* + n_i^t. \tag{2.1}$$

Here, $z_i^t \in \mathbb{R}$ is the observation, $\mathbf{h}_i \in \mathbb{R}^M$ is the deterministic, non-zero regression vector known only by agent i and $n_i^t \in \mathbb{R}$ is the observation noise. The underlying topology is modeled via a graph $G = (V, E)$, where $V = \{1, ..., N\}$ is the set of agents and E is the set of links, i.e., $\{i, j\} \in E$ if there exists a link between agents i and j. We also define the set of all arcs E_d in the following way: if $\{i, j\} \in E$ then $(i, j) \in E_d$ and $(j, i) \in E_d$. We denote by $\Omega_i = \{j \in V : \{i, j\} \in E\}$ set of neighbors of agent i (excluding i) and by $\mathbf{D} = \mathrm{Diag}(\{d_i\})$ the degree matrix, where $d_i = |\Omega_i|$ is the number of neighbors of agent i. The graph Laplacian matrix $\mathbf{L}$ is defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{A}$ is the adjacency matrix, which is a zero-one symmetric matrix with zero diagonal, such that, for $i \neq j$, $\mathbf{A}_{ij} = 1$ if and only if $\{i, j\} \in E$. Let us denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space.
We make the following assumptions.

Assumtion 2.1. **Network model and Observability:**
1. Graph $G = (V, E)$ is undirected, simple (no self or multiple links) and static;
2. The matrix $\sum_{i=1}^N \mathbf{h}_i \mathbf{h}_i^\top$ is invertible;

The condition 2 in Assumption 2.1 ensures that (2.1) is observable, i.e., a centralized estimator (e.g., least squares) that collects all $z_i^t, i = 1, 2, ..., N$, for all t, and has knowledge of all vectors $\mathbf{h}_i, i = 1, 2, ..., N$, is consistent.

Assumtion 2.2. **Observation noise:**
1. For each agent $i = 1, ..., N$, the observation noise sequence $\{n_i^t\}$ in (2.1), is independent identically distributed (i.i.d.);
2. At each agent $i = 1, ..., N$ at each time $t = 0, 1, ...,$ noise n_i^t has the same probability density function p_{o}.
3. Random variables n_i^t and n_j^s are mutually independent whenever the tuple (i, t) is different from (j, s);
4. The pdf p_{o} is symmetric, i.e. $p_{\mathrm{o}}(u) = p_{\mathrm{o}}(-u)$, for every $u \in \mathbb{R}$, and $p_{\mathrm{o}}(u) > 0$ for $|u| \leq c_{\mathrm{o}}$, for some constant $c_{\mathrm{o}} > 0$;
5. There holds that with $\int |u| p_{\mathrm{o}}(u) du < \infty$.

If there is an arc between agents i and j, i.e., $(i, j) \in E_d$, we denote by $\boldsymbol{\xi}_{ij}^t$ communication noise that is injected when agent j communicates to agent i at time instant t

(see ahead algorithm (3.1)).

Assumtion 2.3. **Communication noise:**
1. Additive communication noise $\{\boldsymbol{\xi}_{ij}^t\}$, $\boldsymbol{\xi}_{ij}^t \in \mathbb{R}^M$ is i.i.d. in time t, and independent across different arcs $(i, j) \in E_d$.
2. Each random variable $[\boldsymbol{\xi}_{ij}^t]_\ell$, for each $t = 0, 1...$, for each arc (i, j), for each entry $\ell = 1, ..., M$, has the same probability density function p_c.
3. The pdf p_c is symmetric, i.e. $p_o(u) = p_c(-u)$, for every $u \in \mathbb{R}$ and $p_c(u) > 0$ for $|u| \leq c_c$, for some constant $c_c > 0$;
4. There holds that $\int |u| p_c(u) du < \infty$.

Remark 2.4. Notice here that from the symmetry of the probability density functions p_o and p_c, it follows that both of the distributions are zero mean. Moreover, notice that we do not assume that observation and communication noises are mutually independent for a fixed t. However, they are both i.i.d. in time.

Remark 2.5. Condition 2 in Assumptions 2.2 and 2.3 can be relaxed in the sense that it can be assumed that $\mathbf{n}^t$ has joint probability density function p_o and $\boldsymbol{\xi}_{ij}^t$ has the joint probability density function $p_{c,ij}$. (see Appendix C in [34]). The reason why there is condition 4 in the Assumption 2.2 and condition 3 in the Assumption 2.3 will become clear later.

For future reference, a compact vector form of (2.1) is:

$$\mathbf{z}^t = \mathbf{H} \left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) + \mathbf{n}^t, \tag{2.2}$$

where, $\mathbf{z}^t = [z_1^t, z_2^t, ..., z_N^t]^\top \in \mathbb{R}^N$ is the observation vector, $\mathbf{H} \in \mathbb{R}^{N \times (MN)}$ is the regression matrix whose i-th row vector equals $[\mathbf{0}, ..., \mathbf{0}, \mathbf{h}_i^\top, \mathbf{0}, .., \mathbf{0}] \in \mathbb{R}^{MN}$, where the i-th block of size M equals $\mathbf{h}_i^\top$, and the other M-size blocks are the zero vectors; and $\mathbf{n}^t = [n_1^t, n_2^t, ..., n_N^t]^\top \in \mathbb{R}^N$ is the noise vector at time t.

3. Proposed algorithm. In order to estimate the unknown parameter $\boldsymbol{\theta}^* \in \mathbb{R}^M$, in the presence of heavy-tailed observation noise and heavy-tailed communication noise, each agent uses a nonlinear consensus+innovations strategy. Therein, the impact of the two heavy-tailed noises is mitigated by nonlinearities that have been added to both consensus and innovation steps.
In more detail, each agent i at each time $t = 0, 1, ...$, generates a sequence of estimates $\{\mathbf{x}_i^t\}_{t \geq 0}$ of unknown parameter $\boldsymbol{\theta}^*$ by the following algorithm:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha_t \left(\frac{b}{a} \sum_{j \in \Omega_i} \boldsymbol{\Psi}_c \left(\mathbf{x}_i^t - \mathbf{x}_j^t + \boldsymbol{\xi}_{ij}^t \right) - \mathbf{h}_i \Psi_o \left(z_i^t - \mathbf{h}_i^\top \mathbf{x}_i^t \right) \right). \tag{3.1}$$

Here, α_t is a step-size, and $a, b > 0$ are constants. We consider a family of decaying step-size choices $\alpha_t = a/(t+1)^\delta$, $\delta \in (0.5, 1]$. As shown later, the step-size (values of a and δ) should be designed appropriately in order for good properties (e.g., a.s. convergence, MSE rate guarantees) of the algorithm to hold. Functions $\Psi_o : \mathbb{R} \to \mathbb{R}$ and $\boldsymbol{\Psi}_c : \mathbb{R}^M \to \mathbb{R}^M$ are non-linear functions and function $\boldsymbol{\Psi}_c$ operates component-wise by abusing notation, i.e., for $\mathbf{y} \in \mathbb{R}^M$, we set that $\boldsymbol{\Psi}_c(\mathbf{y}) = [\Psi_c(\mathbf{y}_1), \Psi_c(\mathbf{y}_2), ..., \Psi_c(\mathbf{y}_M)]$. Also, functions Ψ_c and Ψ_o satisfy Assumption 3.1. We compare the proposed method (3.1) with the $\mathcal{LU}$ scheme in [18] and the scheme in [15]. Compared with these schemes, (3.1) introduces a nonlinearity in the innovation step as well. $\mathcal{LU}$ is obtained from (3.1) by setting both of the nonlinearities Ψ_o and Ψ_o to identity functions and $\delta = 1$, the method in [15] is recovered from (3.1) by setting Ψ_o to the identity function and $\delta = 1$.

Assumtion 3.1. **Nonlinearity Ψ:**
The non-linear function $\Psi : \mathbb{R} \to \mathbb{R}$ satisfies the following properties:
1. Function Ψ is odd, i.e., $\Psi(a) = -\Psi(-a)$, for any $a \in \mathbb{R}$;
2. $\Psi(a) > 0$, for any $a > 0$.
3. Function Ψ is a monotonically nondecreasing function;
4. Ψ is continuous, except possibly on a point set with Lebesque measure of zero. Moreover, Ψ is piecewise differentiable;
5. $|\Psi(a)| \le c_1$, for some constant $c_1 > 0$.
6. Ψ is either discontinuous at zero, or $\Psi(u)$ is strictly increasing for $u \in (-c_2, c_2)$, for some $c_2 > 0$.

As it will become clear ahead, the role of $\Psi_{\rm c}$ and $\Psi_{\rm o}$ is to lower the impact of the heavy-tailed noise that occurs in the regression model and in the communication between agents. As it is presented in [15], there are many nonlinear functions which satisfy Assumption 3.1. Now, we add more assumptions on the observation and communication noises through the following assumption.
At each time $t = 0, 1, ...$, a compact vector form of algorithm (3.1) is

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a} \mathbf{L}_{\mathbf{\Psi}_{\rm c}}(\mathbf{x}) - \mathbf{H}^\top \mathbf{\Psi}_{\rm o} \left(\mathbf{z}^t - \mathbf{H}\mathbf{x}^t \right) \right). \tag{3.2}$$

Here, $\mathbf{x}^t = [\mathbf{x}_1^t, \mathbf{x}_2^t, ..., \mathbf{x}_N^t]^\top \in \mathbb{R}^{MN}$, map $\mathbf{L}_{\mathbf{\Psi}_{\rm c}}(\mathbf{x}) : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is defined by

$$\mathbf{L}_{\mathbf{\Psi}_{\rm c}}(\mathbf{x}) = \begin{bmatrix} \vdots \\ \sum_{j \in \Omega_i} \mathbf{\Psi}_{\rm c}(\mathbf{x}_i - \mathbf{x}_j + \boldsymbol{\xi}_{ij}) \\ \vdots \end{bmatrix},$$

where, the blocks $\sum_{j \in \Omega_i} \mathbf{\Psi}_{\rm c}(\mathbf{x}_i - \mathbf{x}_j + \boldsymbol{\xi}_{ij}) \in \mathbb{R}^M$ are stacked one on top of another for $i = 1, ..., N$.

4. Theoretical results. In subsection 4.1 we express algorithm (3.2) in more general way, that will be used in the following subsections. Subsection 4.2 presents the statement and the proof of almost sure convergence of algorithm (3.1). In subsection 4.3 we state and prove asymptotic normality and calculate the corresponding asymptotic variance. Subection 4.4 presents and proves results on MSE rates.

4.1. Setting up analysis. In this subsection we rewrite algorithm (3.1) in the form suitable for stating the main results. To do that, firstly we define function $\varphi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(a) = \int \Psi(a + w) p(w) dw, \tag{4.1}$$

where $\Psi : \mathbb{R} \to \mathbb{R}$ is a nonlinear function that satisfies Assumption 3.1, and p is a probability density function that satisfies Assumptions 2.2 or 2.3.

Remark 4.1. The mapping φ has all key properties of function Ψ (see Lemma 6.2 in Appendix B in [34], see also [29]). Moreover, it has a strictly positive derivative at zero, i.e., $\varphi'(0) > 0$, which is necessary to prove our results. The facts that the nonlinearity Ψ is discontinuous at zero or that it has a positive derivative at zero, together with condition 4 from Assumptions 2.2 and condition 3 from 2.3, are crucial to ensure that φ has a positive derivative at zero (see Appendix B in [34], see also [15, 29]). Notice that the requirement that the pdf p is positive in the vicinity of the zero is not restrictive, since it holds true for a broad classes of non-zero noise pdfs.

Next, we define functions $\boldsymbol{\varphi}_{\mathrm{o}} : \mathbb{R}^N \to \mathbb{R}^N$, $\boldsymbol{\varphi}_{\mathrm{c}} : \mathbb{R}^M \to \mathbb{R}^M$ as $\boldsymbol{\varphi}_{\mathrm{o}}(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N) = [\varphi_{\mathrm{o}}(\mathbf{y}_1), \varphi_{\mathrm{o}}(\mathbf{y}_2), ..., \varphi_{\mathrm{o}}(\mathbf{y}_N)]$, $\boldsymbol{\varphi}_{\mathrm{c}}(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, ..., \hat{\mathbf{y}}_M) = [\varphi_{\mathrm{c}}(\hat{\mathbf{y}}_1), \varphi_{\mathrm{c}}(\hat{\mathbf{y}}_2), ..., \varphi_{\mathrm{c}}(\hat{\mathbf{y}}_M)]$, where $\mathbf{y} \in \mathbb{R}^N$, $\hat{\mathbf{y}} \in \mathbb{R}^M$ and functions φ_{o} and φ_{c} are transformations defined by (4.1) that correspond to Ψ_{o} and Ψ_{c}, respectively. For the a.s. convergence and asymptotic normality results, we will follow the stochastic approximation framework from [28, 18] (see Theorem 4 in Appendix A in [34]). That is, we represent algorithm (3.1) in the form suitable for stochastic approximation analysis. We start by substituting regression model (2.2) into algorithm (3.2), we get

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a} \mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x}) - \mathbf{H}^\top \boldsymbol{\Psi}_{\mathrm{o}} \left(\mathbf{H} \left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) + \mathbf{n}^t - \mathbf{H}\mathbf{x}^t \right) \right). \tag{4.2}$$

Define $\boldsymbol{\zeta}^t \in \mathbb{R}^N$ and $\boldsymbol{\eta}^t \in \mathbb{R}^{MN}$ by

$$\boldsymbol{\zeta}^t = \boldsymbol{\Psi}_{\mathrm{o}}(\mathbf{H} \left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) + \mathbf{n}^t - \mathbf{H}\mathbf{x}^t) - \boldsymbol{\varphi}_{\mathrm{o}} \left(\mathbf{H} \left(\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) - \mathbf{x}^t \right) \right), \quad \boldsymbol{\eta}^t = \begin{bmatrix} \vdots \\ \sum_{j \in \Omega_i} \boldsymbol{\eta}_{ij}^t \\ \vdots \end{bmatrix}, \tag{4.3}$$

where $\boldsymbol{\eta}_{ij}^t = \boldsymbol{\Psi}_{\mathrm{c}}(\mathbf{x}_i^t - \mathbf{x}_j^t + \boldsymbol{\xi}_{ij}^t) - \boldsymbol{\varphi}_{\mathrm{c}}(\mathbf{x}_i^t - \mathbf{x}_j^t)$. Now, since φ is defined by (4.1), it can be shown that $\mathbb{E}[\boldsymbol{\zeta}^t] = \mathbb{E}[\boldsymbol{\eta}^t] = 0$, where the expectation is taken with respect to $\mathcal{F}$ (see Appendix B in [34]). Furthermore, we define function $\mathbf{L}_{\boldsymbol{\varphi}_{\mathrm{c}}} : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ as $\mathbf{L}_{\boldsymbol{\varphi}_{\mathrm{c}}}(\cdot) = \mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\cdot) - \boldsymbol{\eta}^t$, i.e., its i-th block of size M is $\sum_{j \in \Omega_i} \boldsymbol{\varphi}_{\mathrm{c}}(\mathbf{x}_i - \mathbf{x}_j)$. for $i = 1, 2, ..., N$.

Finally, substituting (4.3) into (4.2), we rewrite algorithm (3.2) by

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a} \mathbf{L}_{\boldsymbol{\varphi}_{\mathrm{c}}}(\mathbf{x}^t) - \mathbf{H}^\top \boldsymbol{\varphi}_{\mathrm{o}} \left(\mathbf{H} \left(\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) - \mathbf{x}^t \right) \right) - \mathbf{H}^\top \boldsymbol{\zeta}^t + \frac{b}{a} \boldsymbol{\eta}^t \right). \tag{4.4}$$

Now, we are ready to establish following results.

4.2. Almost sure convergence.

We have the following Theorem.

THEOREM 4.2 (Almost sure convergence). *Let Assumptions 2.1-3.1 hold and* $\alpha_t = a/(t+1)^\delta$, $\delta \in (0.5, 1]$. *Then, for each agent* $i = 1, ..., N$, *the sequence of iterates* $\{\mathbf{x}_i^t\}$ *generated by algorithm* (3.1) *converges almost surely to the true vector parameter* $\boldsymbol{\theta}^*$.

Theorem 4.2 establishes almost sure convergence of the proposed algorithm (3.1), whether observation or communication noises have finite or infinite moments of order greater then one. On the other hand, if we set at least one of the functions $\Psi_{\mathrm{o}}, \Psi_{\mathrm{c}}$ to be identity functions (and thus recover either the $\mathcal{LU}$ scheme from [18] or the method from [15]), the resulting method fails to converge (See Appendix D in [34]). In other words, the methods in [18] and [15] fail to converge under the simultaneous presence of heavy-tailed observation and communication noises.

Proof. (Proof of Theorem 4.2)
The proof consists of verifying conditions B1–B5 of Theorem 4 in [34] (See Appendix A in [34]). First, we define quantities $\mathbf{r}(\mathbf{x})$ and $\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)$ by:

$$\mathbf{r}(\mathbf{x}) = -\frac{b}{a} \mathbf{L}_{\boldsymbol{\varphi}_{\mathrm{c}}}(\mathbf{x}) - \mathbf{H}^\top \boldsymbol{\varphi}_{\mathrm{o}} \left(\mathbf{H} \left(\mathbf{x} - \left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) \right) \right), \tag{4.5}$$

$$\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega) = -\frac{b}{a} \boldsymbol{\eta}^t + \mathbf{H}^\top \boldsymbol{\zeta}^t. \tag{4.6}$$

Here, ω denotes a canonical element of the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Condition B1 holds because $\mathbf{r}(\cdot)$ is $\mathcal{B}^{MN}$ measurable and $\boldsymbol{\gamma}(t+1, \cdot, \cdot)$ is $\mathcal{B}^{MN} \otimes \mathcal{F}$ measurable for each t, where $\mathcal{B}^{MN}$ is the Borel sigma algebra on $\mathbb{R}^{MN}$. Consider the filtration $\mathcal{F}_t$, $t = 1, 2, ...$, where $\mathcal{F}_t$ is the σ- algebra generated by $\{\mathbf{n}^s\}_{s=0}^{t-1}$ and

$\{\boldsymbol{\xi}_{ij}^s\}_{s=0}^{t-1}$. We have that the family of random vectors $\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)$ is $\mathcal{F}_t$ measurable, zero-mean and independent of $\mathcal{F}_{t-1}$. Hence, condition B2 holds.

