Distributed recursive estimation under heavy-tail communication noise

Dusan Jakovetic[*], Manojlo Vukovic[°],
Dragana Bajovic[†], Anit Kumar Sahu[‡], Soummya Kar[▷]

[*] 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 Fundamental Sciences (manojlo.vukovic@uns.ac.rs)
[†] University of Novi Sad, Faculty of Technical Sciences, Department of Power, Electronic and Communication Engineering(dbajovic@uns.ac.rs)
[‡]Amazon Alexa AI (anit.sahu@gmail.com)
[▷] Department of Electrical and Computer Engineering, Carnegie Mellon University (soummyak@andrew.cmu.edu)

February 10, 2022

Abstract

We consider distributed recursive estimation of an unknown vector parameter $\boldsymbol{\theta}^* \in \mathbb{R}^M$ in the presence of impulsive communication noise. That is, we assume that inter-agent communication is subject to an additive communication noise that may have heavy-tails or is contaminated with outliers. To combat this effect, within the class of consensus+innovations distributed estimators, we introduce for the first time a nonlinearity in the consensus update. We allow for a general class of nonlinearities that subsumes, e.g., the sign function or component-wise saturation function. For the general nonlinear estimator and a general class of additive communication noises – that may have infinite moments of order higher than one – we establish almost sure (a.s.) convergence to the parameter $\boldsymbol{\theta}^*$. We further prove asymptotic normality and evaluate the corresponding asymptotic covariance. These results reveal interesting tradeoffs between the negative effect of "loss of information" due to incorporation of the nonlinearity, and the positive effect of communication noise reduction. We also demonstrate and quantify benefits of introducing the nonlinearity in high-noise (low signal-to-noise ratio) and heavy-tail communication noise regimes.

Distributed inference; distributed estimation; recursive estimation; heavy-tail noise; consensus+innovations; stochastic approximation.
93E10, 93E35, 60G35, 94A13, 62M05

1 Introduction

We consider distributed inference in networked systems, whe-re each agent in a generic network continuously (over time instances $t = 0, 1, ...,$) makes noisy linear observations of an unknown vector parameter $\boldsymbol{\theta}^* \in \mathbb{R}^M$. Each agent, at each time t, generates a local estimate of $\boldsymbol{\theta}^*$ through the so-called consensus+innovations strategy, i.e., by 1) weight-averaging its current solution estimate with those of its neighbors, and 2) assimilating its new observation.

In this paper, we are interested in consensus+innovations distributed estimation in the presence of an impulsive communication noise, e.g., when the communication noise that corresponds to inter-neighbor communications is heavy-tailed or contaminated with outliers. It is highly relevant to consider impulsive communication noise in many application scenarios. For example, edge devices in Internet of Things (IoT) systems or sensor networks can be subject to impulsive noise distributions that may not have finite moments of order higher than one, e.g., [8, 32, 13, 37, 12, 9]. In this work, we allow the communication noise to be a zero-mean random variable that may have infinite moments of order α, for any $\alpha > 1$. In particular, communication noise may have an infinite variance. To the best of our knowledge, such scenarios have not been studied in the past work, wherein communication noise in consensus+innovations inference is always assumed to have a finite moment of at least second order (finite variance). Actually, as demonstrated ahead in the paper, existing consensus+innovation estimators – that are always *linear* in the consensus update part – can fail to converge under a heavy-tail communication noise. To combat the effect of the (impulsive or high variance) communication noise, we introduce for the first time a general nonlinearity in the consensus update. More precisely, we apply a nonlinear operator (e.g., a sign function, a saturation-like function, or a sigmoid function) on the difference between an agent's current iterate and a noisy version of its neighbor's iterate, for every agent in the neighborhood set. We establish, under a general setting for the nonlinearity and the additive communication noise, almost sure (a.s.) convergence of the nonlinear estimator to the true parameter $\boldsymbol{\theta}^*$. We also prove asymptotic normality and evaluate the corresponding asymptotic covariance in terms of the underlying network topology, observation noise, communication noise, and the employed nonlinearity. The results reveal interesting interplay among these different problem dimensions. Most notably, we show that, provided that the nonlinearity has uniformly bounded outputs, the nonlinear estimator converges a.s. and achieves a finite asymptotic covariance, even when the communication noise has no finite moments of order α for any $\alpha > 1$. We then demonstrate that, in the same regime, the corresponding linear consensus+innovations estimator has an infinite asymptotic covariance. We further provide several studies in the finite communication noise variance case that highlight the regimes where employing the nonlinearity strictly improves performance of consensus+innovations estimation over linear schemes. Typically, there is a threshold on the communication noise variance above which the nonlinear scheme achieves a strictly better performance over a linear counterpart.

We now review existing literature to help us contrast our contributions with respect to existing work. There has been extensive work on consensus+innovation distributed estimation, e.g., [17, 15, 16] and related distributed estimation methods, e.g., [20, 22, 23, 27, 31, 24, 38]. For example, reference [17] derives distributed estimators for both linear and nonlinear observation models, and establishes a.s. convergence and asymptotic normality of the methods under a general setting for inter-agent communication and observation noises. Specifically, their network model accounts for random link failures and dithered quantization, which, from the analysis perspective, effectively translates into an additive communication noise. Reference [15] considers consensus+innovations distributed estimation in the presence of random link failures without quantization or additive noise and develops estimators that are asymptotically efficient, i.e., that achieve the best achievable asymptotic covariance. The authors of [16] propose adaptive asymptotically efficient estimators, wherein the innovation gains are adaptively learned during the algorithm progress. There have been several recent works that consider robust distributed estimation in the presence of impulsive *observation (sensing)* noise; see [26] for a very recent survey and the references therein. To develop robust estimators, various techniques have been utilized, including, e.g., distributed estimators based on Wilcoxon norm, e.g., [19], Huber loss, e.g., [21], and mean error minimization, e.g., [36], and novel robust variants of gradient descent [30]. Reference [1] also considers distributed recursive estimation in the presence of heavy-tail (impulsive) *sensing (observation) noise*and develops a distributed estimator that seeks the unknown parameter while at the same time identifying the optimal error nonlinearity. Reference [6] considers distributed estimation under measurement attacks. In this setting, the authors develop a consensus+innovations estimator that employs a saturation nonlinearity in the *innovations update.* References [1, 6] utilize nonlinearities in the *innovations update* to combat the *observation attacks or heavy-tail noise.* This is in contrast with the current paper that employs a general nonlinearity in the *consensus update* to combat the heavy-tail communication noise. Reference [7] (see also [35]) considers robust distributed estimation methods based on adaptive subgradient projections. They are also not concerned with combating the effect of heavy-tail inter-agent communication noise. There have also been several works on consensus+innovations and related distributed detection methods, e.g., [25, 3, 2, 14] . In particular, reference [14] considers consensus+innovations distributed detection in the presence of Gaussian additive communication noise. In summary, with respect to existing work on consensus+innovations distributed inference, we employ for the first time a general nonlinearity in the consensus update, we allow for the first time for heavy-tail additive communication noise, and establish for the considered setting strong convergence guarantees, namely a.s. convergence and asymptotic normality.

The idea of employing a nonlinearity into a "baseline" linear scheme has also been used in nonlinear versions of the standard average consensus algorithm, e.g., [18, 33, 9]. Average consensus is a distributed algorithm that compute a network-wide average of scalar values, e.g. [5, 10, 11]. In more detail, the authors of [18] introduce a trigonometric nonlinearity into a standard linear consensus dynamics and show an improved dependence of the method on initial conditions. References [33, 9] employ a general nonlinearity in the linear consensus dynamics and show that it improves the method's resilience to additive communication noise. The above works are different from ours as they focus on the average consensus problem, where the observations are given to agents beforehand; the corresponding consensus algorithms hence involve only a consensus step and not an innovation step in the iterative update rule. In contrast, we consider here the consensus+innovations framework, where new observations are assimilated at each time instant (algorithm iteration). This technically leads to a very different analysis with respect to [18, 33, 9], and to qualitatively very different results. For example, asymptotic performance of the nonlinear consensus+innovations estimators is determined by an interplay between the effects of network topology, observation noise and communication noise; observation noise is a model dimension not present in standard average consensus.

There have also been works that employ a specific nonlinearity in the consensus update within distributed optimization problems. In this context, the authors of [34] modify the linear consensus update by taking out from the averaging operation the maximal and minimal estimates among the estimates from all neighbors of an agent. Reference [4] employs the sign nonlinearity in the consensus update part for distributed consensus optimization. The works [4, 34] contrast from ours in that they employ a specific nonlinearity, while we consider a general nonlinearity class. Furthermore, these works assume deterministic functions in the corresponding distributed consensus optimization problem, that effectively translates into having the observation data available beforehand. On the other hand, we consider a streaming data scenario that corresponds to the innovations update part in the algorithm we study.

Paper organization. Section 2 describes the distributed estimation model that we consider and presents the nonlinear consensus+innovations estimator that we propose. Section 3 explains our main results on the almost sure convergence and the asymptotic normality of the proposed distributed estimator. Section 4 provides several analytical and numerical examples that demonstrate benefits of the proposed nonlinear estimator over the linear counterpart in high and heavy-tail noise regimes. Finally, Section 5 concludes the paper.

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. Further, $\mathrm{Diag}(\mathbf{a})$ is the diagonal matrix whose diagonal entries are the elements of vector $\mathbf{a}$; $\mathrm{Tr}(\mathbf{A})$ the trace of matrix $\mathbf{A}$; $\mathbf{J}$ 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 $\mathrm{sign}(a)$ the sign function, i.e., $\mathrm{sign}(a) = 1$, for $a > 0$, $\mathrm{sign}(a) = -1$, for $a < 0$, and $\mathrm{sign}(0) = 0$. 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 Model and Algorithm

Subsection 2.1 explains the network and observation models that we assume. Subsection 2.2 presents the nonlinear consensus+inno-vations distributed estimator that we propose and states the technical assumptions needed for subsequent analysis presented in Section 3.

2.1 Problem model

Consider a network of N agents (sensors). Each agent i at each time $t = 0, 1, ...,$ collects a linear transformation of the parameter of interest $\boldsymbol{\theta}^* \in \mathbb{R}^M$, corrupted by noise, as follows:

$$z_i^t = \mathbf{h}_i^\top \boldsymbol{\theta}^* + n_i^t. \tag{1}$$

Here, $z_i^t \in \mathbb{R}$ is the observation, $\mathbf{h}_i \in \mathbb{R}^M$ is the deterministic, non-zero linear transformation vector and $n_i^t \in \mathbb{R}$ is a scalar zero-mean noise. The above update in (1) can be written in a compact form as follows:

$$\mathbf{z}^t = \mathbf{H} \left(\mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) + \mathbf{n}^t. \tag{2}$$

Here, $\mathbf{z}^t = [z_1^t, z_2^t, ..., z_N^t]^\top \in \mathbb{R}^N$ is the observation vector. $\mathbf{H}$ is the $N \times (MN)$ 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 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.
The agents constitute a network $G = (V, E)$, where $V = \{1, ..., N\}$ is the set of agents, and E is the set of (undirected) inter-agent communication links (edges) $\{i, j\}$. For future reference, introduce the $N \times N$ graph Laplacian matrix $\mathbf{L}$, defined by $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ is the adjacency matrix. That is, $\mathbf{D} = \mathrm{Diag}(\{d_i\})$, where d_i is the degree (number of neighbors) of agent i, and $\mathbf{A}$ is a zero-one symmetric matrix with zero diagonal, such that, for $i \neq j$, $\mathbf{A}_{ij} = 1$ if and only if $\{i, j\} \in E$. Also, denote by Ω_i the set of neighbors of agent i (excluding i). For an undirected edge $\{i, j\} \in E$, we denote by (i, j) the arc that points from j to i, and similarly, (j, i) is the arc that points from i to j. Following this convention, the communication noise injected when agent j communicates to agent i will be indexed by subscript ij (see ahead (3)).

2.2 Proposed algorithm and technical assumptions

The agents perform an iterative consensus+innovations distributed algorithm to collaboratively estimate the unknown vector parameter $\boldsymbol{\theta}^* \in \mathbb{R}^M$ in the presence of noisy communication links. We assume that communication noise may be heavy-tailed, e.g., [8, 32, 13, 37, 12, 9]. To combat the heavy-tail communication noise, we introduce for the first time a nonlinear consensus step in consensus+innovations-type methods. More precisely, the proposed distributed estimator is as follows. At each time $t = 0, 1, ...,$ each agent i updates its estimate $\mathbf{x}_i^t \in \mathbb{R}^M$ of the parameter $\boldsymbol{\theta}^*$ in the following fashion:

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha_t \left(\frac{b}{a} \sum_{j \in \Omega_i} \Psi \left(\mathbf{x}_i^t - \mathbf{x}_j^t + \boldsymbol{\xi}_{ij}^t \right) - \mathbf{h}_i \left(z_i^t - \mathbf{h}_i^\top \mathbf{x}_i^t \right) \right). \tag{3}$$

Here, $\alpha_t = a/(t+1)$ is a step-size, $a, b > 0$ are constants, $\boldsymbol{\xi}_{ij}^t \in \mathbb{R}^M$ is a zero-mean additive communication noise that models the imperfect communication from agent j to agent i. Next, $\boldsymbol{\Psi} : \mathbb{R}^M \to \mathbb{R}^M$ is a non-linear map that operates component-wise on any vector as follows:

$$\boldsymbol{\Psi}(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_M) = [\Psi(\mathbf{y}_1), \Psi(\mathbf{y}_2), ..., \Psi(\mathbf{y}_M)]^\top,$$

where, abusing notation, $\Psi : \mathbb{R} \to \mathbb{R}$ is a component-wise non-linear function. With algorithm (3), upon reception of the noisy version of agent j's parameter estimate $\widehat{\mathbf{x}}_{ij}^t = \mathbf{x}_j^t - \boldsymbol{\xi}_{ij}^t$, agent i applies the nonlinearity $\boldsymbol{\Psi} : \mathbb{R}^M \to \mathbb{R}^M$ on the consensus contribution $\left(\mathbf{x}_i^t - \widehat{\mathbf{x}}_{ij}^t \right)$. Intuitively, the role of Ψ is to combat the communication noise effect (e.g., truncate large values) while maintaining sufficient useful information flow. When in algorithm (3) we set $\boldsymbol{\Psi} : \mathbb{R}^M \to \mathbb{R}^M$ to be the identity map, we recover the $\mathcal{LU}$ (linear estimator) in [17].