We now show that condition B3 also holds. We use the following Lyapunov function $V:\mathbb{R}^{MN}\to\mathbb{R}$,

$$V(\mathbf{x}) = ||\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*||^2, \tag{4.7}$$

which is clearly twice continuously differentiable and has uniformly bounded second order partial derivatives. The gradient of V equals $\nabla V(\mathbf{x}) = 2\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)$. We must show that

$$\sup_{\mathbf{x}\in S_\epsilon}\langle \mathbf{r}(\mathbf{x}),\nabla V(\mathbf{x})\rangle < 0, \tag{4.8}$$

where $S_\epsilon = \{\mathbf{x}\in\mathbb{R}^{MN}:\ \|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|\in(\epsilon,1/\epsilon)\}$. For any $\mathbf{x}\in\mathbb{R}^{MN}$, we have:

$$\begin{aligned}\langle \mathbf{r}(\mathbf{x}),\nabla V(\mathbf{x})\rangle &= 2\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top\left(-\frac{b}{a}\mathbf{L}_{\boldsymbol{\varphi}_{\rm c}}(\mathbf{x}) - \mathbf{H}^\top\boldsymbol{\varphi}_{\rm o}\left(\mathbf{H}\left(\mathbf{x}^t-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)\right)\\ &= -\frac{2b}{a}\underbrace{\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top\mathbf{L}_{\boldsymbol{\varphi}_{\rm c}}(\mathbf{x})}_{T_1(\mathbf{x})} - \underbrace{\left(\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)^\top\boldsymbol{\varphi}_{\rm o}\left(\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)}_{T_2(\mathbf{x})}.\end{aligned} \tag{4.9}$$

The terms $T_1(\mathbf{x})$ and $T_2(\mathbf{x})$ can be written respectively as

$$T_1(\mathbf{x}) = \sum_{\{i,j\}\in E,\, i<j}\left(\mathbf{x}_i-\mathbf{x}_j\right)^\top\boldsymbol{\varphi}_{\rm c}\left(\mathbf{x}_i-\mathbf{x}_j\right) = \sum_{\{i,j\}\in E,\, i<j}\mathbf{g}\top\boldsymbol{\varphi}_{\rm c}\left(\mathbf{g}\right),$$

$$T_2(\mathbf{x}) = \sum_{i=1}^N \hat{\mathbf{g}}_i\,\varphi_{\rm o}(\hat{\mathbf{g}}_i),$$

where $\hat{\mathbf{g}} = \mathbf{H}^\top\boldsymbol{\varphi}_{\rm o}\left(\mathbf{H}\left(\mathbf{x}^t-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)$, $\mathbf{g}=\mathbf{x}_i-\mathbf{x}_j$ and $\mathbf{g}\top\boldsymbol{\varphi}_{\rm c}\left(\mathbf{g}\right) = \sum_{\ell=1}^M \mathbf{g}_\ell\varphi_{\rm c}\left(\mathbf{g}_\ell\right)$. Using the fact that both of the functions $\varphi_{\rm c}$ and $\varphi_{\rm o}$ are odd functions, for which we have that $\varphi(a)>0$ if $a>0$, we have that $\langle \mathbf{r}(\mathbf{x}),\nabla V(\mathbf{x})\rangle \geq 0$ for all $\mathbf{x}\in\mathbb{R}^{MN}$ (see Appendix B in [34]). Moreover, recalling the fact that function $\varphi_{\rm c}$ is continuous at zero, and equal to zero only at zero, we have that $T_1(\mathbf{x})$ is equal to zero if and only if $\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^* = \mathbf{1}_N\otimes\mathbf{m}$, for $\mathbf{m}\in\mathbb{R}^M$ (see Lemma 6 in Appendix B in [34]). We only consider the case when $\mathbf{m}\neq 0$, since from $\mathbf{m}=0$ we have that $\mathbf{x}=\mathbf{1}_N\otimes\boldsymbol{\theta}^*$, which is not in the set S_ϵ. However, for that choice of $\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*$ we have that

$$\begin{aligned}T_2(\mathbf{1}_N\otimes\boldsymbol{\theta}^*+\mathbf{1}_N\otimes\mathbf{m}) &= \left(\mathbf{H}\,\mathbf{1}_N\otimes\mathbf{m}\right)^\top\boldsymbol{\varphi}_{\rm o}\left(\mathbf{H}\,\mathbf{1}_N\otimes\mathbf{m}\right)\\ &= \sum_{i=1}^N\left(\mathbf{h}_i^\top\mathbf{m}\right)\varphi_{\rm o}\left(\mathbf{h}_i^\top\mathbf{m}\right) > 0,\end{aligned}$$

since $\mathbf{h}_i^\top\mathbf{m}$ and $\varphi_{\rm o}\left(\mathbf{h}_i^\top\mathbf{m}\right)$ have the same sign. Hence, for all $\epsilon>0$ we have that $\sup_{\mathbf{x}\in S_\epsilon}\langle \mathbf{r}(\mathbf{x}),\nabla V(\mathbf{x})\rangle < 0$. Thus, condition B3 also holds.

Now we inspect condition B4. From equation (4.5) we have that

$$\|\mathbf{r}(\mathbf{x})\|^2 \leq \left\|\frac{b}{a}\mathbf{L}_{\boldsymbol{\varphi}_{\rm c}}(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right\|^2 + \left\|\mathbf{H}^\top\boldsymbol{\varphi}_{\rm o}\left(\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)\right\|^2 \leq c_1(1+V(\mathbf{x})), \tag{4.10}$$

for some positive constant c_1 (see Appendix B in [34]). Moreover, we have that

$$\|\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\|^2 \leq \left\|\frac{b}{a}\boldsymbol{\eta}^t\right\|^2 + \left\|\mathbf{H}^\top\boldsymbol{\zeta}^t\right\|^2 \tag{4.11}$$

which leads to

$$\mathbb{E}\left[\left\|\boldsymbol{\gamma}(t+1,\mathbf{x}^t,\omega)\right\|^2\right] \leq c_2(1+V(\mathbf{x})), \tag{4.12}$$

for some positive constant c_2. Finally, we have that

$$\|\mathbf{r}(\mathbf{x})\|^2 + \mathbb{E}\left[\left\|\boldsymbol{\gamma}(t+1,\mathbf{x}^t,\omega)\right\|^2\right] \leq c_3(1+V(\mathbf{x})),$$

for some positive constant c_3. Setting that $\epsilon \to 0^+$ in (4.8), for all $\mathbf{x} \in \mathbb{R}^{MN}$, we have that $\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle \leq 0$. Thus,

$$\|\mathbf{r}(\mathbf{x})\|^2 + \mathbb{E}\left[\left\|\boldsymbol{\gamma}(t+1,\mathbf{x}^t,\omega)\right\|^2\right] \leq c_3(1+V(\mathbf{x})) - k\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle$$

for every $k > 0$. Therefore, condition B4 also holds. Condition B5 holds by the definition of the algorithm (3.1). Thus, almost sure convergence is proved. ∎

4.3. Asymptotic normality. We now consider asymptotic normality of the proposed estimator (3.1). We have the following theorem.

THEOREM 4.3 (Asymptotic normality). *Let Assumptions 2.1-3.1 hold. Consider algorithm (3.1) with step-size $\alpha_t = a/(t+1)^\delta$, $t = 0, 1, ...$, $a > 0$, with $\delta = 1$. Then, the normalized sequence of iterates $\{\sqrt{t+1}(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*)\}$ converges in distribution to a zero-mean multivariate normal random vector, i.e., the following holds:*

$$\sqrt{t+1}(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{S}),$$

where the asymptotic covariance matrix $\mathbf{S}$ equals:

$$\mathbf{S} = a^2 \int_0^\infty e^{\boldsymbol{\Sigma} v} \mathbf{S}_0 e^{\boldsymbol{\Sigma}^\top v} dv. \tag{4.13}$$

Here, $\mathbf{S}_0 = \frac{b^2}{a^2}\sigma_{\mathrm{c}}^2 \,\mathrm{Diag}\left(\{d_i\, \mathbf{I}_M\}\right) - \frac{b}{a}\mathbf{K}_{\mathrm{c,o}}\mathbf{H} - \frac{b}{a}\mathbf{H}^\top\mathbf{K}_{\mathrm{c,o}}^\top + \sigma_{\mathrm{o}}^2\mathbf{H}^\top\mathbf{H}$; $\sigma_{\mathrm{o}}^2 = \int |\Psi_{\mathrm{o}}(w)|^2 d\Phi_{\mathrm{o}}(w)$ is the effective observation noise variance after passing through the nonlinearity Ψ_{o}; $\sigma_{\mathrm{c}}^2 = \int |\Psi_{\mathrm{c}}(w)|^2 d\Phi_{\mathrm{c}}(w)$ is the effective communication noise variance after passing through the nonlinearity Ψ_{c}; $\mathbf{K}_{\mathrm{c,o}} \in \mathbf{R}^{MN\times N}$ is the effective cross-covariance matrix between the observation and the communication noise after passing through the appropriate nonlinearity, i.e., the (k,s) element of the matrix $\mathbf{K}_{\mathrm{c,o}}$ is given by $[(\mathbf{K}_{\mathrm{c,o}})]_{ks} = \sum_{j\in\Omega_i} \int\int \Psi_{\mathrm{c}}(w_{ij\ell})\Psi_{\mathrm{o}}(w_k) p_{k,ij\ell}^{\mathrm{c,o}}(w_{ij\ell}, w_k) dw_{ij\ell} dw_k$. Here, ℓ satisfies the following: $s = M(i-1)+\ell$; and $p_{k,ij\ell}^{\mathrm{c,o}}$ is the joint probability density function for the k-th observation noise n_k and the ℓ-th element of the communication noise $[(\boldsymbol{\xi}_{ij})]_\ell$. We also recall the observation matrix $\mathbf{H}$ in (2.2); functions φ_{c}, φ_{o} appropriate versions of function φ in (4.1); and $\Sigma = \frac{1}{2}\mathbf{I} - a(\frac{b}{a}\varphi_{\mathrm{c}}'(0)\mathbf{L}\otimes\mathbf{I}_M + \varphi_{\mathrm{o}}'(0)\mathbf{H}^\top\mathbf{H})$; here, a is taken large enough such that matrix $\boldsymbol{\Sigma}$ is stable.

Remark 4.4. Notice that, for the assumed setting, σ_{c}^2 and σ_{o}^2 are finite. Also, $\mathbf{K}_{\mathrm{c,o}}$ is finite, i.e., $\|\mathbf{K}_{\mathrm{c,o}}\| < \infty$, since we have that

$$\left|\int\int \Psi_{\mathrm{c}}(w_1)\Psi_{\mathrm{o}}(w_2) d\Phi^{\mathrm{c,o}}\right| \leq \int\int |\Psi_{\mathrm{c}}(w_1)\Psi_{\mathrm{o}}(w_2)| d\Phi^{\mathrm{c,o}} < \frac{1}{2}\sigma_{\mathrm{c}}^2 + \frac{1}{2}\sigma_{\mathrm{o}}^2.$$

Remark 4.5. If we assume that observation and communication noise are mutually independent, the only difference from the previous theoretical results occurs in the $\mathbf{A}(t,\mathbf{x})$, i.e., in the $\mathbf{S}_0$. Under this setting, matrix $\mathbf{S}_0$ is now equal to

$$\mathbf{S}_0 = \frac{b^2}{a^2}\sigma_{\mathrm{c}}^2 \,\mathrm{Diag}\left(\{d_i\, \mathbf{I}_M\}\right) + \sigma_{\mathrm{o}}^2\mathbf{H}^\top\mathbf{H},$$

which is expected, since the effective cross-covariance matrix $\mathbf{K}_{\mathrm{c,o}}$ is now equal to zero.

Theorem 4.3 establishes asymptotic normality of the proposed method. This is achieved with heavy-tailed observation and communication noise an the nonlinearities Ψ_{o} and Ψ_{c} with uniformly bounded outputs. Moreover, the theorem explicitly evaluates the corresponding asymptotic variance. When the two noises are mutually independent, Ψ_{o} is identity, and observation noise variance is finite, we recover the result in [15], Theorem 3.5, as a special case. That is, a notable difference with respect to [15] is the ability to handle here mutually correlated observation and communication noises. The effect of correlation is complex in general, however, as shown in Section 5 later, generally a stronger positive noises correlation leads to a lower asymptotic variance. Intuitively, at an extreme, a full positive correlation practically means that only one effective noise exists in the system, and hence it can be suppressed more easily. Further, note that Theorem 4.3 establishes a local asymptotic rate $O(1/t)$ of $\mathbf{x}^t$ to zero, in the weak convergence sense, when $\alpha_t = a/(t+1)$. We show later (see Theorem 4.6) that a global MSE rate $O(1/t^{\hat{\delta}})$ with a lower (worse) degree $\hat{\delta}$ can be established when step-size $\alpha_t = a/(t+1)^{\delta}$, $\delta \in (0.5, 1)$, is used.

We next discuss asymptotic efficiency[2] of the proposed estimator. We first briefly review the relevant existing work to better position our results. First, consider the best linear centralized estimator $\mathbf{x}^t_{\text{cent}}$ of $\boldsymbol{\theta}^\star$, that has access to measurements from all sensors (nodes) $n = 1, 2, ..., N$ at all times $t = 0, 1, ...$ In the general case, additionally assuming that observation noise has finite variance, the best linear centralized estimator $\mathbf{x}^t_{\text{cent}}$ is asymptotically normal and has the lowest asymptotic covariance matrix $\mathbf{S}_{\text{cent}}$ among all estimators of $\boldsymbol{\theta}^\star$ when the only knowledge of observation noise is variance and no other information of noise distribution is known. Moreover, its asymptotic covariance matrix $\mathbf{S}_{\text{cent}}$ attains the Cramér-Rao lower bound if the observation noise is Gaussian (see for example [17]). On the other hand, when the probability density function is known, the centralized estimator in [29] can be tuned to the pdf of the observation noise so that it achieves the Cramér-Rao bound. In the distributed setting, when there is no communication noise, the authors of [17] develop an estimator which is asymptotically normal and has the optimal asymptotic covariance matrix $\mathbf{S}_{\text{cent}}$ (optimal in the sense that the asymptotic covariance matrix is the same as for the best linear centralized estimator $\mathbf{x}^t_{\text{cent}}$). We now discuss the asymptotic covariance matrix $\mathbf{S}$ of the proposed estimator (3.1). This quantity depends on the system parameters, including network topology and communication noise. Therefore, in the general case, the proposed estimator (3.1) is not asymptotically efficient, i.e., $\mathbf{S} \neq \mathbf{I}^{-1}(\boldsymbol{\theta}^\star)$, where $\mathbf{I}(\boldsymbol{\theta}^\star)$ is the Fisher information matrix. However, with respect to the proposed distributed recursive estimator, we make the following observations. 1) First, the estimator is order-optimal in the weak convergence sense; that is, its (weak convergence sense) rate of error decay is the same as that of the asymptotically efficient estimator. 2) The corresponding "convergence constant," i.e., the asymptotic covariance, is different from that of the centralized Cramér-Rao-optimal estimator, and it is hence not optimal. We note that the paper provides major contributions with respect to state of the art, as it gives the first distributed estimator that ensures almost sure convergence in the presence of infinite variance correlated sensing and communication noises; moreover, its weak convergence sense rate of convergence is order-optimal. It remains an interesting future work direction to explore whether an optimal asymptotic covariance can be achieved in this setting via distributed estimators. In view of the results [29] for the centralized setting, it is likely that this cannot be achieved unless the nonlinearities are tuned to the noise pdfs that in turn have to be known.

Proof. (Proof of Theorem 4.3)

[2]An estimator $\mathbf{y}^t$ of an unknown parameter $\boldsymbol{\theta}^\star$, for which we have that $\sqrt{t+1}(\mathbf{y}^t - \boldsymbol{\theta}^*) \Rightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, is said to be asymptotically efficient if $\mathbf{S} = \mathbf{I}^{-1}(\boldsymbol{\theta}^\star)$, where $\mathbf{I}(\boldsymbol{\theta}^\star)$ is the Fisher information matrix. The Fisher information matrix represents the best achievable asymptotic covariance by any estimator, as determined by the well-known Cramer-Rao bound (see [28]).