For future reference, we write algorithm (3) in compact form.

Let $\mathbf{x}^t = [\mathbf{x}_1^t, \mathbf{x}_2^t, ..., \mathbf{x}_N^t]^\top \in \mathbb{R}^{MN}$. Furthermore, for $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^\top \in \mathbb{R}^{MN}$ and $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, ..., \boldsymbol{\xi}_N]^\top \in \mathbb{R}^{MNN}$, where $\boldsymbol{\xi}_i = [\boldsymbol{\xi}_{i1}, \boldsymbol{\xi}_{i2}, ..., \boldsymbol{\xi}_{iN}]^\top \in \mathbb{R}^{MN}$ and $\boldsymbol{\xi}_{ij} = 0$ if $j \notin \Omega_i$, define $\mathbf{L}_{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\xi})$ by

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

That is, the map $\mathbf{L}_{\boldsymbol{\Psi}}(\mathbf{x}, \boldsymbol{\xi}) : \mathbb{R}^{MN} \times \mathbb{R}^{MNN} \to \mathbb{R}^{MN}$ stacks the N vectors of size M, $\sum_{j \in \Omega_i} \boldsymbol{\Psi}(\mathbf{x}_i - \mathbf{x}_j + \boldsymbol{\xi}_{ij})$, $i = 1, 2, ..., N$, one on top of another. Then, algorithm (3) can be written as:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a} \mathbf{L}_{\boldsymbol{\Psi}}(\mathbf{x}^t, \boldsymbol{\xi}^t) - \mathbf{H}^\top \left(\mathbf{z}^t - \mathbf{H}\mathbf{x}^t \right) \right), \tag{4}$$

for $t = 0, 1, ...$.

We make the following assumptions on the underlying network, non-linear map, observation noise, and communication noise. The assumed nonlinearity class is similar to that in [29].

Assumption 1. ***Network model:***
Graph $G = (V, E)$ is undirected, simple and static.

Assumption 2. ***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.*

Also, $\Psi : \mathbb{R} \to \mathbb{R}$ satisfies one of the following two properties:

5. *$|\Psi(a)| \le c_1(1 + |a|)$, for any $a \in \mathbb{R}$, for some constant $c_1 > 0$;*

5′. *$|\Psi(a)| \le c_2$, for some constant $c_2 > 0$.*

There are many interesting examples of nonlinearities that satisfy Assumption 2, including, e.g., the following:

- **(NL1)** Sign function: $\Psi(a) = \mathrm{sign}(a)$;
- **(NL2)** Saturation or clipping function: $\Psi(a) = a$, for $|a| \le m$; and $\Psi(a) = m\,\mathrm{sign}(a)$, for $|a| > m$, for some constant $m > 0$;
- **(NL3)** Relay function with insensitivity zone: $\Psi(a) = 0$, for $|a| \le r$; and $\Psi(a) = \mathrm{sign}(a)$, for $|a| > r$, for some constant $r > 0$.

Assumption 3. ***Observation model:***

1. *For each agent $i = 1, ..., N$, the observation noise sequence $\{n_i^t\}$ in* (1)*, is zero-mean and independent identically distributed (i.i.d.);*

2. *Random variables n_i^t and n_j^s are mutually independent whenever the tuple (i, t) is different from (j, s);*

3. *Random variable n_i^t has a finite variance equal to σ_{obs}^2, for any $t = 0, 1, ...$ and for any $i = 1, ..., N$;*

4. *The matrix $\sum_{i=1}^N \mathbf{h}_i \mathbf{h}_i^\top$ is invertible.*

The condition 4 in Assumption 3 is a standard global observability assumption, see. e.g. [17]; if it does not hold, then a central estimator that collects all observations according to (1) for each $t = 0, 1, ...$ and for each $i = 1, ..., N$, is not able to provide a consistent sequence of estimates over times $t = 0, 1, ...$

Assumption 4. ***Communication noise:***

1. *Additive communication noise $\{\boldsymbol{\xi}_{ij}^t\}$, $\boldsymbol{\xi}_{ij}^t \in \mathbb{R}^M$ in* (3)*, is i.i.d. in time t, independent of the observation noise family $\{n_i^t\}$, $i = 1, ..., N$, $t = 0, 1...$, and independent across different arcs (i, j) of graph G;*

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 cumulative distribution function Φ;*

3. *The distribution function Φ is symmetric, i.e., for all $a \in \mathbb{R}$ we have that $\Phi(a) = 1 - \Phi(-a)$, and has strictly positive second moment.*

We assume that at least one of the conditions 4. *or* 4′. *below holds.*

4. *Function Ψ is strictly increasing (from Assumption 2) and functions Φ and Ψ have a common growth point, i.e.,*

$$\Psi(a_0 + \varepsilon) \geq \Psi(a_0 - \varepsilon),$$
$$[\Phi_{ij}]_l(a_0 + \varepsilon) \geq [\Phi_{ij}]_l(a_0 - \varepsilon),$$

for some $a_0 \in \mathbb{R}$ and all $\varepsilon > 0$;

4′. *Distribution Φ has a pdf $p(u)$, $p : \mathbb{R} \to \mathbb{R}$, that is strictly unimodal, i.e., there holds $p(0) < +\infty$ and $p(u_1) < p(u_2)$ for $|u_1| > |u_2|$;*

5. *There holds that $\int |a| d\Phi(a) < \infty$, and the communication noise is zero-mean, i.e., $\int a d\Phi(a) = 0$;*

6. *If part* 5 *of Assumption 2 holds, then we additionally require that communication noise has a finite variance, i.e.:*
$$\int a^2 d\Phi(a) < \infty;$$

7. *Distribution Φ has a well-defined pdf $p : \mathbb{R} \to \mathbb{R}$ in the vicinity of discontinuity points of function $\Psi : \mathbb{R} \to \mathbb{R}$ from Assumption 2.*

For notational simplicity and a clearer presentation, we assume that the communication noise has the same distribution Φ across all arcs (i, j) such that $\{i, j\} \in E$. We additionally assume that each element of communication noise vector $[\boldsymbol{\xi}_{ij}^t]_\ell$, $\ell = 1, 2, ..., M$, has the same cumulative distribution function Ψ, and that $[\boldsymbol{\xi}_{ij}^t]_\ell$ and $[\boldsymbol{\xi}_{ij}^t]_s$ are mutually independent for $\ell \neq s$. Extensions to heterogeneous choices of nonlinearity Ψ across links and heterogeneous communication noises with mutually dependent $[\boldsymbol{\xi}_{ij}^t]_\ell$ and $[\boldsymbol{\xi}_{ij}^t]_s$ for $\ell \neq s$, are presented in Remark 1 in Section 3.1 (see also Supplementary material C). Similarly, we assume that the observation noise has the same variance across all agents i; analogous extensions to different agents' observation noise variances can be performed as well.

3 Main results

Subsection 3.1 states and proves almost sure convergence of the proposed nonlinear consensus+innovations distributed estimator in (3). Subsection 3.2 establishes asymptotic normality of the estimator and evaluates the corresponding asymptotic variance.

3.1 Almost sure convergence

We have the following Theorem.

Theorem 1 (Almost sure convergence). *Let Assumptions 1-4 hold. Then, for each agent $i = 1, ..., N$, the sequence of iterates $\{\mathbf{x}_i^t\}$ generated by algorithm (3) converges almost surely to the true vector parameter $\boldsymbol{\theta}^*$.*

Theorem 1 establishes, for a nonlinearity Ψ with bounded outputs (e.g., the nonlinearities NL1-3 introduced in Section 2), almost sure convergence of the proposed algorithm (3) under heavy-tail communication noise that may not have finite moments of order greater than one. In contrast, it can be shown that the corresponding linear $\mathcal{LU}$ scheme in [17] (obtained by taking Ψ to be the identity function in (3)) generates a sequence of iterates with unbounded second moments for all $t = 1, 2, ...$ (see Supplementary material B). The Theorem also establishes almost sure convergence of (3) for nonlinearities with unbounded outputs, more precisely, those that satisfy part 5 of Assumption 2, when the communication noise has finite second moment. As a special case, by taking Ψ to be the identity map, we recover for the letter case almost sure convergence of the linear estimator (the $\mathcal{LU}$ algorithm) in [17].

Setting up the proof. We next outline our strategy for proving Theorem 1. We base our analysis on stochastic approximation arguments. More precisely, we use Theorem 29 in [17] adapted from [28] (see also Theorem 3 in the supplementary material) to establish a.s. convergence of $\mathbf{x}^t$ to $\mathbf{1}_N \otimes \boldsymbol{\theta}^*$ by verifying assumptions B1–B5 of Theorem 29 in [17].

The proof strategy is as follows. We first prove a.s. convergence of algorithm (3) for the case without communication noise, i.e., by setting $\boldsymbol{\xi}_{ij}^t \equiv 0$ in (3). In this setting, we first prove the result assuming a continuous function $\Psi : \mathbb{R} \mapsto \mathbb{R}$. Then, we handle the case with discontinuous Ψ by additionally assuming that we can associate to $\Psi : \mathbb{R} \mapsto \mathbb{R}$ a "lower bound" surrogate function $\underline{\Psi} : \mathbb{R} \mapsto \mathbb{R}$ that is *continuous*, satisfies assumption 2, and the following holds:

$$|\Psi(a)| \geq |\underline{\Psi}(a)|, \text{ for any } a \in \mathbb{R}. \tag{5}$$

This enables us to complete the proof for the noiseless case. To transition to the noisy communications case, a key argument is to consider an auxiliary function $\varphi : \mathbb{R} \to \mathbb{R}$, defined by

$$\varphi(a) = \int \Psi(a + w) d\Phi(w). \tag{6}$$

Intuitively, $\varphi : \mathbb{R} \to \mathbb{R}$ is a convolution-like transformation of nonlinearity $\Psi : \mathbb{R} \to \mathbb{R}$, where the convolution is taken with respect to the communication noise cumulative distribution function Φ.

As we will demonstrate ahead, function $\varphi : \mathbb{R} \to \mathbb{R}$ in the noisy communications case effectively plays the role that function $\Psi : \mathbb{R} \to \mathbb{R}$ has in the noiseless case. Moreover, function φ inherits all the key properties of function Ψ. More precisely, we exploit the following Lemma in [29] (see Lemmas 1-6 in [29]).

Lemma 1 ([29]). *Consider function φ in (6), where function $\Psi : \mathbb{R} \to \mathbb{R}$, satisfies Assumption 2. Then, the following holds:*

1. *φ is odd;*
2. *If $|\Psi(\nu)| \leq c_1$, for any $\nu \in \mathbb{R}$, then $|\varphi(a)| \leq c_2'$, for any $a \in \mathbb{R}$, for some $c_1' > 0$;*
3. *If $|\Psi(\nu)| \leq c_2(1 + |\nu|)$, for any $\nu \in \mathbb{R}$, then $|\varphi(a)| \leq c_2'(1 + |a|)$, for any $a \in \mathbb{R}$, for some $c_2' > 0$;*
4. *$\varphi(a)$ is monotonically nondecreasing;*

5. *$\varphi(a) > 0$, for any $a > 0$.*

6. *φ is continuous at zero;*

7. *φ is differentiable at zero, with a strictly positive derivative at zero, equal to:*

$$\varphi'(0) = \sum_{i=1}^{s} \left(\Psi(\nu_i + 0) - \Psi(\nu_i - 0)\right) p(\nu_i) + \sum_{i=0}^{s} \int_{\nu_i}^{\nu_{i+1}} \Psi'(\nu) p(\nu) d\nu, \tag{7}$$

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 distribution Φ (see Assumption 2).

Lemma 1 allows that the treatment of the noisy case becomes completely analogous to the noiseless case, by replacing function Ψ with φ. Finally, to address the case when φ may not be continuous over $\mathbb{R}$, we make use of the following Lemma that is a trivial corollary of Lemma 1.

Lemma 2. *Consider φ in (6). Then, there exists a positive constant ξ such that $|\varphi(a)| \geq \frac{1}{2}\varphi'(0)\,|a|$, for $|a| \leq \xi$.*

Lemma 2 allows us to define a continuous function $\underline{\varphi} : \mathbb{R} \mapsto \mathbb{R}$,

$$\underline{\varphi}(a) = \left\{ \begin{array}{ccc} \frac{1}{2}\varphi'(0)\, a & , & |a| \leq \xi \\ \xi \, \mathrm{sign}(a) & , & \text{else} \end{array} \right. ,$$

that satisfies Assumption 2 and obeys the property:

$$|\varphi(a)| \geq |\underline{\varphi}(a)|, \text{ for any } a \in \mathbb{R}. \tag{8}$$

Function $\underline{\varphi}$ will then clearly play the role of function $\underline{\Psi}$ in (5) in the noiseless case. We are now ready to prove Theorem 1.