We prove Theorem 4.3 in the same manner as Theorem 4.2 is proved, i.e., by verifying assumptions C1-C5 of Theorem 4 in [34] (see Appendix A in [34]). Function $\mathbf{r}(\cdot)$ defined by (4.5) can be written as

$$\mathbf{r}(\mathbf{x}) = -\frac{b}{a}\varphi_{\mathrm{c}}'(0)\mathbf{L}\otimes\mathbf{I}_M\ (\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*) - \varphi_{\mathrm{o}}'(0)\mathbf{H}^\top\mathbf{H}\,(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*) + \boldsymbol{\delta}(\mathbf{x}),$$

Here, mapping $\boldsymbol{\delta}:\mathbb{R}^{MN}\to\mathbb{R}^{MN}$ is given by:

$$\boldsymbol{\delta}(\mathbf{x}) = -\frac{b}{a}\mathbf{L}_{\boldsymbol{\delta}_{\mathrm{c}}}(\mathbf{x}) - \mathbf{H}^\top\boldsymbol{\delta}_{\mathrm{o}}\left(\mathbf{H}\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right). \tag{4.14}$$

Next, mapping $\mathbf{L}_{\boldsymbol{\delta}_{\mathrm{c}}}(\mathbf{x}):\mathbb{R}^{MN}\to\mathbb{R}^{MN}$ is vector of size MN such that the i-th M-size block equals $\sum_{j\in\Omega_i}\boldsymbol{\delta}_{\mathrm{c}}(\mathbf{x}_i-\mathbf{x}_j)$, $i=1,2,...,N$, mappings $\boldsymbol{\delta}_{\mathrm{c}}:\mathbb{R}^M\to\mathbb{R}^M$, $\boldsymbol{\delta}_{\mathrm{o}}:\mathbb{R}^N\to\mathbb{R}^N$ are component-wise maps of δ_{c} and δ_{o} are first order residuals that corresponds to φ_{c} and φ_{o} respectively, i.e., $\boldsymbol{\delta}_{\mathrm{c}}(\mathbf{y}_1,\mathbf{y}_1,...,\mathbf{y}_M)=[\delta_{\mathrm{c}}(\mathbf{y}_1),\delta_{\mathrm{c}}(\mathbf{y}_2),...,\delta_{\mathrm{c}}(\mathbf{y}_M)]^\top$ and $\boldsymbol{\delta}_{\mathrm{o}}(\hat{\mathbf{y}}_1,\hat{\mathbf{y}}_1,...,\hat{\mathbf{y}}_N)=[\delta_{\mathrm{o}}(\hat{\mathbf{y}}_1),\delta_{\mathrm{o}}(\hat{\mathbf{y}}_2),...,\delta_{\mathrm{o}}(\hat{\mathbf{y}}_M)]^\top$ for $\mathbf{y}\in\mathbb{R}^N,\hat{\mathbf{y}}\in\mathbb{R}^M$ (see Appendix B in [34]).
Thus, $\mathbf{r}(\mathbf{x})$ admits representation in (36) of Theorem 4 in [34] for $\mathbf{B}=-\frac{b}{a}\varphi_{\mathrm{c}}'(0)\mathbf{L}\otimes\mathbf{I}_M-\varphi_{\mathrm{o}}'(0)\mathbf{H}^\top\mathbf{H}$ and mapping $\boldsymbol{\delta}(\cdot)$ defined by (4.14). Therefore, condition C1 holds. Since we use that $\alpha_t=\frac{a}{t+1}$, condition C2 trivially holds. Furthermore, $\boldsymbol{\Sigma}=a\mathbf{B}+\frac{1}{2}\mathbf{I}$ is stable if a is large enough, because matrix $-\mathbf{B}$ is positive definite (See [18]). Thus, condition C3 also holds.
For $\mathbf{A}(t,\mathbf{x})=\mathbb{E}\big[\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)\boldsymbol{\gamma}^\top(t+1,\mathbf{x},\omega)\big]$ it is easy to show that

$$\lim_{t\to\infty,\mathbf{x}\to\boldsymbol{\theta}^*}\mathbf{A}(t,\mathbf{x}) = \frac{b^2}{a^2}\sigma_{\mathrm{c}}^2\,\mathrm{Diag}\left(\{d_i\,\mathbf{I}_M\}\right) - \frac{b}{a}\mathbf{K}_{\mathrm{c,o}}\mathbf{H} - \frac{b}{a}\mathbf{H}^\top\mathbf{K}_{\mathrm{c,o}}^\top + \sigma_{\mathrm{o}}^2\mathbf{H}^\top\mathbf{H}.$$

Therefore, condition C4 also holds. To show that condition C5 holds, it is suffice to show that the family of random variables $\{\|\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1,\mathbf{x},\omega)\|^2\}_{t=0,1,...,\,\|\mathbf{x}-\boldsymbol{\theta}^*\|<\epsilon}$ is uniformly integrable. To do that, follow the arguments as in e.g., [18] and [15]. ∎

4.4. Mean squared error convergence.

In this subsection, we state and prove a result on the mean squared error (MSE) convergence rate when both nonlinearities Ψ_{o} and Ψ_{c} satisfy part 5' of Assumption 3.1, i.e., $|\Psi_{\mathrm{o}}|\le c_{\mathrm{o}}$, $|\Psi_{\mathrm{c}}|\le c_{\mathrm{c}}$, for some positive constants c_{o} and c_{c}. Moreover, we set the step size to $\alpha_t=\frac{a}{(t+1)^\delta}$, for $\delta\in(\frac{1}{2},1)$. We have the following theorem.

THEOREM 4.6 (MSE convergence). *Let Assumptions 2.1-3.1 hold. Then, for the sequence of iterates $\{\mathbf{x}^t\}$ generated by algorithm (3.2), provided that the step-size sequence $\{\alpha_t\}$ is given by $\alpha_t=a/(t+1)^\delta$, $a>0,\delta\in(0.5,1)$, there exists $\hat{\delta}\in(0,1)$ such that $\mathbb{E}[\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2]=O(1/t^{\hat{\delta}})$.*

Theorem 4.6 establishes a MSE convergence rate of the proposed estimator (3.2) under the simultaneous presence of heavy-tailed (possibly infinite variance) observation and communication noises, when both the observation and communication nonlinearities have uniformly bounded outputs. This is in contrast with recent studies on distributed estimation in heavy-tailed noises like [15] that only establishes a.s. and asymptotic normality results. We refer to the proof of Theorem 4.6 for the exact value of the convergence rate power $\hat{\delta}$.

Setting up the proof. We now prove Theorem 4.6 through a sequence of intermediate results (Lemmas). Recall quantities $\mathbf{r}(\cdot)$, $\boldsymbol{\gamma}(\cdot,\cdot,\cdot)$ and $V(\cdot)$ from (4.5), (4.6) and (4.7) respectively. The proof will be based on establishing a sufficient decay on quantity $\mathbb{E}[V(\mathbf{x}^t)]$. First, notice that algorithm (4.4) can be written as

$$\mathbf{x}^{t+1}=\mathbf{x}^t+\alpha_t\left(\mathbf{r}(\mathbf{x}^t)+\boldsymbol{\gamma}(t+1,\mathbf{x}^t,\omega)\right).$$

Moreover, we have that

$$\begin{aligned} V(\mathbf{x}^{t+1}) &= V(\mathbf{x}^t) + 2\alpha_t \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)^\top \left(\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)\right) \\ &\quad + \alpha_t^2 \|\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)\|^2 \\ &= V(\mathbf{x}^t) + 2\alpha_t \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)^\top \left(\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)\right) + \alpha_t^2\, c', \end{aligned}$$

for positive constant $c' = \|\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)\|^2 < \infty$. Therefore, taking a conditional expectation with respect to $\mathcal{F}_t$, we have:

$$\mathbb{E}[V(\mathbf{x}^{t+1}) | \mathcal{F}_t] = V(\mathbf{x}^t) + 2\alpha_t \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)^\top \mathbf{r}(\mathbf{x}^t) + \alpha_t^2\, c'. \tag{4.15}$$

Also, from equation (4.9), it follows that

$$\left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)^\top \mathbf{r}(\mathbf{x}^t) = -\frac{b}{a} T_1(\mathbf{x}^t) - T_2(\mathbf{x}^t). \tag{4.16}$$

We next need to show that the quantity in (4.16) is "sufficiently negative", relative to quantity $V(\mathbf{x}^t)$. This is achieved through a sequence of lemmas. First, we upper bound quantities $\|\mathbf{x}^t\|$ and $\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\|$.

Lemma 4.7. *Let Assumptions 2.1-3.1 hold. Then, for the sequence of iterates $\{\mathbf{x}^t\}$ generated by algorithm (3.2), provided that the step-size sequence $\{\alpha_t\}$ is given by $\alpha_t = a/(t+1)^\delta$, $a > 0, \delta \in (0.5, 1)$, we have that, for any outcome ω:*

$$\|\mathbf{x}^t\| \le g_t = \|\mathbf{x}^0\| + \left(b\sqrt{MN} d\, c_{\mathrm{c}} + a\, \|\mathbf{H}\| \sqrt{N} c_{\mathrm{o}}\right) \frac{t^{1-\delta}}{1-\delta}, \tag{4.17}$$

$$\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\| \le g_t' = \|\mathbf{x}^0 - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\| + \left(b\sqrt{MN} d\, c_{\mathrm{c}} + a\, \|\mathbf{H}\| \sqrt{N} c_{\mathrm{o}}\right) \frac{t^{1-\delta}}{1-\delta}. \tag{4.18}$$

Consequently, $\|\mathbf{H}\left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\| \le \|\mathbf{H}\|\, g_t'$.

Proof. Using the boundness of the nonlinearities, we have that $\|\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x})\|^2 \le \sqrt{MN} d\, c_{\mathrm{c}}$ and $\|\mathbf{H}^\top \boldsymbol{\Psi}_{\mathrm{o}}\left(\mathbf{H}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* - \mathbf{x}^t\right) + \mathbf{n}^t\right)\| \le \|\mathbf{H}\| \sqrt{N} c_{\mathrm{o}}$, where $d = \max_i d_i$. Therefore, recalling the algorithm (4.2), for all $t > 0$ we have that

$$\begin{aligned} \|\mathbf{x}^t\| &\le \|\mathbf{x}^{t-1}\| + \alpha_{t-1} \underbrace{\left(\frac{b}{a}\sqrt{MN} d\, c_{\mathrm{c}} + \|\mathbf{H}\| \sqrt{N} c_{\mathrm{o}}\right)}_{c} \le \|\mathbf{x}^{t-2}\| + \alpha_{t-2}\, c + \alpha_{t-1}\, c \\ &\le \|\mathbf{x}^0\| + c \sum_{j=0}^{t-1} \frac{a}{(1+j)^\delta} \le \|\mathbf{x}^0\| + c \int_0^{t-1} \frac{a}{(1+s)^\delta} ds \le \|\mathbf{x}^0\| + c\, a\, \frac{t^{1-\delta}}{1-\delta}. \end{aligned}$$

Analogously, for all $t > 0$, we have that $\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\| \le g_t'$, and as a consequence $\|\mathbf{H}\left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)\| \le \|\mathbf{H}\|\, g_t'$. ∎

Next, we have the following Lemma that bounds quantities $T_1(x)$ and $T_2(x)$.

Lemma 4.8. *Let Assumptions 2.1-3.1 hold. Then, for the sequence of iterates $\{\mathbf{x}^t\}$ generated by algorithm (3.2), provided that the step-size sequence $\{\alpha_t\}$ is given by $\alpha_t = a/(t+1)^\delta$, $a > 0, \delta \in (0.5, 1)$, we have that there exist positive constants G_c and G_0 such that, for any outcome ω:*

$$\begin{aligned} T_1(\mathbf{x}^t) &\ge \frac{\varphi_c'(0) G_{\mathrm{c}}}{4 g_t} \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)^\top \mathbf{L} \otimes \mathbf{I} \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right), \\ T_2(\mathbf{x}^t) &\ge \frac{\varphi_o'(0) G_o}{2\|\mathbf{H}\| g_t'} \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)^\top \mathbf{H}^\top \mathbf{H} \left(\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right), \end{aligned}$$

To prove Lemma 4.8, we make use of the following Lemma from [13] (see Lemma 5.5 in [13]).

Lemma 4.9. *Consider function φ in (4.1), there exists a positive constant G such* *that* $|\varphi(a)| \leq \frac{\varphi'(0)\, G|a|}{2\, g}$, *for all* $|a| < g$.

Proof. (Proof of Theorem 4.8) Using Lemma 4.9 for function φ_c we get that there exists a positive constant $G_{\rm c}$ such that

$$
\begin{aligned}
T_1(\mathbf{x}^t) &= \sum_{\{i,j\}\in E,\, i<j} \left(\mathbf{x}_i^t - \mathbf{x}_j^t\right)^\top \boldsymbol{\varphi}_{\rm c}\left(\mathbf{x}_i^t - \mathbf{x}_j^t\right) \\
&= \sum_{\{i,j\}\in E,\, i<j} \sum_{\ell=1}^{M} \left((\mathbf{x}_i^t)_\ell - (\mathbf{x}_j^t)_\ell\right)^\top \boldsymbol{\varphi}_{\rm c}\left((\mathbf{x}_i^t)_\ell - (\mathbf{x}_j^t)_\ell\right) \\
&\geq \frac{\varphi_c'(0)G_{\rm c}}{4g_t} \sum_{\{i,j\}\in E,\, i<j} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2 = \frac{\varphi_c'(0)G_{\rm c}}{4g_t}(\mathbf{x}^t)^\top(\mathbf{L}\otimes\mathbf{I})\mathbf{x}^t \qquad (4.19) \\
&= \frac{\varphi_c'(0)G_{\rm c}}{4g_t}\left(\mathbf{x}^t - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \mathbf{L}\otimes\mathbf{I}\left(\mathbf{x}^t - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right),
\end{aligned}
$$

since, from Lemma 4.7 we have $\|\mathbf{x}^t\| \leq g_t$. Analogously, from Lemma 4.9 we have that for the function φ_o there exists a positive constant G_o such that

$$
\begin{aligned}
T_2(\mathbf{x}) &= \sum_{i=1}^{N} \left(\mathbf{H}\left(\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right)_i \varphi_{\rm o}\left(\left(\mathbf{H}\left(\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\right)_i\right) \\
&\geq \frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\left(\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \mathbf{H}^\top\mathbf{H}\left(\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right), \qquad (4.20)
\end{aligned}
$$

since, from Lemma 4.7 we have $\|\mathbf{H}\left(\mathbf{x}^t - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\| \leq \|\mathbf{H}\|\, g_t'$. ∎

We next have the following theorem that analyzes positive definiteness of the matrix $\frac{\varphi_c'(0)G_{\rm c}}{4g_t}\mathbf{L}\otimes\mathbf{I} + \frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\mathbf{H}^\top\mathbf{H}$.

LEMMA 4.10. *Let Assumptions 2.1-3.1 hold. The following is true for any* $\mathbf{x} \in \mathbb{R}^{MN}$*:*

$$
\begin{aligned}
&\left(\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \left(\frac{\varphi_c'(0)G_{\rm c}}{4g_t}\mathbf{L}\otimes\mathbf{I} + \frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\mathbf{H}^\top\mathbf{H}\right)\left(\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right) \\
&\geq \min\left\{\frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\left(\frac{\lambda_{\rm H}}{N} - \frac{2S_{\rm H}}{\sqrt{N}}k\right), \frac{b\,\varphi_c'(0)G_{\rm c}}{4ag_t}\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\right\}\|\mathbf{x} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2,
\end{aligned}
$$

where g_t *and* g_t' *are defined in Lemma 4.7,* $G_{\rm c}$ *and* G_o *in Lemma 4.8,* $S_{\rm H} = \sum_{i=1}^{N}\|\mathbf{h}_i\|^2$, $\lambda_{\rm H} = \lambda_1\left(\sum_{i=1}^{N}\mathbf{h}_i\mathbf{h}_i^\top\right) > 0$ *is the smallest eigenvalue of regular matrix* $\sum_{i=1}^{N}\mathbf{h}_i\mathbf{h}_i^\top$ *(see Assumption 2.1) and recalling that* $\lambda_2(\mathbf{L}) > 0$ *is the smallest positive eigenvalue of Laplacian matrix* $\mathbf{L}$.

Proof. Let us consider matrix $\mathbf{L}\otimes\mathbf{I}+\mathbf{H}^\top\mathbf{H}$ and follow argument as in Appendix A of [32]. For any $\mathbf{x}\in\mathbb{R}^{MN}$, we have that there exist vectors $\mathbf{u}\in\operatorname{span}\{\mathbf{1}\otimes\mathbf{m}|\mathbf{m}\in\mathbb{R}^M\}$ and $\mathbf{v}\in\operatorname{span}\{\mathbf{1}\otimes\mathbf{m}|\mathbf{m}\in\mathbb{R}^M\}^\perp$ such that $\mathbf{x}=\mathbf{u}+\mathbf{v}$. Firstly, we have that

$$
\begin{aligned}
\left(\mathbf{u} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \mathbf{H}^\top\mathbf{H}\left(\mathbf{u} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\right) &= \sum_{i=1}^{N}(\hat{\mathbf{u}} - \boldsymbol{\theta}^\star)^\top\mathbf{h}_i\mathbf{h}_i^\top(\hat{\mathbf{u}} - \boldsymbol{\theta}^\star) \\
&= (\hat{\mathbf{u}} - \boldsymbol{\theta}^\star)^\top\left(\sum_{i=1}^{N}\mathbf{h}_i\mathbf{h}_i^\top\right)(\hat{\mathbf{u}} - \boldsymbol{\theta}^\star) \\
&\geq \lambda_{\rm H}\|\hat{\mathbf{u}} - \boldsymbol{\theta}^\star\|^2,
\end{aligned}
$$

where $\hat{\mathbf{u}}\in\mathbb{R}^M$ such that $\mathbf{u}=\mathbf{1}\otimes\hat{\mathbf{u}}$. Notice here that $\|\mathbf{u} - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\| = \sqrt{N}\|\hat{\mathbf{u}} - \boldsymbol{\theta}^\star\|$. Secondly, $(\mathbf{x}-\mathbf{u})^\top\mathbf{H}^\top\mathbf{H}(\mathbf{x}-\mathbf{u}) \geq 0$, since $\mathbf{H}^\top\mathbf{H}$ is positive semi-definite matrix.