Proof. (Proof of Theorem 1)
Step 1: No communication noise. We start the proof by verifying conditions B1–B5 of Theorem 29 in [17] for the case without communication noise. We use the following Lyapunov function $V : \mathbb{R}^{MN} \to \mathbb{R}$, $V(\mathbf{x}) = ||\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*||^2$. For this, only for condition B3, we need to analyze separately the case with continuous Ψ and the case when Ψ may not be continuous. Also, it can be shown that (3) can be put in the form required by Theorem 29 in [17] (see also (36) in the supplementary material) by letting

$$\mathbf{r}(\mathbf{x}) = -\mathbf{H}^\top \mathbf{H} \left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right) - \frac{b}{a} \mathbf{L}_{\boldsymbol{\Psi}}(\mathbf{x}, \mathbf{0}), \tag{9}$$

$$\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega) = \mathbf{H}^\top \mathbf{n}^t, \tag{10}$$

where ω denotes an element of the underlying probability space.
Consider the filtration $\mathcal{F}_t$, $t = 1, 2, ...$, where $\mathcal{F}_t$ is the σ- algebra generated by $\{\mathbf{n}^s\}_{s=0}^{t-1}$. Denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the underlying probability space that generates random vectors $\mathbf{n}^t$, $t = 0, 1, 2, ...$, and by $\omega \in \Omega$ its arbitrary element. Clearly, for each t, function $\boldsymbol{\gamma}(t+1, \cdot, \cdot)$ is $\mathcal{B}^{MN} \otimes \mathcal{F}$ measurable, where $\mathcal{B}^{MN}$ is the Borel sigma algebra on $\mathbb{R}^{MN}$. Also, $\mathbf{r}(\cdot)$ is $\mathcal{B}^{MN}$ measurable. Hence, condition B1 holds. Further, the family of random vectors $\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)$ is $\mathcal{F}_t$ measurable, zero-mean and independent of $\mathcal{F}_{t-1}$. Thus, condition B2 holds.
We now inspect condition B3. Assume first that function $\Psi : \mathbb{R} \mapsto \mathbb{R}$ is continuous. The gradient of V equals $\nabla V(\mathbf{x}) = 2\left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right)$. Clearly, function $V(\cdot)$ is twice continuously differentiable and has uniformly bounded second order partial derivatives. We consider

$$S = \sup_{||\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*|| \in (\epsilon, 1/\epsilon)} \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x}) \rangle, . \tag{11}$$

We will show that $S < 0$, thus verifying condition B3. We have, for any $\mathbf{x} \in \mathbb{R}^{MN}$:

$$\begin{aligned}\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x}) \rangle &= -2 \left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right)^\top \left(\mathbf{H}^\top \mathbf{H} \left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) + \frac{b}{a} \mathbf{L}_{\boldsymbol{\Psi}}(\mathbf{x}) \right) \\ &= -2 \underbrace{\left(\left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right)^\top \mathbf{H}^\top \mathbf{H} \left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) \right)}_{T_1(\mathbf{x})} - 2 \frac{b}{a} \underbrace{\left(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right)^\top \mathbf{L}_{\boldsymbol{\Psi}}(\mathbf{x})}_{T_2(\mathbf{x})}. \end{aligned} \tag{12}$$

Clearly $T_1 = T_1(\mathbf{x}) \geq 0$. We will also show that $T_2 = T_2(\mathbf{x}) \geq 0$. Utilizing the fact that, $\Psi(\cdot)$ is an odd function, we have that,

$$T_2 = \sum_{\{i,j\} \in E,\, i<j} \left(\mathbf{x}_i - \mathbf{x}_j \right)^\top \boldsymbol{\Psi} \left(\mathbf{x}_i - \mathbf{x}_j \right) \geq 0, \tag{13}$$

as for $\mathbf{g} = (\mathbf{x}_i - \mathbf{x}_j)$, we have that,

$$\left(\mathbf{x}_i - \mathbf{x}_j \right)^\top \boldsymbol{\Psi} \left(\mathbf{x}_i - \mathbf{x}_j \right) = \sum_{\ell=1}^{M} \mathbf{g}_\ell \Psi \left(\mathbf{g}_\ell \right) \geq 0, \tag{14}$$

because $\mathbf{g}_\ell$ and $\Psi\left(\mathbf{g}_\ell\right)$ have the same sign, by Assumption 2. Therefore,

$$\langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x}) \rangle = -2\, T_1 - 2\, \frac{b}{a}\, T_2 \leq 0,$$

for any $\mathbf{x} \in \mathbb{R}^{MN}$.

We will further show that S in (11) is strictly less than 0. First, consider the set $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^{MN} : \|\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\| \in [\epsilon, 1/\epsilon]\}$. Note that set $\mathcal{C}$ is nonempty and compact. Clearly, we have that:

$$S \leq S_{\mathcal{C}} := \sup_{\mathbf{x} \in \mathcal{C}} \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x}) \rangle, \tag{15}$$

It is thus sufficient to show that $S_{\mathcal{C}} < 0$. Suppose the contrary is true, i.e., suppose that $S_{\mathcal{C}} = 0$. As set $\mathcal{C}$ is compact and function $\mathbf{x} \mapsto \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x}) \rangle$ is continuous, by the Weierstrass theorem, we have that $S_{\mathcal{C}} = 0$ is equivalent to having $\langle \mathbf{r}(\mathbf{x}^\bullet), \nabla V(\mathbf{x}^\bullet) \rangle = 0$, for some point $\mathbf{x}^\bullet \in \mathcal{C}$. In this case, $\mathbf{x}^\bullet$ has to be of the form, $\mathbf{x}^\bullet = \mathbf{1}_N \otimes \mathbf{m}$, where $\mathbf{m} \in \mathbb{R}^M$. As otherwise, we would have that, T_2 is strictly positive. But then, we have, $T_1 = \left(\left(\mathbf{x}^\bullet - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right)^\top \mathbf{H}^\top \mathbf{H} \left(\mathbf{x}^\bullet - \mathbf{1}_N \otimes \boldsymbol{\theta}^* \right) \right) = \left(\mathbf{m} - \boldsymbol{\theta}^\star \right)^\top \left(\sum_{i=1}^{N} \mathbf{h}_i \mathbf{h}_i^\top \right) \left(\mathbf{m} - \boldsymbol{\theta}^\star \right) > 0$, which is a contradiction in view of (11). Hence, we conclude that, for a continuous function Ψ, it holds that $S < 0$, and that condition B3 holds, i.e.,

$$\sup_{\|\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\| \in (\epsilon, 1/\epsilon)} \langle \mathbf{r}(\mathbf{x}), \nabla V(\mathbf{x}) \rangle < 0.$$

Now, we verify condition B3 for function Ψ that is not continuous but to which we can associate function $\underline{\Psi}$ that obeys Assumption 2 and for which condition (5) holds. Then, the verification of condition B3 follows analogously to the case with continuous Ψ by replacing T_2 in (13) with the following lower bound of T_2

$$\underline{T}_2 = \sum_{\{i,j\} \in E,\, i<j} \left(\mathbf{x}_i - \mathbf{x}_j \right)^\top \underline{\boldsymbol{\Psi}} \left(\mathbf{x}_i - \mathbf{x}_j \right), \tag{16}$$

where $\underline{\boldsymbol{\Psi}}(\mathbf{a}) = [\underline{\Psi}(\mathbf{a}_1), \underline{\Psi}(\mathbf{a}_2), ..., \underline{\Psi}(\mathbf{a}_M)]^\top$. Hence, condition B3 is verified.

We next verify condition B4. Recalling the definition of $\mathbf{r}(\mathbf{x})$ in (9), we have,

$$\|\mathbf{r}(\mathbf{x})\|^2 \leq c_3 V(\mathbf{x}) + c_4 \|\boldsymbol{\Psi}(\mathbf{x})\|^2, \tag{17}$$

where $c_3 = 2a^2 \left\|\mathbf{H}^\top \mathbf{H}\right\|^2$ and $c_4 = 2b^2 \left\|\mathbf{L}\right\|^2$.
We also have that,

$$\|\mathbf{\Psi}(\mathbf{x})\| \le c_5 \sum_{\{i,j\}\in E} \left(|\mathbf{x}_i - \boldsymbol{\theta}^*| + |\mathbf{x}_j - \boldsymbol{\theta}^*|\right) + c_6 \le c_7 \left\|\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\right\| + c_6,$$

for some positive constants c_5, c_6, c_7. Therefore, we have

$$\left\|\mathbf{\Psi}(\mathbf{x})\right\|^2 \le 2c_7 V(\mathbf{x}) + 2c_8^2, \tag{18}$$

for some positive constant c_8.
Thus, we have that,

$$\left\|\mathbf{r}(\mathbf{x})\right\|^2 \le c_9 V(\mathbf{x}) + c_{10},$$

form some positive constants c_9, c_{10}. Recall $\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)$ in (10). Using the boundedness of the second moment of the observation noise, we finally have that,

$$\left\|\mathbf{r}(\mathbf{x})\right\|^2 + \mathbb{E}\left[\left\|\boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)\right\|^2\right] \le c_{11}\left(V(\mathbf{x}) + 1\right),$$

for some positive constant c_{11}. Hence, condition B4 is satisfied. Finally, condition B5 clearly holds. Therefore, we conclude that $\mathbf{x}^t \to \mathbf{1}_N \otimes \boldsymbol{\theta}^*$, almost surely.
Step 2: The case with communication noise. We proceed by considering algorithm (3) under communication noise.
We clarify the steps needed to transition from the noiseless to the noisy case. If we write

$$\mathbf{\Psi}(\mathbf{x}_i^t - \mathbf{x}_j^t + \boldsymbol{\xi}_{ij}^t) = \boldsymbol{\varphi}(\mathbf{x}_i^t - \mathbf{x}_j^t) + \boldsymbol{\eta}_{ij}^t,$$

where $\boldsymbol{\eta}_{ij}^t = \left[\mathbf{\Psi}(\mathbf{x}_i^t - \mathbf{x}_j^t + \boldsymbol{\xi}_{ij}^t) - \boldsymbol{\varphi}(\mathbf{x}_i^t - \mathbf{x}_j^t)\right]$ and $\boldsymbol{\varphi} : \mathbb{R}^M \to \mathbb{R}^M$ is component-wise map defined as $\boldsymbol{\varphi}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_M) = \left[\varphi(\mathbf{x}_1), \varphi(\mathbf{x}_2), ..., \varphi(\mathbf{x}_M)\right]^\top$. We will see that quantity $\boldsymbol{\eta}_{ij}^t$ is a key ingredient of $\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)$ in Theorem 29 in [17] (see also Theorem 3 in the supplementary material).
The algorithm (3) can be written in compact form:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a}\mathbf{L}_{\boldsymbol{\varphi}}(\mathbf{x}^t) - \mathbf{H}^T(\mathbf{z}^t - \mathbf{H}\mathbf{x}^t) + \frac{b}{a}\boldsymbol{\eta}^t\right). \tag{19}$$

Here,

$$\mathbf{L}_{\boldsymbol{\varphi}}(\mathbf{x}^t) = \begin{bmatrix} \vdots \\ \sum\limits_{j\in\Omega_i} \boldsymbol{\varphi}(\mathbf{x}_i^t - \mathbf{x}_j^t) \\ \vdots \end{bmatrix} \in \mathbb{R}^{MN}, \quad \boldsymbol{\eta}^t = \begin{bmatrix} \vdots \\ \sum\limits_{j\in\Omega_i} \boldsymbol{\eta}_{ij}^t \\ \vdots \end{bmatrix} \in \mathbb{R}^{MN}, \tag{20}$$

where the $M \times 1$ blocks $\sum\limits_{j\in\Omega_i} \boldsymbol{\varphi}(\mathbf{x}_i^t - \mathbf{x}_j^t)$ and $\sum\limits_{j\in\Omega_i} \boldsymbol{\eta}_{ij}^t$ are stacked one on top of another for $j = 1, ..., N$.
The differences of (19) with respect to the case without additive communication noise are that $\mathbf{L}_{\boldsymbol{\varphi}}$ replaces $\mathbf{L}_{\Psi}$ and the term $\frac{b}{a}\,\alpha_t\,\boldsymbol{\eta}^t$ is added.
We define the Lyapunov function $V : \mathbb{R}^{MN} \to \mathbb{R}$, and quantities $\mathbf{r}_{\boldsymbol{\varphi}}(x)$ and $\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t, \mathbf{x}, \omega)$ as follows:

$$V(\mathbf{x}) = ||\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*||^2 \tag{21}$$

$$\mathbf{r}_{\boldsymbol{\varphi}}(\mathbf{x}) = -\mathbf{H}^T\mathbf{H}(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*) - \frac{b}{a}\mathbf{L}_{\boldsymbol{\varphi}}(\mathbf{x}), \tag{22}$$

$$\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1, \mathbf{x}, \omega) = \mathbf{H}^\top \mathbf{n}^t - \frac{b}{a}\boldsymbol{\eta}^t, \tag{23}$$

Now, make the following identification with respect to the transition from the noiseless to the noisy case. Quantity $\mathbf{H}^\top \mathbf{n}^t$ in the noiseless case is replaced with quantity $\mathbf{H}^\top \mathbf{n}^t - \frac{b}{a}\boldsymbol{\eta}^t$ in the noisy case. The map $\mathbf{L}_{\boldsymbol{\Psi}}(\cdot, \mathbf{0}) : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ in (4) is replaced with the map $\mathbf{L}_{\boldsymbol{\varphi}} : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ given in (20).
The proof proceeds analogously by again verifying Assumptions B1–B5. We only clarify the differences in verifying these conditions with respect to the noiseless case.
The filtration $\mathcal{F}_t$ is replaced with the filtration $\mathcal{G}_t$, $t = 1, 2, ...,$, which is generated not only by $\{\mathbf{n}^s\}_{s=0}^{t-1}$ but also by $\{\boldsymbol{\xi}_{ij}^s\}_{s=0}^{t-1}$ for $(i,j) \in E$. Clearly, for each t, function $\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1; \cdot; \cdot)$ is $\mathcal{B}^{MN} \otimes \mathcal{F}$ measurable. Also, $\mathbf{r}_{\boldsymbol{\varphi}}(\cdot)$ is $\mathcal{B}^{MN}$ measurable. Hence, condition B1 holds. Further, the family of random vectors $\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1, \mathbf{x}, \omega)$ is $\mathcal{F}_t$ measurable, zero-mean and independent of $\mathcal{F}_{t-1}$. Thus, condition B2 holds. As function φ is odd, non-decreasing, strictly positive for its positive arguments, and has a positive derivative at zero by Lemma 1, condition B3 is derived analogously to the noiseless case. Conditions B4 and B5 hold analogously to the noiseless case. Thus, the result is verified. □