Thirdly, following also holds

$$
\begin{aligned}
\left(\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \mathbf{H}^\top\mathbf{H}\left(\mathbf{x}-\mathbf{u}\right) &= \sum_{i=1}^{N}(\hat{\mathbf{u}}-\boldsymbol{\theta}^\star)^\top\mathbf{h}_i\mathbf{h}_i^\top(\mathbf{x}_i-\hat{\mathbf{u}})\\
&\geq -\sum_{i=1}^{N}\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|\|\mathbf{h}_i\|^2\|\mathbf{x}_i-\hat{\mathbf{u}}\|\\
&\geq -\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|\|\mathbf{v}\|S_{\mathrm{H}}.
\end{aligned}
$$

Analogously, we have that $\left(\mathbf{x}-\mathbf{u}\right)^\top \mathbf{H}^\top\mathbf{H}\left(\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\geq -\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|\|\mathbf{v}\|S_{\mathrm{H}}$. Therefore,

$$
\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \mathbf{H}^\top\mathbf{H}\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\geq \lambda_{\mathrm{H}}\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|^2-2S_{\mathrm{H}}\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|\|\mathbf{v}\|.
$$

We also have that $\mathbf{u}-\mathbf{1}\otimes\boldsymbol{\theta}^\star\in\mathrm{null}(\mathbf{L}\otimes\mathbf{I})$ and $\mathbf{v}\in\mathrm{Range}(\mathbf{L}\otimes\mathbf{I})$ and, hence, we have that

$$
\begin{aligned}
\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top \mathbf{L}\otimes\mathbf{I}\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right) &= \left(\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*+\mathbf{v}\right)^\top\mathbf{L}\otimes\mathbf{I}\left(\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*+\mathbf{v}\right)\\
&= \mathbf{v}^\top\mathbf{L}\otimes\mathbf{I}\,\mathbf{v}\geq\lambda_2(\mathbf{L}\otimes\mathbf{I})\|\mathbf{v}\|^2=\lambda_2(\mathbf{L})\|\mathbf{v}\|^2.
\end{aligned}
$$

Let $k>0$ be arbitrarily chosen. If $\|\mathbf{v}\|\leq k\|\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|$, then we have that

$$
\begin{aligned}
&\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top\left(\mathbf{L}\otimes\mathbf{I}+\mathbf{H}^\top\mathbf{H}\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\\
&\quad\geq \lambda_{\mathrm{H}}\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|^2-2S_{\mathrm{H}}\|\hat{\mathbf{u}}-\boldsymbol{\theta}^\star\|\|\mathbf{v}\|+\lambda_2(\mathbf{L})\|\mathbf{v}\|^2\\
&\quad\geq\left(\frac{\lambda_{\mathrm{H}}}{N}-\frac{2S_{\mathrm{H}}}{\sqrt{N}}k\right)\|\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2+\lambda_2(\mathbf{L})\|\mathbf{v}\|^2\\
&\quad\geq\min\{\frac{\lambda_{\mathrm{H}}}{N}-\frac{2S_{\mathrm{H}}}{\sqrt{N}}k,\lambda_2(\mathbf{L})\}\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2,
\end{aligned}
$$

where in the last inequality we used the fact that $\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2=\|\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2+\|\mathbf{v}\|^2$. If $\|\mathbf{v}\|\geq k\|\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|$, then

$$
\begin{aligned}
\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top\left(\mathbf{L}\otimes\mathbf{I}+\mathbf{H}^\top\mathbf{H}\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)&\geq 0+\lambda_2(\mathbf{L})\|\mathbf{v}\|^2\\
&\geq\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\|\mathbf{v}\|^2+\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\|\mathbf{u}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2\\
&\geq\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2.
\end{aligned}
$$

Therefore, regardless of vector $\mathbf{v}$, we have that

$$
\begin{aligned}
&\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top\left(\mathbf{L}\otimes\mathbf{I}+\mathbf{H}^\top\mathbf{H}\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\\
&\quad\geq\min\left\{\frac{\lambda_{\mathrm{H}}}{N}-\frac{2S_{\mathrm{H}}}{\sqrt{N}}k,\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\right\}\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2.
\end{aligned}
$$

Following the same idea, we get that

$$
\begin{aligned}
&\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)^\top\left(\frac{\varphi_c'(0)G_{\mathrm{c}}}{4g_t}\mathbf{L}\otimes\mathbf{I}+\frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\mathbf{H}^\top\mathbf{H}\right)\left(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right)\\
&\quad\geq\min\left\{\frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\left(\frac{\lambda_{\mathrm{H}}}{N}-\frac{2S_{\mathrm{H}}}{\sqrt{N}}k\right),\frac{b\,\varphi_c'(0)G_{\mathrm{c}}}{4ag_t}\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\right\}\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2. \qquad (4.21)\quad\square
\end{aligned}
$$

Finally, to prove Theorem 4.6, we make use of the following Lemma from [13] (see Theorem 5.2 in [13]).

LEMMA 4.11. *Let z^t be a nonnegative (deterministic) sequence satisfying*

$$
z^{t+1}\leq(1-r_1^t)z^t+r_2^t,
$$

for all $t\geq t'$, for some $t'>0$, with some $z^{t'}\geq 0$. Here, $\{r_1^t\}$ and $\{r_2^t\}$ are deterministic sequences with $\frac{a_1}{t+1}\leq r_1^t\leq 1$ and $r_2^t\leq\frac{a_2}{(t+1)^\delta}$, with $a_1,a_2>0$, and $\delta>0$. Then, the *following holds: (1)* $z^t = O(\frac{1}{t^{\delta-1}})$ *provided that* $a_1 > \delta - 1$; *(2) if* $a_1 \leq \delta - 1$, *them* $z^t = O(\frac{1}{t^s})$, *for any* $s < a_1$.

We are finally ready to finalize the proof of Theorem 4.6.

Proof. (Proof of Theorem 4.6) From equations (4.21) and (4.16) we get that

$$(\mathbf{x}-\mathbf{1}_N \otimes \boldsymbol{\theta}^*)^\top \mathbf{r}(\mathbf{x}^t)$$
$$\leq -\min\left\{\frac{\varphi_o'(0)G_o}{2\|\mathbf{H}\|g_t'}\left(\frac{\lambda_{\mathrm{H}}}{N} - \frac{2S_{\mathrm{H}}}{\sqrt{N}}k\right), \frac{b\,\varphi_c'(0)G_{\mathrm{c}}}{4ag_t}\frac{\lambda_2(\mathbf{L})}{1+\frac{1}{k^2}}\right\}\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2.$$

Therefore, taking the expectation in (4.15), we have that

$$\mathbb{E}[V(\mathbf{x}^{\mathbf{t+1}})] \leq \left(1 - \frac{a_1}{t+1}\right)\mathbb{E}[V(\mathbf{x}^{\mathbf{t}})] + \frac{a_2}{(1+t)^{2\delta}},$$

where

$$a_1 = \min\left\{\frac{\varphi_o'(0)G_o a(1-\delta)\left(\lambda_{\mathrm{H}} - 2S_{\mathrm{H}}\sqrt{N}k\right)}{\|\mathbf{H}\|N\left(\|\mathbf{x}^0 - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\| + b\sqrt{MN}d\,c_{\mathrm{c}} + a\,\|\mathbf{H}\|\sqrt{N}c_{\mathrm{o}}\right)},\right.$$
$$\left.\frac{b\,\varphi_c'(0)G_{\mathrm{c}}(1-\delta)\lambda_2(\mathbf{L})k^2}{2(k^2+1)\left(\|\mathbf{x}^0\| + b\sqrt{MN}d\,c_{\mathrm{c}} + a\,\|\mathbf{H}\|\sqrt{N}c_{\mathrm{o}}\right)}\right\}$$

and $a_2 = a^2c'$. Therefore, using the Lemma 4.11, $\hat{\delta}$ is any positive number such that

$$\hat{\delta} < \min\left\{2\delta - 1, \frac{\varphi_o'(0)G_{\mathrm{o}} a(1-\delta)\left(\lambda_{\mathrm{H}} - 2S_{\mathrm{H}}\sqrt{N}k\right)}{\|\mathbf{H}\|N\left(\|\mathbf{x}^0 - \mathbf{1}_N\otimes\boldsymbol{\theta}^*\| + b\sqrt{MN}d\,c_{\mathrm{c}} + a\,\|\mathbf{H}\|\sqrt{N}c_{\mathrm{o}}\right)},\right.$$
$$\left.\frac{b\,\varphi_c'(0)G_{\mathrm{c}}(1-\delta)\lambda_2(\mathbf{L})k^2}{2(k^2+1)\left(\|\mathbf{x}^0\| + b\sqrt{MN}d\,c_{\mathrm{c}} + a\,\|\mathbf{H}\|\sqrt{N}c_{\mathrm{o}}\right)}\right\}.$$

Therefore, using Lemma 4.11 we obtain MSE convergence with rate $O(1/t^{\hat{\delta}})$. ∎

Remark 4.12. Even though, we see that the convergence factor $\hat{\delta}$ depends on the system parameters, i.e., on the network and sensing model and also on the innovation and consensus nonlinearities, it is easy to see that $\hat{\delta} \in (0,1)$ regardless of the system parameters. Recall that Theorem 4.3 shows that the proposed estimator (3.1) obtains rate $1/t$ in the *weak convergence sense*, while Theorem 4.6 shows that (3.1) obtains a slower convergence rate, but in the sense of the *mean squared convergence*. Note that this is not a contradiction, and Theorem 4.6 adds information with respect to Theorem 4.3. Namely, it is well known that mean squared convergence implies convergence in distribution; therefore, with the same assumptions as in Theorem 4.6, the convergence rate $1/t^{\hat{\delta}}$ is also attainable for convergence in distribution. In contrast, from Theorem 4.3, we can not conclude that the rate of the mean squared convergence is also $1/t$.

Remark 4.13. In fact, we next show that, in the presence of the heavy-tailed observation noise considered here, the MSE convergence rate cannot be as fast as $1/t$, for any estimator (even not for centralized ones). In this sense, the fact that quantity $\hat{\delta}$ is strictly smaller than one is not a consequence of loose bounds, but it is rather due to the intrinsic difficulty of the estimation problem. To be specific, we consider here the special case where each agent i observes a scalar parameter $\theta^\star \in \mathbb{R}$ according to

$$z_i(t) = \theta^\star + n_i^t, \tag{4.22}$$

where n_i^t satisfies Assumption 2.2. In this case, the proposed estimator (3.1) can be viewed as a mean estimator of the probability density function $p_{\mathrm{o}}(u - \theta^\star)$. Let us denote by $\mathcal{P}$ the class of all probability density functions $p_{\mathrm{o}}(u-\theta^\star)$ such that p_{o} is the pdf of the observation noise that satisfies Assumption 2.2, for any $\theta^\star \in \mathbb{R}$. Extending the results from [8] (see Appendix G in [34]), we prove that, for any $\theta^\star \in \mathbb{R}$, and for any mean estimator $\hat{\theta}_t$, the following holds:

$$\sup_t \sup_{p\in\mathcal{P}} tN\mathbb{E}[|\hat{\theta}_t-\theta^\star|^2] = +\infty. \tag{4.23}$$

On the other hand, Theorem 4.6 shows that, with the proposed distributed estimator (3.1), the following holds:

$$\sup_t \sup_{p\in\mathcal{P}} (tN)^{\hat{\delta}}\mathbb{E}[|\hat{\theta}_t-\theta^\star|^2] < +\infty,$$

for some $\hat{\delta} \in (\frac{1}{2}, 1)$[3]

Remark 4.14. Theorems 4.2, 4.3 and 4.6 continue to hold even if the linear transformation vectors $\mathbf{h}_i$ in (2.1) are no longer static (see Appendix H in [34]). That is, we can allow that each agent i at each time $t = 0, 1, ...,$ makes the observation by:

$$z_i^t = (\mathbf{h}_i^t)^\top \boldsymbol{\theta}^\star + n_i^t. \tag{4.24}$$

Here, for each agent i, for each time step t, the linear transformation vector $\mathbf{h}_i^t$ is a random variable that satisfies the following assumptions.

1. For each agent i and each time step $t = 0, 1, ...,,$ the linear transformation vector is given by $\mathbf{h}_i^t = \overline{\mathbf{h}}_i + \widetilde{\mathbf{h}}_i^t$, where the vector $\overline{\mathbf{h}}_i \in \mathbf{R}^M$ is deterministic, and vector $\widetilde{\mathbf{h}}_i^t \in \mathbf{R}^M$ is a random vector;
2. The sequence of vectors $\{[\widetilde{\mathbf{h}}_1^t, \widetilde{\mathbf{h}}_2^t, ..., \widetilde{\mathbf{h}}_N^t]\}$ is i.i.d., with finite second moment, and it is independent of the sequences $\mathbf{n}^t$ and $\boldsymbol{\xi}_{ij}^t$ for $\{i,j\} \in E$;
3. At each agent $i = 1, ..., N$ at each time $t = 0, 1, ...,,$ each entry $\ell = 1, 2, ...M$ $[\widetilde{\mathbf{h}}_i^t]_\ell$ has the same probability density function p_{h};
4. The pdf p_{h} is symmetric, i.e. $p_{\mathrm{h}}(u) = p_{\mathrm{h}}(-u)$, for every $u \in \mathbb{R}$ and $p_{\mathrm{h}}(u) > 0$ for $|u| \leq c_{\mathrm{h}}$, for some constant $c_{\mathrm{h}} > 0$;
5. The matrix $\sum_{i=1}^N \overline{\mathbf{h}}_i \left(\overline{\mathbf{h}}_i\right)^\top$ is invertible.

5. Analytical and numerical examples.

In this section we provide analytical and numerical examples that illustrate results from Section 4.

Example 1: We consider the network where each agent i observes a scalar parameter $\theta^\star \in \mathbb{R}$ following the linear regression model:

$$z_i(t) = h\theta^\star + n_i^t, \tag{5.1}$$

where $h \neq 0$ and $n_i(t)$ is zero mean and i.i.d. in time and across agents. For simplicity, we assume that the underlying graph of the network is regular, with degree d. We assume that there is no communication noise between agents, i.e., $\boldsymbol{\xi}_{ij} \equiv 0$ for $(i,j) \in E_d$. We additionally assume that the nonlinearity on the consensus part Ψ_{c} in (3.1) is the identity function and the nonlinearity on the innovation part is $\Psi_{\mathrm{o}}(w) = B\tanh(w/B)$, for $B > 0$. Therefore, algorithm (3.1) is now given by:

$$x_i^{t+1} = x_i^t - \alpha_t\left(\frac{b}{a}\sum_{j\in\Omega_i}\left(x_i^t - x_j^t\right) - h\Psi_{\mathrm{o}}\left(z_i^t - hx_i^t\right)\right), \tag{5.2}$$

for each agent i and each time t. From Theorem 4.3, we have that the asymptotic covariance matrix is given by (4.13) and matrix $\mathbf{S}_0$ is now given by $\mathbf{S}_0 = \sigma_{\mathrm{o}}^2 h^2 \mathbf{I}$ and $\sigma_{\mathrm{o}}^2 = \int |\Psi_{\mathrm{o}}(w)|^2 d\Phi_{\mathrm{o}}(w)$ is the effective observation noise. Following the same procedure as in [18, 15], for $\boldsymbol{\Sigma} = \frac{1}{2}\mathbf{I} - a\varphi_{\mathrm{o}}'(0)h^2\mathbf{I}$, we have that the average per-agent

[3] Notice that in the centralized case, the observations are collected in batches of fixed size N. That is, after t time steps, there are Nt observations. Henceforth, we include quantity N in (4.23) for a precise statement. Note that, since N is constant and the supremum is taken with respect to t, the inclusion of N is not necessary.

asymptotic variance, denoted by $\sigma_B^2 = \frac{1}{N}\operatorname{Tr}(\mathbf{S})$, is equal to $\sigma_B^2 = \frac{a^2\sigma_{\mathrm{o}}^2 h^2}{2ah^2\varphi_{\mathrm{o}}'(0)-1}$, for $a > \frac{1}{2h^2\varphi_{\mathrm{o}}'(0)}$ (see Appendix F in [34]). Therefore, we need to change the constant a when changing B, i.e., we define $a = a(B) = \frac{1}{2h^2\varphi_{\mathrm{o}}'(0)(B)} + \epsilon$[4], for some constant $\epsilon > 0$, we rewrite σ_B^2 as follows (see Appendix F in [34]), $\sigma_B^2 = \frac{(1+2h^2\varphi_{\mathrm{o}}'(0)\epsilon)^2\sigma_{\mathrm{o}}^2(B)}{8h^4\varphi_{\mathrm{o}}'(0)^3\epsilon}$. For the nonlinearity Ψ_{o} that is considered here, we have that $\sigma_{\mathrm{o}}^2 = \int\limits_{-\infty}^{+\infty} B^2\tanh^2\left(\frac{w}{B}\right) f(w)dw$, and $\varphi_{\mathrm{o}}'(0) = \int\limits_{-\infty}^{+\infty}\Psi'(w)f(w)dw = \int\limits_{-\infty}^{+\infty}\frac{1}{\cosh^2(\frac{w}{B})}f(w)dw$. Notice that both functions σ_{o}^2 and $\varphi_{\mathrm{o}}'(0)$ are increasing with respect to B (see Appendix F in [34]). Since we have that $|B^2\tanh^2(\frac{w}{B})f(w)| \le |w^2 f(w)|$ and $|\frac{1}{\cosh^2(\frac{w}{B})}f(w)| \le |f(w)|$ for all $w \in \mathbb{R}$ and all $B > 0$, using the Lebesgue's dominated convergence theorem, we have that $\lim\limits_{B\to 0^+}\sigma_{\mathrm{o}}^2 = 0$, $\lim\limits_{B\to+\infty}\sigma_{\mathrm{o}}^2 = \sigma_\eta^2$, $\lim\limits_{B\to 0^+}\varphi_{\mathrm{o}}'(0) = 0$, $\lim\limits_{B\to+\infty}\varphi_{\mathrm{o}}'(0) = 1$, where σ_η^2 is the variance of the observation noise η. Therefore, we have that $\sigma_0^2 = \lim\limits_{B\to 0^+}\sigma_B^2 = +\infty$ (see Appendix F in [34]), and $\sigma_\infty^2 = \lim\limits_{B\to+\infty}\sigma_B^2 = \frac{(1+2h^2\epsilon)^2\sigma_\eta^2}{8h^4\epsilon}$. Suppose now that the variance of the observation noise η is infinite, i.e. $\sigma_\eta^2 = +\infty$. This means that $\sigma_\infty^2 = +\infty$. For the continuous function σ_B^2, defined for all $B \in (0, +\infty)$, we have that $\lim\limits_{B\to 0^+}\sigma_B^2 = \lim\limits_{B\to+\infty}\sigma_B^2 = +\infty$. Therefore, there exists an optimal $B^\star$ such that $\sigma_{B^\star}^2 = \inf\limits_{B\in(0,\infty)}\sigma_B^2$. Note that the case $B\to\infty$ corresponds to a $\mathcal{LU}$ scheme from [18], while the case $B \to 0$ corresponds to each agent working in isolation. Therefore, we show analytically on the simple class of nonlinearities Ψ_{o} (hyperbolic tangent), that cooperation through a nonlinear mapping Ψ_{o} strictly improves performance with respect to both using linear and non-cooperative schemes.