Remark 1: Theorem 1 continues to hold under the following generalizations:

- A different nonlinear function $\Psi_{ij,\ell} : \mathbb{R} \to \mathbb{R}$ is assigned to each arc (i,j) and to each element $\ell = 1, ..., M$ of the communication noise $[\boldsymbol{\xi}_{ij}^t]_\ell$. Each function $\Psi_{ij,\ell}$ obeys Assumption 2.
- The observation noise $\sigma^2_{obs,i}$ is different for each agent $i = 1, 2, ..., N$.
- The communication noise $\boldsymbol{\xi}_{ij}^t$ has the joint cumulative distribution function $\boldsymbol{\Phi}_{ij}$ such that:

$$\int\limits_{\mathbf{a}\in\mathbb{R}^M} \|\mathbf{a}\| d\boldsymbol{\Phi}_{ij}(\mathbf{a}) < \infty, \quad \int\limits_{\mathbf{a}\in\mathbb{R}^M} \mathbf{a} d\boldsymbol{\Phi}_{ij}(\mathbf{a}) = 0,$$

  and $\boldsymbol{\Phi}_{ij}(\mathbf{a}) = 1 - \boldsymbol{\Phi}_{ij}(-\mathbf{a})$, for all $\mathbf{a} \in \mathbb{R}^M$.

All the remaining assumptions in 1-4 continue to hold.
Note that the above means that the communication noise $\boldsymbol{\xi}_{ij}^t$ may have mutually dependent elements $[\boldsymbol{\xi}_{ij}^t]_\ell$, for $\ell = 1, ..., M$.
For the above generalization, it can be shown that Theorem 1 continues to hold (see Supplementary material C).

3.2 Asymptotic normality

We now present our results on asymptotic normality of estimator (3).

Theorem 2 (Asymptotic normality). *Let Assumptions* $1-4$ *hold. Consider algorithm* (3) *with step-size* $\alpha_t = a/(t+1)$, $t = 0, 1, ...,$ $a > 0$. *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:*

$$\mathbf{S} = a^2 \int\limits_0^\infty e^{\boldsymbol{\Sigma} v} \mathbf{S}_0 e^{\boldsymbol{\Sigma}^\top v} dv. \tag{24}$$

Here, $\mathbf{S}_0 = \sigma^2_{obs} \mathbf{H}^\top \mathbf{H} + \frac{b^2}{a^2} \sigma^2 \operatorname{Diag}(\{d_i \mathbf{I}_M\})$, *where we recall that* d_i *is the degree of agent* i; $\sigma^2 = \int |\Psi(w)|^2 d\Phi(w)$ *is the effective communication noise variance after passing through the nonlinearity* Ψ; *we recall the observation matrix* $\mathbf{H}$ *in* (2); *the observation noise variance* σ^2_{obs} *in* (1); *function* ϕ *in* (6); *and* $\Sigma = \frac{1}{2}\mathbf{I} - a(\mathbf{H}^\top \mathbf{H} + \frac{b}{a}\varphi'(0)(\mathbf{L} \otimes \mathbf{I}_M))$, *where* a *is taken large enough such that matrix* $\boldsymbol{\Sigma}$ *is stable (i.e., real parts of* $\boldsymbol{\Sigma}$*'s eigenvalues are negative).*

Theorem 2 shows that, for the communication noise with finite variance and unbounded nonlinearities that satisfy part 5 of Assumption 2, the variance with the proposed nonlinear estimator (3) decays (in the weak convergence sense) at (the best achievable) rate $O(1/t)$. In particular, by taking Ψ to be the identity function, we recover the asymptotic normality result in [17] of the corresponding linear estimator (the $\mathcal{LU}$ scheme in [17]). Note also that the asymptotic variance expression in (24) for the indentity function $\Psi(a) = a$ coincides with that in [17] for the $\mathcal{LU}$ scheme.

Theorem 2 further demonstrates that, even under a heavy-tailed communication noise (with unbounded variance) with a bounded nonlinearity (e.g., nonlinearities NL1-3 in Section 2), the variance with algorithm (3) still decays at rate $O(1/t)$. In contrast, the corresponding linear scheme (obtained by taking Ψ in (3) to be the identity function) generates a sequence with unbounded variances for each $t = 1, 2, ...$ More precisely, we then have that $E[||\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*||^2] = \infty$, for any $t = 1, 2, ...$ (see Supplementary material C).

Theorem 2 explicitly quantifies asymptotic variance of (3). This also allows, in the finite communication noise regime, to compare the nonlinear versus the linear scheme (when both schemes achieve a finite asymptotic variance). See Subsection 4.1 for details.

Theorem 2 also reveals an interesting tradeoff when including the nonlinearity Ψ into the consensus update. On the one hand, nonlinearity makes a beneficial effect in that the communication noise plays the role only through the effective variance $\sigma^2 = \int |\Psi(w)|^2 d\Phi(w)$. In contrast, with the linear scheme, σ^2 is replaced with $\int w^2 d\Phi(w)$ that is infinite under a heavy tail setting. On the other hand, the nonlinearity Ψ makes a negative effect in that it "reduces quality" of matrix $\mathbf{\Sigma}$ through the quantity $\varphi'(0)$ that is typically less than one with a nonlinear scheme and equal to one with the linear scheme. Clearly, the tradeoff goes in favor of the nonlinear scheme in the heavy tail setting (finite variance with the nonlinear estimator versus infinite variance with the linear estimator). In the finite communication noise variance setting, the nonlinear scheme typically improves performance under a sufficiently low communication signal to noise ratio (SNR); see also Subsection 4.1. We are now ready to prove Theorem 2.