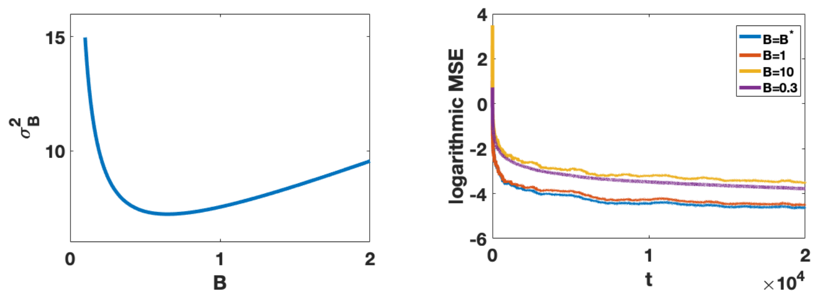
To numerically illustrate the above results, we now consider a sensor (agents) network with $N = 8$ agents, setting that the underlying topology is given by a regular graph with degree $d = 3$. The true parameter is $\theta^* = 1$, the observation parameter is $h = 1$, and the observation noise for each agent's measurements has the following pdf

$$f(w) = \frac{\beta - 1}{2\,(1+|w|)^\beta}, \tag{5.3}$$

with $\beta = 2.05$, which has an infinite variance. Recall that we assumed that there is no communication noise between agents. We set the consensus parameter as $b = 1$ and the innovation parameter as $a = a(0.3) = \frac{1}{2h^2\varphi_{\mathrm{o}}'(0)(0.3)} + 0.1$. Figure 1a shows the average per-agent asymptotic variance σ_B^2 versus B. As it can be seen, optimal $B^\star$ approximately equals $B^\star = 0.65$. Using Monte Carlo simulations, we compare numerically an estimated per-sensor MSE across iterations, for the optimal $B^\star$ and for some sub-optimal choices of B. We can see that the algorithm performs better for the optimal value $B^\star$ than for the other considered suboptimal choices of B (see Figure 1b), hence confirming the theory.

Example 2: In this example we provide analysis in the terms of the average per node variance with respect to the level of the mutual dependence of observation and communication noise. Once more, we consider the network where each agent i observes a scalar parameter $\theta^\star \in \mathbb{R}$ following the linear regression model (5.1) and we assume that the underlying graph of the network is regular, with degree d. As it is said, we now allow observation and communication noise to be mutually dependent. For

[4] ϵ is added since we need to have that $a > \frac{1}{2h^2\varphi_{\mathrm{o}}'(0)}$.



(a) (b)

FIG. 1. *(a) Average per-agent asymptotic variance σ_B^2 versus B (b) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for the different choices of B*

simplicity, we consider the case when that dependence between communication noise ξ_{ij}^t and observation noise n_i is given by $\xi_{ij} = \rho\, n_i^t + \sqrt{1-\rho^2}\, \hat{n}_i^t$, at each time $t = 0, 1, ..$ and for all tuples $\{i, j\} \in E$, where, $\rho \in (-1, 1)$, sequence $\{\hat{n}_i^t\}$ is independently identically distributed in time t and across all agents i. Moreover, n_i^t are $\hat{n}_j^s$ mutually independent whenever $(i, t) \neq (j, s)$. Here, it is easy to see that we have strong positive correlation if $\rho \to 1$, strong negative correlation if $\rho \to -1$ and we do not have any correlation if $\rho = 0$. Moreover, we set that $\Psi_{\mathrm{o}}(w) = \Psi_{\mathrm{c}}(w) = \operatorname{sign} w$, and hence, algorithm (3.1) is given by

(5.4)

$$x_i^{t+1} = x_i^t - \alpha_t \left(\frac{b}{a} \sum_{j \in \Omega_i} \Psi_{\mathrm{c}} \left(x_i^t - x_j^t + \rho\, n_i^t + \sqrt{1-\rho^2}\, \hat{n}_i^t \right) - h \Psi_{\mathrm{o}} \left(h(\theta^\star - x_i^t) + n_i \right) \right).$$

Analogously to the previous example, we have that the average per-agent asymptotic variance σ_ρ^2 is given by

$$\sigma_\rho^2 = \frac{b^2 \sigma_{\mathrm{c}}^2 d^2 + a^2 h^2 \sigma_{\mathrm{o}}^2 - 2abhd\sigma_{\mathrm{oc}}}{N\left(2ah^2 \varphi_{\mathrm{o}}'(0) - 1\right)} \tag{5.5}$$

$$+ \frac{b^2 \sigma_{\mathrm{c}}^2 d^2 + a^2 h^2 \sigma_{\mathrm{o}}^2 - 2abhd\sigma_{\mathrm{oc}}}{N} \sum_{i=2}^{N} \frac{1}{2b\varphi_{\mathrm{c}}'(0)\lambda_i + 2ah^2\varphi_{\mathrm{o}}'(0) - 1}, \tag{5.6}$$

since $\mathbf{S}_0 = \left(\frac{b^2}{a^2} \sigma_{\mathrm{c}}^2 d^2 + \sigma_{\mathrm{o}}^2 h^2 - 2\frac{b}{a} hd\sigma_{\mathrm{oc}} \right) \mathbf{I}$ and $\mathbf{\Sigma} = \frac{1}{2}\mathbf{I} - a\left(\frac{b}{a} \varphi_{\mathrm{c}}'(0) \mathbf{L} + \varphi_{\mathrm{o}}'(0) h^2 \mathbf{I} \right)$. Here, regardless of ρ we have that $\sigma_{\mathrm{o}}^2 = \sigma_{\mathrm{c}}^2 = 1$ and $\varphi_{\mathrm{o}}'(0) = 2p_n(0)$ (see [15]). On the other hand, σ_{oc} which is effective cross-covariance between the observation and the communication noise after passing through the appropriate nonlinearity and $\varphi_{\mathrm{c}}'(0)$

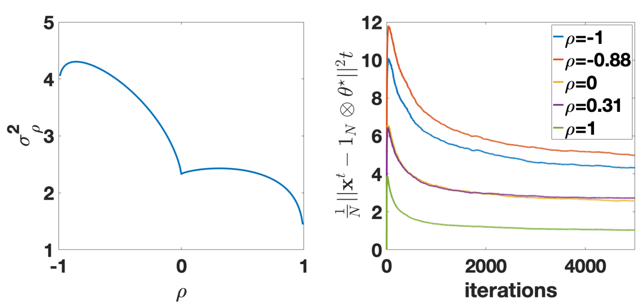
are functions with respect to ρ. We have that

$$\sigma_{\mathrm{oc}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Psi_{\mathrm{c}}(\rho x + \sqrt{1-\rho^2}y)\Psi_{\mathrm{o}}(x)p_{\hat{n}}(y)p_n(x)dxdy \tag{5.7}$$

$$= \int_{0}^{+\infty}\int_{\frac{-\rho x}{\sqrt{1-\rho^2}}}^{\infty} p_{\hat{n}}(y)p_n(y)dydx - \int_{0}^{+\infty}\int_{-\infty}^{\frac{-\rho x}{\sqrt{1-\rho^2}}} p_{\hat{n}}(y)p_n(y)dydx \tag{5.8}$$

$$- \int_{-\infty}^{0}\int_{\frac{-\rho x}{\sqrt{1-\rho^2}}}^{\infty} p_{\hat{n}}(y)p_n(y)dydx + \int_{-\infty}^{0}\int_{-\infty}^{\frac{-\rho x}{\sqrt{1-\rho^2}}} p_{\hat{n}}(y)p_n(y)dydx, \tag{5.9}$$

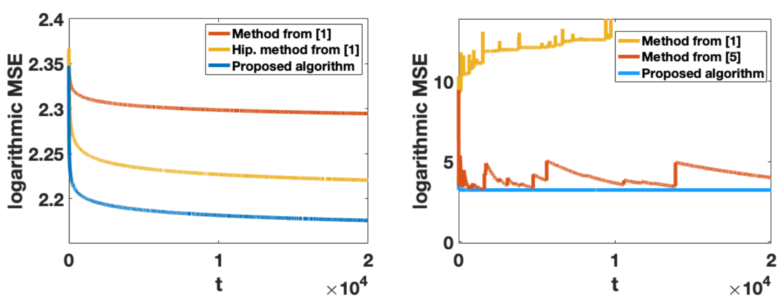
and we see that $\sigma_{\mathrm{oc}} \to 0$ as $\rho \to 0$, $\sigma_{\mathrm{oc}} \to 1$ as $\rho \to 1$ and $\sigma_{\mathrm{oc}} \to -1$ as $\rho \to -1$. Moreover, we have that $\varphi_{\mathrm{c}}'(0) = 2\int_{-\infty}^{\infty} p_{\hat{n}}(-\rho x)p_n(\sqrt{1-\rho^2}x)dx$, and again, it is easy to see that, $\varphi_{\mathrm{c}}'(0) \to 2p_n(0)$ as $\rho \to \pm 1$ and $\varphi_{\mathrm{c}}'(0) \to 2p_{\hat{n}}(0)$ as $\rho \to 0$. To demonstrate the above results, again we consider a sensor (agents) network with $N = 8$ agents, setting that the underlying topology is given by a regular graph with degree $d = 3$. The true parameter is $\theta^* = 1$, the observation parameter, the innovation parameter and consensus parameter are $h = a = b = 1$. We set that for all i, n_i and $\hat{n}_i$ have the pdf as in (5.3) with $\beta = 2.05$. Figure 2a shows σ_ρ^2 with respect to ρ. As it can be seen, the lowest σ_ρ^2 is attained at $\rho = 1$, also σ_ρ^2 has two local maxima at $\rho \approx -0.88$ and at $\rho \approx 0.31$. Figure 2b shows the comparison of Monte Carlo simulation for $\frac{1}{N}\|\mathbf{x}^t - \mathbf{1}\otimes\theta\|^2\, t$ for different choices of ρ. Moreover, Figure 2b justifies the results presented in 2a, in the sense that $\frac{1}{N}\|\mathbf{x}^t - \mathbf{1}\otimes\theta\|^2\, t$ is minimal for $\rho = 1$ and maximal for $\rho = -0.88$. Finally, we note that, while the two local maxima obtained here are specific for the simplistic correlation and sensing model assumed here for analytical tractability, we observe numerically for more general models that the general trend of this example is preserved, in the sense that higher (more positive) correlations lead to a better performance.



(a) (b)

FIG. 2. *(a) Average per-agent asymptotic variance σ_B^2 versus B (b) Monte Carlo-estimation of $\frac{1}{N}\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^\star\|^2\, t$ for different choices of ρ*

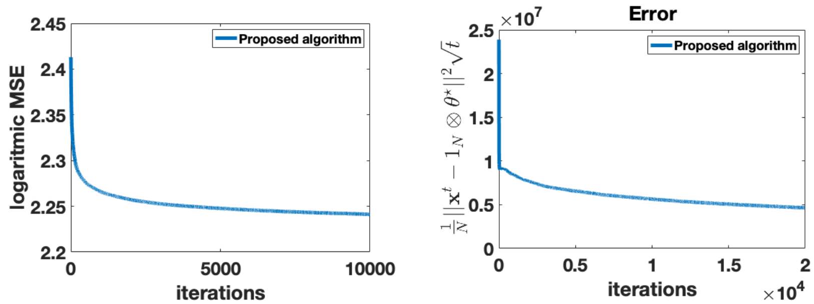
5.1. Numerical simulations. In this subsection, we demonstrate the performance of proposed consensus+innovations estimator in a larger sensor network. We consider a sensor network with $N = 40$ agents where the underlying topology is an instance of a random geometric graph; we used randomly generated true parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^{10}$, whose entries are drawn mutually independently form the uniform distribution on $[-10, 10]$; we used randomly generated observation vectors $\mathbf{h}_i \in \mathbb{R}^{10}$, for which the condition 2 of Assumption 2.1 is verified to be true. We set the consensus parameter as $b = 1$ and and step-size parameter as $\delta = 1$. First, we compare the proposed consensus+innovations estimator with the method from [1] and its hypothetical variant in the case when there is no communication noise, but in the presence of heavy-tailed observation noise with pdf as in (5.3) for $\beta = 2.05$. Here, we used the same algorithm settings and the same nonlinearities for the proposed algorithm as in Example 1, with a slight change, i.e., we set that $B = 10$ and $a = 0.2$ For method from [1] and its hypothetical variant (see Appendix E in [34]), we set that $B_i = 2$, $\phi_{i,1}(x) = x$ and $\phi_{i,2}(x) = \tanh(x)$ for all agents i. Furthermore, we set that weighting coefficients are chosen according to $a_{ij} = \frac{\tilde{\mathbf{A}}_{ij}}{\sum\limits_{\ell \in \mathcal{N}_i} \tilde{\mathbf{A}}_{\ell i}}$, where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. Moreover, for the smoothing recursions, zero initial conditions are assumed, ν_i is set to 0.9 for every agent i and $\epsilon = 10^{-2}$. We can see all methods manage to (slowly) decrease MSE over iterations, with the proposed method exhibiting the best performance among the three methods considered. Figure 3b shows Monte Carlo simulation of the MSE for the proposed algorithm, algorithm from [1] and the algorithm in [15], when communication between agents is also contaminated with heavy-tailed communication noise. Here, for the proposed algorithm we set that both nonlinearities are $\Psi_{\mathrm{o}}(w) = \Psi_{\mathrm{c}}(w) = B \tanh(w/B)$, for $B = 10$ and $a = 1$. Further, we use the same algorithm setting for the method in [1] as in the previous simulation example, and we use the same nonlinearity on the consensus part and the same B for algorithm from [15] as in the proposed algorithm. We can see that both [15] and [1] here fail to converge, while the proposed method still effectively reduces MSE.



(a) (b)

FIG. 3. *(a) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for proposed algorithm for $B = 10$, method from [1] and its hypothetical variant (b) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for proposed algorithm, algorithm form [1] and algorithm from [15]*

We next present the scenario where the observation and communication noises are mutually dependent. To do this, we set that the i-th element of the observation noise $\mathbf{n}$ is given by $\mathbf{n}_i = \mathbf{v}_i \exp\left(\frac{h}{2}\mathbf{v}_i^2\right)$, where $\mathbf{v}$ has standard normal distribution and h is a heavy-tail parameter (see [9]). Moreover, the ℓ-th element of the communication noise $\boldsymbol{\xi}_{ij}$ is given by $[\boldsymbol{\xi}_{ij}]_\ell = [\mathbf{w}_{ij}]_\ell \exp\left(\frac{h}{2}[\mathbf{w}_{ij}]_\ell^2\right)$, where $\mathbf{w}_{ij}$ is the linear transformation of $\mathbf{v}$, i.e., $\mathbf{w}_{ij} = \mathbf{W}_{ij}\mathbf{v}$ and $\mathbf{W}_{ij} \in \mathbb{R}^{M \times N}$ is a randomly generated matrix independent of the observation noise. Figure 4a presents Monte Carlo estimates of per-agent MSE across iterations. Figure 4b shows Monte Carlo simulation of quantity $\frac{1}{N}\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^\star\|^2 \sqrt{t}$. For this numerical setting, from the Figure 4b, we can deduce that $E[\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^\star\|^2]$ decreases at least as fast as $O(\frac{1}{\sqrt{t}})$, hence confirming our MSE rate theory.



(a) (b)

FIG. 4. *(a) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for proposed algorithm when link failures can occur for $B = 1$ and $h = 10$ (b) Monte Carlo-estimation of $\frac{1}{N}\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^\star\|^2 \sqrt{t}$ for $B = 10$ and $h = 2$.*

6. Conclusion. We have studied distributed consensus+innovations estimation under the simultaneous presence of heavy-tailed (infinite variance) correlated sensing and communication noises. This setting is in contrast with existing work that either always assumes a finite-variance sensing noise. We developed a nonlinear estimator and established its almost sure convergence and asymptotic normality. Furthermore, we showed that the estimator achieves a sublinear MSE convergence rate $O(1/t^{\kappa})$, and we explicitly charaterized the rate $\kappa i \in (0, 1)$ in terms of system parameters. Analytical examples illustrate the role of the nonlinearities incorporated in the method and the effects of noises correlation. Finally, numerical simulations corroborate our findings and demonstrate that the proposed distributed estimator converges under the simultaneous presence of heavy-tailed (infinite variance) correlated sensing and communication noises, while, for the same setting, existing distributed estimators fail to converge.

Acknowledgements The work of D. Bajovic, D. Jakovetic and M. Vukovic is supported by the Ministry of Education, Science and Technological Development, Republic of Serbia. Moreover, the work of D. Bajovic and D. Jakovetic is supported by the European Union's Horizon 2020 Research and Innovation program under grant agreement No. 957337. The paper reflects only the view of the authors and the Commission is not responsible for any use that may be made of the information it contains. The work of D. Jakovetic is also supported by the Provincial Secretariat for Higher Education and Scientific Research, grant no 142-451-2593/2021-01/2.

REFERENCES

[1] S. Al-Sayed, A. M. Zoubir, and A. H. Sayed, *Robust distributed estimation by networked agents*, IEEE Transactions on Signal Processing, 65 (2017), pp. 3909–3921.

[2] D. Bajovic, D. Jakovetic, J. M. Moura, J. Xavier, and B. Sinopoli, *Large deviations performance of consensus+ innovations distributed detection with non-gaussian observations*, IEEE Transactions on Signal Processing, 60 (2012), pp. 5987–6002.

[3] D. Bajovic, D. Jakovetic, J. Xavier, B. Sinopoli, and J. Moura, *Distributed detection via gaussian running consensus: Large deviations asymptotic analysis*, Signal Processing, IEEE Transactions on, 59 (2011), pp. 4381 – 4396.

[4] Y. Chen, S. Kar, and J. Moura, *Resilient distributed field estimation*, SIAM Journal on Control and Optimization, 58 (2020), pp. 1429–1456.

[5] S. Chouvardas, K. Slavakis, and S. Theodoridis, *Adaptive robust distributed learning in diffusion sensor networks*, Signal Processing, IEEE Transactions on, 59 (2011), pp. 4692 – 4707.

[6] L. Clavier, T. Pedersen, I. Larrad, M. Lauridsen, and M. Egan, *Experimental evidence*

for heavy tailed interference in the IoT, IEEE Communications Letters, 25 (2021), pp. 692–695.

[7] S. Dasarathan, C. Tepedelenlioglu, M. Banavar, and A. Spanias, *Robust consensus in the presence of impulsive channel noise*, Signal Processing, IEEE Transactions on, 63 (2014).

[8] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira, *Sub-gaussian mean estimators*, The Annals of Statistics, 44 (2016), pp. 2695–2725.

[9] G. Goerg, *The lambert way to gaussianize heavy-tailed data with the inverse of tukey's h transformation as a special case*, The Scientific World Journal, 2015 (2015).

[10] M. Haenggi and R. Ganti, *Interference in large wireless networks*, Foundations and Trends in Networking, 3 (2009), pp. 127–248.

[11] B. Hughes, *Alpha-stable models of multiuser interference*, in 2000 IEEE International Symposium on Information Theory (Cat. No.00CH37060), 2000, pp. 383–.

[12] J. Ilow and D. Hatzinakos, *Analytic alpha-stable noise modeling in a poisson field of interferers or scatterers*, Signal Processing, IEEE Transactions on, 46 (1998), pp. 1601 – 1611.

[13] D. Jakovetic, D. Bajovic, A. K. Sahu, S. Kar, N. Milosevic, and D. Stamenkovic, *Nonlinear gradient mappings and stochastic optimization: A general framework with applications to heavy-tail noise*. arXiv preprint, https://arxiv.org/abs/2204.02593, 2022.

[14] D. Jakovetic, J. M. F. Moura, and J. Xavier, *Distributed detection over noisy networks: Large deviations analysis*, IEEE Transactions on Signal Processing, 60 (2012), pp. 4306–4320.

[15] D. Jakovetic, M. Vukovic, D. Bajovic, A. K. Sahu, and S. Kar, *Distributed recursive estimation under heavy-tail communication noise*, SIAM Journal on Control and Optimization, 61 (2023), pp. 1582–1609.

[16] S. Kar and J. Moura, *Asymptotically efficient distributed estimation with exponential family statistics*, IEEE Transactions on Information Theory, 60 (2014), pp. 4811–4831.

[17] S. Kar, J. Moura, and H. V. Poor, *Distributed linear parameter estimation: Asymptotically efficient adaptive strategies*, SIAM Journal on Control and Optimization, 51 (2013), pp. 2200–2229.

[18] S. Kar, J. M. F. Moura, and K. Ramanan, *Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication*, IEEE Transactions on Information Theory, 58 (2012), pp. 3575–3605.

[19] S. Kumar, U. K. Sahoo, A. K. Sahoo, and D. P. Acharya, *Diffusion minimum-wilcoxon-norm over distributed adaptive networks: Formulation and performance analysis*, Digital Signal Processing, 51 (2016), pp. 156–169.

[20] A. Lalitha, T. Javidi, and A. D. Sarwate, *Social learning and distributed hypothesis testing*, IEEE Transactions on Information Theory, 64 (2018), pp. 6161–6179.

[21] Z. Li and S. Guan, *Diffusion normalized huber adaptive filtering algorithm*, Journal of the Franklin Institute, 355 (2018), pp. 3812–3825.

[22] Q. Liu and A. Ihler, *Distributed estimation, information loss and exponential families*. arXiv preprint, https://arxiv.org/abs/1410.2653, 2014.

[23] C. Lopes and A. Sayed, *Diffusion least-mean squares over adaptive networks: Formulation and performance analysis*, IEEE Transactions on Signal Processing, 56 (2008), pp. 3122–3136.

[24] G. Mateos, I. Schizas, and G. Giannakis, *Distributed recursive least-squares for consensus-based in-network adaptive estimation*, Signal Processing, IEEE Transactions on, 57 (2009), pp. 4583 – 4588.

[25] V. Matta, P. Braca, S. Marano, and A. H. Sayed, *Diffusion-based adaptive distributed detection: Steady-state performance in the slow adaptation regime*, IEEE Transactions on Information Theory, 62 (2016), pp. 4710–4732.

[26] S. Modalavalasa, U. Sahoo, A. Sahoo, and S. Baraha, *A review of robust distributed estimation strategies over wireless sensor networks*, Signal Processing, 188 (2021), p. 108150.

[27] A. Nedic, A. Olshevsky, and C. A. Uribe, *Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs*, in 2015 American Control Conference (ACC), IEEE, 2015, pp. 5884–5889.

[28] M. B. Nevel'son and R. Z. Has' minskii, *Stochastic approximation and recursive estimation*, vol. 47, American Mathematical Soc., 1976.

[29] B. Polyak and Y. Tsypkin, *Adaptive estimation algorithms: Convergence, optimality, stability*, Automation and Remote Control, 1979 (1979).

[30] S. Ram, V. Veeravalli, and A. Nedic, *Distributed and Recursive Parameter Estimation*, Springer Science & Business Media, 2009, pp. 17–38.

[31] B. Selim, M. S. Alam, V. Carvalho, G. Kaddoum, and B. L. Agba, *Noma-based iot net-*

works: Impulsive noise effects and mitigation, IEEE Communications Magazine, 58 (2020), pp. 69–75.

[32] W. Shi, Q. Ling, G. Wu, and W. Yin, *Extra: An exact first-order algorithm for decentralized consensus optimization*. arXiv preprint, https://arxiv.org/abs/1404.6264, 2014.

[33] S. Theodoridis, K. Slavakis, and I. Yamada, *Adaptive learning in a world of projections*, Signal Processing Magazine, IEEE, 28 (2011), pp. 97 – 123.

[34] M. Vukovic, D. Jakovetic, D. Bajovic, and S. Kar, *Nonlinear consensus+innovations under correlated heavy-tailed noises: Mean square convergence rate and asymptotics*. arXiv preprint, https://arxiv.org/abs/2212.11959, 2022.

[35] F. Wen, *Diffusion least mean p-power algorithms for distributed estimation in alpha-stable noise environments*, Electronics Letters, 49 (2013).

[36] M. Win, P. Pinto, and L. Shepp, *A mathematical theory of network interference and its applications*, Proceedings of the IEEE, 97 (2009), pp. 205 – 230.

[37] X. Yang and A. Petropulu, *Co-channel interference modeling and analysis in a poisson field of interferers in wireless communications*, IEEE Transactions on Signal Processing, 51 (2003), pp. 64–76.

[38] X. Zhao, S.-Y. Tu, and A. H. Sayed, *Diffusion adaptation over networks under imperfect information exchange and non-stationary data*, IEEE Transactions on Signal Processing, 60 (2012), pp. 3460–3475.

Appendix.

A. Some results on Stochastic approximation. We make use of the following standard stochastic approximation result, see [28], see also [18].

THEOREM 6.1. *Let $\{\mathbf{x}^t \in \mathbb{R}^l\}_{t\geq 0}$ be a random sequence:*

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \alpha_t[\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)], \tag{6.1}$$

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

- ***B1:*** *The function $\boldsymbol{\gamma}(t, \cdot, \cdot) : \mathbb{R}^l \times \Omega \to \mathbb{R}$ is $\mathcal{B}^l \otimes \mathcal{F}$ measurable for every t; $\mathcal{B}^l$ is the Borel algebra of $\mathbb{R}^l$.*
- ***B2:*** *There exists a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ of $\mathcal{F}$, such that, for each t, the family of random vectors $\{\boldsymbol{\gamma}(t, \mathbf{x}, \omega)\}_{\mathbf{x}\in\mathbb{R}^l}$ is $\mathcal{F}_t$ measurable, zero-mean and independent of $\mathcal{F}_{t-1}$.*

(If Assumtions B1, B2 hold, $\{\mathbf{x}(t)\}_{t\geq 0}$, is Markov.)

- ***B3:*** *There exists a twice continuously differentiable $V(\mathbf{x})$ with bounded second order partial derivatives and a point $\mathbf{x}^* \in \mathbb{R}^l$ satisfying*
$$V(\mathbf{x}^*) = 0, V(\mathbf{x}) > 0, \mathbf{x} \neq \mathbf{x}^*, \lim_{||\mathbf{x}||\to\infty} V(\mathbf{x}) = \infty,$$
$$\sup_{\epsilon<||\mathbf{x}-\mathbf{x}^*||<\frac{1}{\epsilon}} \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle < 0, \forall \epsilon > 0.$$
- ***B4:*** *There exists constants $k_1, k_2 > 0$, such that,*
$$||\mathbf{r}(\mathbf{x})||^2 + \mathbb{E}[||\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)||^2] \leq k_1(1 + V(\mathbf{x})) - k_2\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x})\rangle$$
- ***B5:*** *The weight sequence $\{\alpha(t)\}_{t\geq 0}$ satisfies*
$$\alpha_t > 0, \sum_{t\geq 0} \alpha_t = \infty, \sum_{t\geq 0} \alpha_t^2 < \infty.$$
- ***C1:*** *The function $\mathbf{r}(\mathbf{x})$ admits the representation*
$$\mathbf{r}(\mathbf{x}) = \mathbf{B}(\mathbf{x} - \mathbf{x}^*) + \boldsymbol{\delta}(\mathbf{x}), \tag{6.2}$$
where
$$\lim_{\mathbf{x}\to\mathbf{x}^*} \frac{||\boldsymbol{\delta}(\mathbf{x})||}{||\mathbf{x} - \mathbf{x}^*||} = 0. \tag{6.3}$$
(Note, in particular, if $\boldsymbol{\delta}(\mathbf{x}) \equiv 0$ then (6.3) is satisfied.)
- ***C2:*** *The weight sequence $\{\alpha_t\}_{t\geq 0}$ is of form*
$$\alpha_t = \frac{a}{t+1}, \forall t \geq 0, \tag{6.4}$$
*where $a > 0$ is a constant (note that **C2** implies **B5**).*
- ***C3:*** *Let $\mathbf{I}$ be the $l \times l$ identity matrix and $a, \mathbf{B}$ as in (6.4) and (6.2), respectively. Then, the matrix $\boldsymbol{\Sigma} = a\mathbf{B} + \frac{1}{2}\mathbf{I}$ is stable.*
- ***C4:*** *The entries of the matrices, $\forall t \geq 0$, $x \in \mathbb{R}^l$,*
$$\mathbf{A}(t, \mathbf{x}) = \mathbb{E}[\boldsymbol{\gamma}(t, \mathbf{x}, \omega)\boldsymbol{\gamma}^\top(t, \mathbf{x}, \omega)],$$
are finite, and the following limit exists:
$$\lim_{t\to\infty, \mathbf{x}\to\mathbf{x}^*} \mathbf{A}(t, \mathbf{x}) = \mathbf{S}_0.$$
- ***C5:*** *There exists $\epsilon > 0$, such that*
$$\lim_{R\to\infty} \sup_{||\mathbf{x}-\mathbf{x}^*||<\epsilon} \sup_{t\geq 0} \int_{||\boldsymbol{\gamma}(t+1,\mathbf{x},\omega)||>R} ||\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)||^2 dP = 0$$

Let Assumptions B1–B5 hold for $\{\mathbf{x}(t)\}_{t\geq 0}$ in (6.1). Them, starting from an arbitrary initial state, the Markov process, $\{\mathbf{x}^t\}_{t\geq 0}$, converges a.s. to $\mathbf{x}^*$. In other words,

$$\mathbf{P}[\lim_{t\to\infty} \mathbf{x}^t = \mathbf{x}^*] = 1.$$

The normalized process, $\{\sqrt{t}(\mathbf{x}^t - \mathbf{x}^*)\}_{t\geq 0}$, is asymptotically normal if, besides As-

sumptions B1–B5, Assumptions C1–C5 are also satisfied. In particular, as $t \to \infty$

$$\sqrt{t}(\mathbf{x}^t - \mathbf{x}^*) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{S}), \tag{6.5}$$

where $\Rightarrow$ *denotes convergence in distribution (weak convergence). Also, asymptotic variance,* $\mathbf{S}$*, in* (6.5) *is*

$$\mathbf{S} = a^2 \int\limits_0^\infty e^{\mathbf{\Sigma} v} \mathbf{S}_0 e^{\mathbf{\Sigma}^\top v} dv$$

B. Additional results on nonlinearity φ. We present some properties of the function φ defined in (4.1). As it is stated in [15], we can intuitively see function φ as a convolution-like transformation of nonlinearity $\Psi : \mathbb{R} \to \mathbb{R}$, where the convolution is taken with respect to the probability density function p of random value w. If w is generated by the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have that expectation of

$$v = \Psi(a + w) - \varphi(a) \tag{6.6}$$

is equal to zero, i.e., $\mathbb{E}[v] = 0$. Here, the expectation is taken with respect to $\mathcal{F}$. Hence, for all $t = 0, 1, ...$, we have that expectation of both of the sequences $\boldsymbol{\zeta}^t$, $\boldsymbol{\eta}^t$ defined in (4.3) is equal to zero, due to the fact that communication noise $\boldsymbol{\xi}^t$ and observation noise $\mathbf{n}^t$, $t = 0, 1, ...$, are generated by underlying probability space.

We have following Lemma (see [29], see also [15]).

Lemma 6.2 ([29]). *Consider function* φ *in* (4.1)*, where function* $\Psi : \mathbb{R} \to \mathbb{R}$*, satisfies Assumption 3.1. Then, the following holds:*

1. φ *is odd;*
2. *If* $|\Psi(\nu)| \le c_1$*, for any* $\nu \in \mathbb{R}$*, then* $|\varphi(a)| \le c_1'$*, for any* $a \in \mathbb{R}$*, for some* $c_1' > 0$*;*
3. $\varphi(a)$ *is monotonically nondecreasing;*
4. $\varphi(a) > 0$*, for any* $a > 0$*.*
5. φ *is continuous at zero;*
6. φ *is differentiable at zero, with a strictly positive derivative at zero, equal to:*

$$\varphi'(0) = \sum_{i=1}^{s} \left(\Psi(\nu_i + 0) - \Psi(\nu_i - 0)\right) p(\nu_i) + \sum_{i=0}^{s} \int_{\nu_i}^{\nu_{i+1}} \Psi'(\nu) p(\nu) d\nu, \tag{6.7}$$

where $\nu_i, i = 1, ..., s$ *are points of discontinuity of* Ψ *such that* $\nu_0 = -\infty$ *and* $\nu_{s+1} = +\infty$*, and we recall that* $p(u)$ *is the pdf of random variable* w*.*

From Lemma 6.2, we have that $\varphi(a) = 0$ if and only if $a = 0$. Moreover, there exists a function $\delta : \mathbb{R} \to \mathbb{R}$, which is continuous in the vicinity of zero, such that

$$\varphi(a) = \varphi(0) + \varphi'(0)a + \delta(a) = \varphi'(0)a + \delta(a), \tag{6.8}$$

and $\lim\limits_{a \to 0} \frac{\delta(a)}{a} = 0$.

We now prove boundedness of the function $\mathbf{r}(\cdot)$ in equation (4.10). If condition 2 of Lemma 6.2 is satisfied for both functions φ_{c} and φ_{o}, then the right hand side of (4.10) would be lesser or equal to some positive constant c, which would led to $\|\mathbf{r}(\mathbf{x})\|^2 \le c_1(1 + V(\mathbf{x}))$. Suppose now that condition 3 of Lemma 6.2 is satisfied for the function φ_{c}, then there exists some positive constant c_1 such that

$$\left\|\frac{b}{a}\mathbf{L}_{\boldsymbol{\varphi}_{\mathrm{c}}}(\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right\|^2 = \left(\frac{b}{a}\right)^2\sum_{i=1}^{N}\left\|\sum_{j\in\Omega_i}\boldsymbol{\varphi}_{\mathrm{c}}(\mathbf{x}_i-\mathbf{x}_j)\right\|^2$$
$$\leq \left(\frac{b}{a}\right)^2\sum_{i=1}^{N}\sum_{j\in\Omega_i}\|\boldsymbol{\varphi}_{\mathrm{c}}(\mathbf{x}_i-\mathbf{x}_j)\|^2$$
$$\leq \left(\frac{b}{a}\right)^2\sum_{i=1}^{N}\sum_{j\in\Omega_i}\left(c\left(1+\|\mathbf{x}_i-\mathbf{x}_j\|^2\right)\right)$$
$$\leq \left(\frac{b}{a}\right)^2\sum_{i=1}^{N}\sum_{j\in\Omega_i}\left(c\left(1+\|\mathbf{x}_i-\boldsymbol{\theta}^\star\|^2+\|\mathbf{x}_j-\boldsymbol{\theta}^\star\|^2\right)\right)$$
$$\leq c_1(1+V(\mathbf{x})),$$

since we have that $\|\mathbf{x}_i-\boldsymbol{\theta}^\star\|^2\leq\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\|^2=V(\mathbf{x})$ for all $i=1,2,...,N$. If we assume that condition 3 of Lemma 6.2 is satisfied for the function φ_{o}, we will get that

$$\left\|\mathbf{H}^\top\boldsymbol{\varphi}_{\mathrm{o}}\left(\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)\right\|^2\leq\|\mathbf{H}\|^2\left\|\boldsymbol{\varphi}_{\mathrm{o}}\left(\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right)\right\|^2$$
$$\leq\|\mathbf{H}\|^2c\left(1+\left\|\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\right)\right\|^2\right)$$
$$\leq\|\mathbf{H}\|^2c\left(1+\|\mathbf{H}\|^2\left\|\mathbf{x}-\mathbf{1}_N\otimes\boldsymbol{\theta}^*\right\|^2\right).$$

Therefore, $\left\|\mathbf{H}^\top\boldsymbol{\varphi}_{\mathrm{o}}\left(\mathbf{H}\left(\mathbf{x}-(\mathbf{1}_N\otimes\boldsymbol{\theta}^*)\leq\right)\right)\right\|^2\leq c_1(1+V(\mathbf{x}))$, for some positive constant c_1. Hence, inequality in (4.10) is proven.

Next we prove boundedness of $\mathbb{E}\left[\|\boldsymbol{\gamma}(t+1,\mathbf{x}^t,\omega)\|^2\right]$ in (4.12). If the function Ψ in (4.1) satisfies condition 5' of Assumption 3.1, whether w in (4.1) has finite or infinite variance, v in (6.6) is bounded, i.e.,

$$|v|^2\leq|\Psi(a+w)|^2+|\varphi(a)|^2\leq c,$$

for some positive constant c. If the function Ψ in (4.1) satisfies condition 5 of Assumption 3.1 and w has finite variance, we get that variance of v in (6.6) is bounded with $c\,(1+|a|^2)$ for some positive constant c, i.e.,

$$\mathbb{E}[|v|^2]\leq\mathbb{E}[|\Psi(a+w)|^2+|\varphi(a)|^2]\leq\mathbb{E}[c_1(1+|a+w|^2)+c_1'(1+|a|^2)]$$
$$\leq c_1(1+|a|^2+\mathbb{E}[|w|^2])+c_1'(1+|a|^2)\leq c\,(1+|a|^2),$$

where c_1 and c_2 are some positive constants. Thus, whether condition 5 or 5' is satisfied for the function Ψ in (4.1), variance of v in (6.6) is bounded with $c\,(1+|a|^2)$ for some positive constant c. Hence, we have that for $\boldsymbol{\zeta}^t$, $\boldsymbol{\eta}^t$ defined in (4.3)

$$\mathbb{E}[\boldsymbol{\zeta}^t]\leq c'(1+V(\mathbf{x}))$$
$$\mathbb{E}[\boldsymbol{\eta}^t]\leq c''(1+V(\mathbf{x})),$$

for all $t=0,1,...$, where c' and c'' are some positive constants.

C. Mutually dependent observation noise and mutually dependent communication noise. In this subsection we relax assumptions on observation and communication noises and show that Theorems 4.2 and 4.3 continue to hold. We let Assumptions 1–6 still hold except those which overlap with the following generalizations:

- The observation noise $\mathbf{n}^t$ has the joint probability density function p_{o} such that:

$$\int\limits_{\mathbf{a}\in\mathbb{R}^N}\|\mathbf{a}\|p_{\mathrm{o}}(\mathbf{a})d\mathbf{a}<\infty,\qquad\int\limits_{\mathbf{a}\in\mathbb{R}^N}\mathbf{a}\,p_{\mathrm{o}}(\mathbf{a})d\mathbf{a},$$

  and $p_{\mathrm{o}}(\mathbf{a})=p_{\mathrm{o}}(-\mathbf{a})$, for all $\mathbf{a}\in\mathbb{R}^N$.

- A (possibly) different nonlinear function $\Psi_{\mathrm{o},i} : \mathbb{R} \to \mathbb{R}$ is assigned to each agent i. Each function $\Psi_{\mathrm{o},i}$ obeys Assumption 3.1.
- The communication noise $\boldsymbol{\xi}_{ij}^t$ has the joint probability density function $p_{\mathrm{c},ij}$ such that:
$$\int_{\mathbf{a}\in\mathbb{R}^M} \|\mathbf{a}\| p_{\mathrm{c},ij}(\mathbf{a}) d\mathbf{a} < \infty, \qquad \int_{\mathbf{a}\in\mathbb{R}^M} \mathbf{a}\, p_{\mathrm{c},ij}(\mathbf{a}) d\mathbf{a} = 0,$$
and $p_{\mathrm{c},ij}(\mathbf{a}) = p_{\mathrm{c},ij}(-\mathbf{a})$, for all $\mathbf{a} \in \mathbb{R}^M$.
- A different nonlinear function $\Psi_{\mathrm{c},ij,\ell} : \mathbb{R} \to \mathbb{R}$ is assigned to each arc $(i, j) \in E_d$ and to each element $\ell = 1, ..., M$ of the communication noise $[\boldsymbol{\xi}_{ij}^t]_\ell$. Each function $\Psi_{\mathrm{c},ij,\ell}$ obeys Assumption 3.1.

This means that observation noises of agents i and j can be mutually dependent. Moreover, the communication noises $\boldsymbol{\xi}_{ij}^t$ may have mutually dependent elements $[\boldsymbol{\xi}_{ij}^t]_\ell$, for $\ell = 1, ..., M$. Further, here, for simplicity, we assume that observation and communication noises are mutually independent.
Let us define functions $\varphi_{\mathrm{o},i} : \mathbb{R} \to \mathbb{R}$ for $i = 1, 2, ..., N$ and $\varphi_{ij,\ell} : \mathbb{R} \to \mathbb{R}$ for $(i, j) \in E$ and $\ell = 1, 2, ..., M$ in the same manner as in (4.1), i.e.,

$$\varphi_{\mathrm{o},i}(a) = \int \Psi_{\mathrm{o},i}(a + w) p_{\mathrm{o},i}(w) dw, \tag{6.9}$$

$$\varphi_{\mathrm{c},ij,\ell}(a) = \int \Psi_{\mathrm{c},ij,\ell}(a + w) p_{\mathrm{c},ij,\ell}(w) dw. \tag{6.10}$$

Here, $p_{\mathrm{o},i}$ and $p_{\mathrm{c},ij,\ell}$ are the marginal probability density functions of random variables $\mathbf{n}_i^t$ and $[\boldsymbol{\xi}_{ij}^t]_\ell$, respectively. Following same steps as in the proofs of Theorems 4.2 and 4.3, almost sure convergence and asymptotic normality can be shown. In the following, we emphasize only differences. First of all, algorithm (4.4) gets replaced by

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a} \hat{\mathbf{L}}_{\boldsymbol{\varphi}_\mathrm{c}}(\mathbf{x}^t) - \mathbf{H}^\top \boldsymbol{\varphi}_\mathrm{o} \left(\mathbf{H} \left((\mathbf{1}_N \otimes \boldsymbol{\theta}^*) - \mathbf{x}^t \right) \right) - \mathbf{H}^\top \boldsymbol{\zeta}^t + \frac{b}{a} \boldsymbol{\eta}^t \right).$$

Now, the map $\hat{\mathbf{L}}_{\boldsymbol{\varphi}_\mathrm{c}} : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is

$$\hat{\mathbf{L}}_{\boldsymbol{\varphi}_\mathrm{c}}(\mathbf{x}) = \begin{bmatrix} \vdots \\ \sum_{j\in\Omega_i} \boldsymbol{\varphi}_{\mathrm{c},ij}(\mathbf{x}_i - \mathbf{x}_j) \\ \vdots \end{bmatrix},$$

for any $\mathbf{x} \in \mathbb{R}^{MN}$, where for all $(i, j) \in E$, function $\boldsymbol{\varphi}_{\mathrm{c},ij} : \mathbb{R}^M \to \mathbb{R}^M$ is given with $\boldsymbol{\varphi}_{\mathrm{c},ij}(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M) = [\varphi_{\mathrm{c},ij,1}(\mathbf{y}_1), \varphi_{\mathrm{c},ij,2}(\mathbf{y}_2), ..., \varphi_{\mathrm{c},ij,M}(\mathbf{y}_M)]^\top$, for $\mathbf{y} \in \mathbb{R}^M$, functions $\varphi_{\mathrm{c},ij,\ell}(a)$ for $(i, j) \in E$ and $\ell = 1, 2, ..., M$ are given by (6.10). Moreover, for $\mathbf{y} \in \mathbb{R}^N$, the map $\boldsymbol{\varphi}_\mathrm{o} : \mathbb{R}^N \to \mathbb{R}^N$ is now given with $\boldsymbol{\varphi}_\mathrm{o}(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N) = [\varphi_{\mathrm{o},1}(\mathbf{y}_1), \varphi_{\mathrm{o},2}(\mathbf{y}_2), ..., \varphi_{\mathrm{o},N}(\mathbf{y}_N)]^\top$. Using the same notation, sequences $\boldsymbol{\zeta}^t \in \mathbb{R}^N$ and $\boldsymbol{\eta}^t \in \mathbb{R}^{MN}$ are appropriate versions of the sequences defined in (4.3). If we define quantities $\hat{\mathbf{r}}(\mathbf{x})$ and $\hat{\boldsymbol{\gamma}}(t + 1, \mathbf{x}, \omega)$ as follows

$$\hat{\mathbf{r}}(\mathbf{x}) = -\frac{b}{a} \hat{\mathbf{L}}_{\boldsymbol{\varphi}_\mathrm{c}}(\mathbf{x}) - \mathbf{H}^\top \boldsymbol{\varphi}_\mathrm{o} \left(\mathbf{H} \left(\mathbf{x} - (\mathbf{1}_N \otimes \boldsymbol{\theta}^*) \right) \right), \tag{6.11}$$

$$\hat{\boldsymbol{\gamma}}(t + 1, \mathbf{x}, \omega) = -\frac{b}{a} \boldsymbol{\eta}^t + \mathbf{H}^\top \boldsymbol{\zeta}^t, \tag{6.12}$$

it is easy to see that all conditions B1–B5 and C1–C5 from Theorem 6.1 still hold (see [15]). The only difference occurs in the asymptotic covariance matrix $\mathbf{S}$, i.e., in $\mathbf{S}_0$, which is now given by

$$\mathbf{S}_0 = \frac{b^2}{a^2} \mathbf{K}_{\boldsymbol{\eta}} + \mathbf{H}^\top \mathbf{K}_{\boldsymbol{\zeta}} \mathbf{H},$$

where $\mathbf{K}_{\boldsymbol{\eta}} \in \mathbb{R}^{N\times N}$ and $\mathbf{K}_{\boldsymbol{\zeta}} \in \mathbb{R}^{MN\times MN}$ are the effective covariance matrices of communication and observation noises after passing through the appropriate nonlinearities (analogously defined as cross-covariance matrix $\mathbf{K}_{\mathrm{c,o}}$ in Theorem 4.3).

D. Heavy-tailed noise and identity function. In this subsection, we show that the algorithm (3.1) does not converge in the presence of heavy-tailed observation and communication noise if at least one of the nonlinearities Ψ_{o} and Ψ_{c} is the identity function. This means that in the presence of heavy-tailed observation and communication noises, the algorithms from [15, 18] do not converge, in fact, they exhibit an infinite variance solution sequence.

THEOREM 6.3 (Infinite variance). *For the sequence of iterates* $\{\mathbf{x}^t\}, t = 1, 2, ...,$ *generated by* (3.1)*, we have that* $\mathbb{E}[\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^\star\|^2] = \infty, t = 1, 2, ...,$ *if at least one of the following statements is true.*

1. *Function* Ψ_{o} *is the identity function, i.e.,* $\Psi_{\mathrm{o}}(a) = a$ *and the observation noise has infinite variance, i.e.,* $\int a^2 d\Phi_{\mathrm{o}} = +\infty$.
2. *Function* Ψ_{c} *is the identity function, i.e.,* $\Psi_{\mathrm{c}}(a) = a$ *and the communication noise has infinite variance, i.e.,* $\int a^2 d\Phi_{\mathrm{c}} = +\infty$.

Proof. For simplicity, we assume that if statement 1 holds there is no communication noise, i.e. $\boldsymbol{\xi}_{ij} \equiv 0$ for all $(i, j) \in E_d$, and *vice versa*, if statement 2 holds we assume that there is no observation noise, i.e., $\mathbf{n} \equiv 0$. If statement 1 holds, in the absence of communication noise, the algorithm (3.2) can be written as

$$\begin{aligned}\mathbf{x}^{t+1} &= \mathbf{x}^t - \alpha_t \left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x}) - \mathbf{H}^\top\left(\mathbf{z}^t - \mathbf{H}\mathbf{x}^t\right)\right)\\ &= \mathbf{x}^t - \alpha_t \left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x}) - \mathbf{H}^\top\left(\mathbf{H}(\mathbf{1}\otimes\boldsymbol{\theta}^\star) + \mathbf{n}^t - \mathbf{H}\mathbf{x}^t\right)\right).\end{aligned}$$

If we define $\mathbf{e}^t = \mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^\star$, $t = 1, 2, ...,$ we have that $\mathbf{e}^{t+1} = \mathbf{F}^t(\mathbf{e}^t) + \alpha_t\mathbf{H}^\top\mathbf{n}^t$, where function $\mathbf{F}^t : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is given by $\mathbf{F}^t(\mathbf{y}) = (\mathbf{I} + \alpha_t\mathbf{H}^\top\mathbf{H})\mathbf{y} - \alpha_t\frac{b}{a}\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{y})$, for $\mathbf{y} \in \mathbb{R}^{MN}$. Therefore, we have that

$$\begin{aligned}\|\mathbf{e}^{t+1}\|^2 &= \|\mathbf{F}^t(\mathbf{e}^t)\|^2 + 2\alpha_t(\mathbf{F}^t(\mathbf{e}^t))^\top\mathbf{H}^\top\mathbf{n}^t + \alpha_t^2\|\mathbf{H}^\top\mathbf{n}^t\|^2\\ &\geq 2\alpha_t(\mathbf{H}\,\mathbf{F}^t(\mathbf{e}^t))^\top\mathbf{n}^t + \alpha_t^2\|\mathbf{H}^\top\mathbf{n}^t\|^2,\end{aligned}$$

and using the fact that $\mathbf{e}^t$ and $\mathbf{n}^t$ are independent, we have that

$$\mathbb{E}[\|\mathbf{e}^{t+1}\|^2] \geq \alpha_t^2\mathbb{E}[\|\mathbf{H}^\top\mathbf{n}^t\|^2] = \infty,$$

which completes the proof of statement 1. Proof of statement 2 follows directly from Appendix B in [15]. ∎

E. Hypothetical variant of algorithm from [1]. Firstly, we give an overview of algorithm that is proposed in [1], for more information see [1]. They considered a network of N agents where each agent $i = 1, 2, ..., N$ at each time $t \geq 0$ collects a linear transformation of unknown vector parameter $\mathbf{w}^0 \in \mathbb{R}^M$ corrupted by noise as follows

$$d_i(t) = \mathbf{u}_{i,t}\mathbf{w}^0 + v_i(t),$$

where $\mathbf{u}_{i,t} \in \mathbb{R}^M$ is a row regression vector and $v_i(t) \in \mathbb{R}$ is wide-sense stationary zero-mean impulsive noise process with variance $\sigma_{v,i}^2$. They introduced an agent-dependent and time-varying error nonlinearity, $h_{i,t}(e_i(t))$, into the adaptation step and proposed following algorithm

$$\begin{aligned}\boldsymbol{\psi}_{i,t} &= \mathbf{w}_{i,t-1} + \mu_i\mathbf{u}_{i,t}^\top h_{i,t}(e_i(t)),\\ \mathbf{w}_{i,t} &= \sum_{\ell\in\mathcal{N}_i} a_{\ell i}\boldsymbol{\psi}_{\ell,t},\end{aligned} \tag{6.13}$$

where μ_i is a step size parameter, $\mathcal{N}_i$ is the set of agents connected to agent i including himself and $a_{\ell i}$ are weighting coefficients. For the error nonlinearity $h_{i,t}(e_i(t))$, they set to be a linear combination of $B_i \geq 1$ preselected sign-preserving basis functions, i.e., $h_{i,t}(e_i(t)) = \boldsymbol{\alpha}_{i,t}^\top \boldsymbol{\varphi}_{i,t}(e_i(t))$. As it is said in [1], if agent i were to run the sand-alone counterpart of the adaptive filter in (6.13), then the optimal nonlinearity that minimizes i-th agent MSE is given by $h_{i,t}^{\text{opt}}(x) = -\frac{p_e'(x)}{p_e(x)}$ in terms of the pdf of the error signal.

Even though the pdf is not available in practice, for the purpose of comparing algorithms in the specific numerical example when we know pdf, we introduce hypothetical variant of algorithm, by finding optimal $\boldsymbol{\alpha}_{i,t}^{\text{opt}}$, for each agent i at each time t, i.e., $\boldsymbol{\alpha}_{i,t}^{\text{opt}} = \underset{\boldsymbol{\alpha}_{i,t}}{\operatorname{argmin}}\, \mathbb{E}[h_{i,t}^{\text{opt}}(e_i(t)) - h_{i,t}(e_i(t))]^2$

F. Derivations and numerical illustrations for Example 1. Derivation for the average per-agent asymptotic variance $\sigma_B^2 = \frac{1}{N}\operatorname{Tr}(\mathbf{S})$ follows

$$\sigma_B^2 = \frac{1}{N}\operatorname{Tr}(a^2 \int_0^{+\infty} e^{\Sigma v}\mathbf{S}_0 e^{\Sigma v} dv) = \frac{1}{N} a^2 \sigma_{\text{o}}^2 h^2 \int_0^{+\infty} \operatorname{Tr}(e^{2\Sigma v} dv)$$

$$= \frac{1}{N} a^2 \sigma_{\text{o}}^2 h^2 \int_0^{+\infty} N e^{(1-2ah^2\varphi_{\text{o}}'(0))v} dv = \frac{a^2\sigma_{\text{o}}^2 h^2}{2ah^2\varphi_{\text{o}}'(0) - 1}.$$

Integral in the last equality converge for $a > \frac{1}{2h^2\varphi_{\text{o}}'(0)}$.

If $a = a(B) = \frac{1}{2h^2\varphi_{\text{o}}'(0)(B)} + \epsilon$, for some constant $\epsilon > 0$, we have that

$$\sigma_B^2 = \frac{\left(\frac{1}{2h^2\varphi_{\text{o}}'(0)} + \epsilon\right)^2 \sigma_{\text{o}}^2 h^2}{2\left(\frac{1}{2h^2\varphi_{\text{o}}'(0)} + \epsilon\right) h^2\varphi_{\text{o}}'(0) - 1} = \frac{\left(\frac{1+2h^2\varphi_{\text{o}}'(0)\epsilon}{2h^2\varphi_{\text{o}}'(0)}\right)^2 \sigma_{\text{o}}^2 h^2}{2\left(\frac{1}{2h^2\varphi_{\text{o}}'(0)} + \epsilon\right) h^2\varphi_{\text{o}}'(0) - 1}$$

$$= \frac{\left(\frac{1+2h^2\varphi_{\text{o}}'(0)\epsilon}{2h^2\varphi_{\text{o}}'(0)}\right)^2 \sigma_{\text{o}}^2 h^2}{1 + 2h^2\varphi_{\text{o}}'(0)\epsilon - 1} = \frac{\left(1 + 2h^2\varphi_{\text{o}}'(0)\epsilon\right)^2 \sigma_{\text{o}}^2}{8h^4\varphi_{\text{o}}'(0)^3\epsilon}.$$

Next, we validate that $\lim_{B\to 0^+} \sigma_B^2 = +\infty$. It is suffice to show that $\lim_{B\to 0^+} \frac{\sigma_{\text{o}}^2}{\varphi_{\text{o}}'(0)^3} = +\infty$, since $\sigma_B^2 = \frac{\sigma_{\text{o}}^2}{8h^4\varphi_{\text{o}}'(0)^3\epsilon} + \frac{4h^2\epsilon\sigma_{\text{o}}^2}{8h^4\varphi_{\text{o}}'(0)^2\epsilon} + \frac{4h^4\epsilon^2\sigma_{\text{o}}^2}{8h^4\varphi_{\text{o}}'(0)\epsilon}$.

$$\lim_{B\to 0^+} \frac{\sigma_{\text{o}}^2}{\varphi_{\text{o}}'(0)^3} = \lim_{B\to 0^+} \frac{B^2 \int_{-\infty}^{+\infty} \tanh^2(\frac{w}{B}) f(w) dw}{\left(\int_{-\infty}^{+\infty} \frac{1}{\cosh^2(\frac{w}{B})} f(w) dw\right)^3} = [\frac{w}{B} = t, dw = dt]$$

$$= \lim_{B\to 0^+} \frac{B^2 \int_{-\infty}^{+\infty} \tanh^2(\frac{w}{B}) f(w) dw}{B^3 \left(\int_{-\infty}^{+\infty} \frac{1}{\cosh^2(w)} f(Bw) dw\right)^3}$$

$$= \lim_{B\to 0^+} \frac{\int_{-\infty}^{+\infty} \tanh^2(\frac{w}{B}) f(w) dw}{B \left(\int_{-\infty}^{+\infty} \frac{1}{\cosh^2(w)} f(Bw) dw\right)^3} = +\infty,$$

since $\lim\limits_{B\to 0^+} \int\limits_{-\infty}^{+\infty} \tanh^2(\frac{w}{B}) f(w) dw = 1$ and $\lim\limits_{B\to 0^+} = \int\limits_{-\infty}^{+\infty} \frac{1}{\cosh^2(w)} f(Bw) dw < +\infty$. We now prove that both of the functions σ_{o}^2 and $\varphi_{\mathrm{o}}'(0)$ are increasing function with respect to B. Suppose that $B_1 < B_2$, then we have that

$$B_1^2 \tanh^2(\frac{w}{B_1}) < B_2^2 \tanh^2(\frac{w}{B_2}),$$

$$\frac{1}{\cosh^2(\frac{w}{B_1})} < \frac{1}{\cosh^2(\frac{w}{B_2})},$$

for all $w \in \mathbb{R}$. Moreover, since $f(w) \geq 0$ for all $w \in \mathbb{R}$, we have that

$$B_1^2 \tanh^2(\frac{w}{B_1}) f(w) < B_2^2 \tanh^2(\frac{w}{B_2}) f(w),$$

$$\frac{1}{\cosh^2(\frac{w}{B_1})} f(w) < \frac{1}{\cosh^2(\frac{w}{B_2})} f(w),$$

for all $w \in \mathbb{R}$. Therefore, we have that

$$\sigma_{\mathrm{o}}^2(B_1) = \int_{-\infty}^{+\infty} B_1^2 \tanh^2(\frac{w}{B_1}) f(w) dw < \int_{-\infty}^{+\infty} B_2^2 \tanh^2(\frac{w}{B_2}) f(w) dw = \sigma_{\mathrm{o}}^2(B_2),$$

$$\varphi_{\mathrm{o}}'(0)(B_1) = \int_{-\infty}^{+\infty} \frac{1}{\cosh^2(\frac{w}{B_1})} f(w) dw < \int_{-\infty}^{+\infty} \frac{1}{\cosh^2(\frac{w}{B_2})} f(w) dw = \varphi_{\mathrm{o}}'(0)(B_2).$$

We now compare, in the presence of heavy-tailed observation noise with pdf as in (5.3) for $\beta = 2.05$, the proposed algorithm (5.2) for the optimal choice of $B^\star$ with the method from [1] and its hypothetical variant (see Appendix E). For those methods we set that $B_i = 2$, $\phi_{i,1}(x) = x$ and $\phi_{i,2}(x) = \tanh(x)$ for all agents. Furthermore, we set that weighting coefficients are chosen according to $a_{ij} = \frac{\tilde{\mathbf{A}}_{ij}}{\sum\limits_{\ell \in \mathcal{N}_i} \tilde{\mathbf{A}}_{\ell i}}$, where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. Moreover, for the smoothing recursions, zero initial conditions are assumed, ν_i is set to 0.9 for every agent i and $\epsilon = 10^{-2}$.

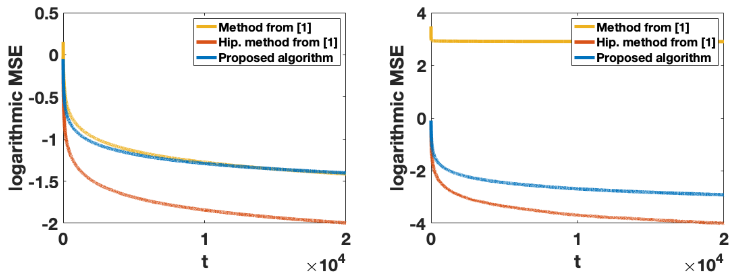
Figure 5a shows Monte Carlo estimation of MSE for step size $\alpha_t = \frac{0.5}{t+1}$ and the Figure 5b shows Monte Carlo estimation of MSE for step size $\alpha_t = \frac{1}{t+1}$. As it can be seen, the hypothetical variant of the method from [1] outperforms the proposed one in both of the scenarios. However, that is because with the hypothetical variant of [1] we optimize the choice of the nonlinearity for each agent at each time, whereas the proposed algorithm (5.2) is optimized only by average per-agent asymptotic variance. Moreover, we see that the method from [1] is not as robust as the proposed algorithm (5.2) with respect to the choice of the step size α_t (constant a).

G. Proof of the assertion in Remark 4.13. Here, we modify Theorem 3.1 from [8] and make it applicable to probability density functions that satisfy Assumption 2.2. We will show that

$$\sup_{p \in \mathcal{P}_{1+\epsilon}^M} \mathbb{P}\left(|\hat{\theta}_t - \theta^\star| > \left(\frac{8^{\frac{1}{\epsilon}} M^{\frac{2}{\epsilon}} \ln 2\delta}{t(\ln 2\delta - 1)} \right)^{\frac{\epsilon}{1+\epsilon}} \right) \geq \delta, \tag{6.14}$$

for any $\theta^\star \in \mathbb{R}$, $\delta \in (0, \frac{1}{2})$, where $\mathcal{P}_{1+\epsilon}^M \subseteq \mathcal{P}$ denotes the subclass of all pdfs from $\mathcal{P}$ such that $1+\epsilon$-central moment equals M for $\epsilon \in (0,1)$. Therefore, using Markov inequality, we get

$$\sup_{p \in \mathcal{P}_{1+\epsilon}^M} t\mathbb{E}[|\hat{\theta}_t - \theta^\star|^2] \geq c_1 t^{\frac{1-\epsilon}{1+\epsilon}},$$



(a) (b)

FIG. 5. *(a) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for the algorithm* (5.2) *for optimal $B^\star$ and for algorithm and its hypothetical variant from [1] for $a = 0.5$ (b) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for the algorithm* (5.2) *for optimal $B^\star$ and for algorithm and its hypothetical variant from [1] for $a = 1$*

for $c_1 = \delta\left(\frac{8^{\frac{1}{\epsilon}} M^{\frac{2}{\epsilon}} \ln 2\delta}{\ln 2\delta - 1}\right)^{\frac{2\epsilon}{1+\epsilon}}$. Using that $\mathcal{P}^M_{1+\epsilon} \subseteq \mathcal{P}$ and taking the supremum with respect to t we get (4.23).

To show that (6.14) holds, we follow the same idea as in [8]. Let us consider the class $\mathcal{P}_{+,-} = \{p_+, p_-\}$ of probability density function p_+ and p_- such that p_+ and p_- are probability density functions of uniform random variables on $[\frac{p^2-p}{2}, \frac{p^2+p}{2}]$ and on $[\frac{-p^2-p}{2}, \frac{p-p^2}{2}]$, respectively, for $p \in (0,1)$. It is easy to see that means of probability density functions p_+ and p_- are $\theta_+ = \frac{p^2}{2}$ and $\theta_- = -\frac{p^2}{2}$, respectively. Moreover, $1+\epsilon$-th central moment of both pdfs is equal to

$$M = \frac{p^{\epsilon+1}}{2^{\epsilon+1}(\epsilon+2)}. \tag{6.15}$$

Let $(X_j, Y_j), j = 1, 2, .., t$ be i.i.d. pairs random variables such that p_+ is pdf of X_1, and $Y_1 = X_1$ if $X_1 \in I = [\frac{p^2-p}{2}, \frac{p-p^2}{2}]$ and $Y_1 = -X_1$ if $X_1 \notin I$. Notice that probability density function of Y_1 is p_-. Since we have that $\mathbb{P}\{X_1 \in I\} = 1 - p$, for $X^t = (X_1, X_2, ..., X_t)$ and $Y^t = (Y_1, Y_2, ..., Y_t)$, we have that

$$\mathbb{P}\{X^t = Y^t\} = (1-p)^t.$$

Using that $1 - p \geq e^{\frac{-p}{1-p}}$, we have that $\mathbb{P}\{X^t = Y^t\} = (1-p)^t \geq 2\delta$, if $p \leq \frac{\ln 2\delta}{\ln 2\delta - t}$. Setting that $p := \frac{\ln 2\delta}{t(\ln 2\delta - 1)}$, we have that $p \in (0,1)$ for all $t = 1, 2, ...$ and $\delta \in (0, \frac{1}{2})$. Let $\hat{\theta}_t = \hat{\theta}_t(\cdot)$ be any estimator, then we have that

$$\begin{aligned}
&\max\left(\mathbb{P}\Big\{|\hat{\theta}_t(X^t) - \theta_+| > \frac{p^2}{2}\Big\}, \mathbb{P}\Big\{|\hat{\theta}_t(Y^t) - \theta_-| > \frac{p^2}{2}\Big\}\right) \\
&\geq \frac{1}{2}\mathbb{P}\Big\{|\hat{\theta}_t(X^t) - \theta_+| > \frac{p^2}{2} \ \ \textit{or} \ \ |\hat{\theta}_t(Y^t) - \theta_-| > \frac{p^2}{2}\Big\} \\
&\geq \frac{1}{2}\mathbb{P}\{\hat{\theta}_t(X^t) = \hat{\theta}_t(Y^t)\} \\
&\geq \frac{1}{2}\mathbb{P}\{X^t = Y^t\} \geq \delta.
\end{aligned}$$

Finally, using (6.15) we get that $\frac{\left(\frac{p^2}{2}\right)^{\frac{\epsilon+1}{2}}}{2\sqrt{2}} \geq \frac{\left(\frac{p^2}{2}\right)^{\frac{\epsilon+1}{2}}}{2^{\frac{\epsilon+1}{2}}(\epsilon+1)} = M \geq Mp^{\frac{\epsilon}{2}}$, which gives us that $\frac{p^2}{2} \geq \left(8^{\frac{1}{\epsilon}} M^{\frac{2}{\epsilon}} p\right)^{\frac{\epsilon}{\epsilon+1}}$ and therefore we have that

$$\max\left(\mathbb{P}\left\{|\hat{\theta}_t(X^t) - \theta_+| > \left(\frac{8^{\frac{1}{\epsilon}} M^{\frac{2}{\epsilon}} \ln 2\delta}{t(\ln 2\delta - 1)}\right)^{\frac{\epsilon}{1+\epsilon}}\right\},\right.$$

$$\left.\mathbb{P}\left\{|\hat{\theta}_t(Y^t) - \theta_-| > \left(\frac{8^{\frac{1}{\epsilon}} M^{\frac{2}{\epsilon}} \ln 2\delta}{t(\ln 2\delta - 1)}\right)^{\frac{\epsilon}{1+\epsilon}}\right\}\right) \geq \delta.$$

Since we have that $\mathcal{P}_{+,-} \subseteq \mathcal{P}_{1+\epsilon}^M$, it follows that (6.14) also holds.

H. Proof of extensions in Remark 4.14. For compact notation, we set that $\overline{\mathbf{H}}$ and $\widetilde{\mathbf{H}}^t$ are the $N \times (MN)$ matrices whose i-th row vectors are equal to $[\mathbf{0}, ..., \mathbf{0}, (\overline{\mathbf{h}}_i)^\top, \mathbf{0}, ..., \mathbf{0}]$ and $[\mathbf{0}, ..., \mathbf{0}, (\widetilde{\mathbf{h}}_i^t)^\top, \mathbf{0}, ..., \mathbf{0}]$, respectively. Hence, for $\mathbf{H}^t = \overline{\mathbf{H}}^t + \widetilde{\mathbf{H}}^t$, we have that (4.24) can be written, in compact form, as

$$\mathbf{z}^t = \mathbf{H}^t(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \mathbf{n}^t = \overline{\mathbf{H}}(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \widetilde{\mathbf{H}}^t(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \mathbf{n}^t. \tag{6.16}$$

Under this setting, we modify algorithm (3.1) such that, at each time $t = 0, 1, ...,$, each agent i updates its estimate $\mathbf{x}_i^t$ according to

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha_t\left(\frac{b}{a}\sum_{j\in\Omega_i}\boldsymbol{\Psi}_{\mathrm{c}}\left(\mathbf{x}_i^t - \mathbf{x}_j^t + \boldsymbol{\xi}_{ij}^t\right) - \overline{\mathbf{h}}_i\Psi_{\mathrm{o}}\left(z_i^t - \overline{\mathbf{h}}_i^\top\mathbf{x}_i^t\right)\right). \tag{6.17}$$

Assuming that all Assumptions 2.1-3.1 still hold (except those which overlap and are hence replaced with assumptions in Remark 4.14), we show that the results in subsections 4.2, 4.3 and 4.4 continue to hold for algorithm (6.17). Following the same idea as in Section 4, we write algorithm (6.17), in compact form, by:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t\left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x}) - \overline{\mathbf{H}}^\top\boldsymbol{\Psi}_{\mathrm{o}}\left(\mathbf{z}^t - \overline{\mathbf{H}}\mathbf{x}^t\right)\right). \tag{6.18}$$

Substituting (6.16) into (6.18), we get that

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t\left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x}) - \overline{\mathbf{H}}^\top\boldsymbol{\Psi}_{\mathrm{o}}\left(\overline{\mathbf{H}}(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \widetilde{\mathbf{H}}^t(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \mathbf{n}^t - \overline{\mathbf{H}}\mathbf{x}^t\right)\right)$$

$$= \mathbf{x}^t - \alpha_t\left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\Psi}_{\mathrm{c}}}(\mathbf{x}) - \overline{\mathbf{H}}^\top\boldsymbol{\Psi}_{\mathrm{o}}\left(\overline{\mathbf{H}}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* - \mathbf{x}^t\right) + \widetilde{\mathbf{H}}^t(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \mathbf{n}^t\right)\right).$$

Recalling $\boldsymbol{\eta}^t \in \mathbb{R}^{MN}$ from (4.3) and defining $\boldsymbol{\zeta}^t \in \mathbb{R}^N$ by

$$\boldsymbol{\zeta}^t = \boldsymbol{\Psi}_{\mathrm{o}}\left(\overline{\mathbf{H}}\left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* - \mathbf{x}^t\right) + \widetilde{\mathbf{H}}^t(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \mathbf{n}^t\right) - \boldsymbol{\varphi}_{\mathrm{o}}\left(\overline{\mathbf{H}}\left((\mathbf{1}_N \otimes \boldsymbol{\theta}^*) - \mathbf{x}^t\right)\right)$$

algorithm (6.18) can be written by

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t\left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\varphi}_{\mathrm{c}}}(\mathbf{x}^t) - \overline{\mathbf{H}}^\top\boldsymbol{\varphi}_{\mathrm{o}}\left(\overline{\mathbf{H}}\left((\mathbf{1}_N \otimes \boldsymbol{\theta}^*) - \mathbf{x}^t\right)\right) - \overline{\mathbf{H}}^\top\boldsymbol{\zeta}^t + \frac{b}{a}\boldsymbol{\eta}^t\right), \tag{6.19}$$

Since random variable $\widetilde{\mathbf{H}}^t(\mathbf{1}_N \otimes \boldsymbol{\theta}^*) + \mathbf{n}^t$ satisfies Lemma 6.2, the rest of the proofs are same as in the Section 4.