Proof. (Proof of Theorem 2) We establish asymptotic normality by verifying assumptions C1-C5 of Theorem 29 in [17] (see also Theorem 3 in the supplementary material). Firstly, we show that condition C1 hold. Since function φ is differentiable at zero, we have that

$$\varphi(a) = \varphi(0) + \varphi'(0)a + \Delta(a) = \varphi'(0)a + \Delta(a), \tag{25}$$

where for the function $\Delta : \mathbb{R} \to \mathbb{R}$, we have that $\lim\limits_{a\to 0} \frac{\Delta(a)}{a} = 0$. Hence, the function $\mathbf{r}_{\boldsymbol{\varphi}}(\mathbf{x})$ admits representation as in Theorem 29 of [17] (see also (37) of Theorem 3 in the supplementary material), with matrix

$$\mathbf{B} = -\mathbf{H}^T\mathbf{H} - \frac{b}{a}\varphi'(0)\big[\mathbf{L} \otimes \mathbf{I}_M\big],$$

and function $\boldsymbol{\delta} : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$, given with $\boldsymbol{\delta}(\mathbf{x}) = -\frac{b}{a}\mathbf{L}_{\boldsymbol{\Delta}}(\mathbf{x})$. Here, function $\mathbf{L}_{\boldsymbol{\Delta}}(\mathbf{x}) : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is defined by

$$\mathbf{L}_{\boldsymbol{\Delta}}(\mathbf{x}) = \begin{bmatrix} \vdots \\ \sum\limits_{j\in\Omega_i} \boldsymbol{\Delta}(\mathbf{x}_i - \mathbf{x}_j) \\ \vdots \end{bmatrix},$$

where function $\boldsymbol{\Delta} : \mathbb{R}^M \to \mathbb{R}^M$ is defined by (25) $\boldsymbol{\Delta}(\mathbf{y}_1, \mathbf{y}_1, ..., \mathbf{y}_M) = [\Delta(\mathbf{y}_1), \Delta(\mathbf{y}_2), ..., \Delta(\mathbf{y}_M)]^\top$, $\mathbf{y} \in \mathbb{R}^M$. Condition C2 trivially holds, if we use that $\alpha_t = \frac{a}{t+1}$. Furthermore, $\mathbf{\Sigma} = a\mathbf{B} + \frac{1}{2}\mathbf{I}$ is stable if a is large enough, because matrix $-\mathbf{B}$ is positive definite (see [17]). Thus, condition C3 also holds.

For $\mathbf{A}(t, \mathbf{x}) = \mathbb{E}\big[\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1, \mathbf{x}, \omega)\boldsymbol{\gamma}_{\boldsymbol{\varphi}}{}^\top(t+1, \mathbf{x}, \omega)\big]$, using the Lebesgue's dominated convergence theorem, it can be shown that

$$\lim_{t\to\infty, \mathbf{x}\to\boldsymbol{\theta}^*} \mathbf{A}(t, \mathbf{x}) = \sigma^2_{\mathrm{obs}}\mathbf{H}^\top\mathbf{H} + \sigma^2\,\mathrm{Diag}\left(\{d_i\,\mathbf{I}_M\}\right).$$

Therefore, condition C4 also holds. It remains to verify condition C5. Recall quantity $\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1,\mathbf{x},\omega)$ in (23). Note that this condition is equivalent to saying that the family of random variables $\{\|\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1,\mathbf{x},\omega)\|^2\}_{t=0,1,...,\,\|\mathbf{x}-\boldsymbol{\theta}^\star\|<\epsilon}$ is uniformly integrable. If the condition 5 in Assumption 2 holds (the case with finite communication noise variance and the nonlinearity with unbounded outputs), then:

$$\|\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1,x,\omega)\|^2 \quad \leq \quad c_{12}+c_{13}\|\mathbf{n}^t\|^2+c_{14}\,\|\boldsymbol{\eta}^t\|^2, \tag{26}$$

for some positive constants c_{12},c_{13},c_{14}.
Consider the family $\{\widetilde{\mathbf{g}}(t+1,\mathbf{x},\omega)\}_{t=0,1,...,\|\mathbf{x}-\boldsymbol{\theta}^\star\|<\epsilon}$, with

$$\widetilde{\mathbf{g}}(t+1,\mathbf{x},\omega)=c_{12}+c_{13}\|\mathbf{n}^t\|^2+c_{14}\,\|\boldsymbol{\eta}^t\|^2. \tag{27}$$

Clearly, $\widetilde{\mathbf{g}}(t+1,x,\omega)$ is integrable, for any $t=0,1,...$, for any $\epsilon>0$, due to the finite second moment of sensing and observation noises. The family $\{\widetilde{\mathbf{g}}(t+1,x,\omega)\}_{t=0,1,...,\|\mathbf{x}-\boldsymbol{\theta}^\star\|<\epsilon}$ is i.i.d. and hence it is uniformly integrable. The family $\{\|\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1,x,\omega)\|^2\}_{t=0,1,...,\|\mathbf{x}-\boldsymbol{\theta}^\star\|<\epsilon}$ is dominated by $\{\widetilde{\mathbf{g}}(t+1,x,\omega)\}_{t=0,1,...,\|\mathbf{x}-\boldsymbol{\theta}^\star\|<\epsilon}$ that is uniformly integrable, and hence $\{\|\boldsymbol{\gamma}_{\boldsymbol{\varphi}}(t+1,x,\omega)\|^2\}_{t=0,1,...,\|x-x^\star\|<\epsilon}$ is also uniformly integrable. An analogous argument can be applied if condition $5'$ in Assumption 2 holds (bounded nonlinearity, communication noise with infinite variance). Hence, condition C5 holds; thus, the result. □

4 Analytical and numerical examples

Subsection 4.1 provides analytical examples, and Subsection 4.2 provides simulation examples, that illustrate the main results presented in Section 3.

4.1 Analytical examples

We provide several analytical examples that illustrate Theorem 2. The examples demonstrate that, in the considered setting, the proposed nonlinear method in (3) achieves a lower asymptotic variance than the corresponding linear scheme, for a low SNR regime, i.e., for the case when the communication noise variance is above a threshold. We also consider optimization of the nonlinearity Ψ for a given nonlinearity class; more precisely, for the given analytical example, we consider optimization of parameter B for the NL2 nonlinearity class in Section 2.
Example 1: We follow a setup similar to [17], but we consider the nonlinear consensus+innovations scheme in (3), with the non-linear operator $\Psi:\mathbb{R}\rightarrow\mathbb{R}$ of the following form (the NL2 nonlinearity):

$$\Psi(w)=\left\{\begin{array}{ccc} w & , & |w|\leq B\\ +B & , & w>B\\ -B & , & w<B\end{array}\right. , \tag{28}$$

for some parameter $B>0$. Notice that letting $B\rightarrow\infty$ in (28) leads to the linear consensus+innovations $\mathcal{LU}$ scheme in [17].
Each agent i observes a scalar parameter $\theta^*\in\mathbb{R}$ according to:

$$z_i(t)=h\theta^*+n_i^t,$$

where $h\neq 0$ and n_i^t is i.i.d. in time and across sensors with variance σ^2_{obs} and zero mean. Communication noise is i.i.d. across arcs and in time and is independent of $\{n_i^t\}$, for all $i=1,2,...,N$. Assume that the communication noise has a probability distribution function $f(w)$ that is strictly positive in the vicinity of zero. Denote the eigenvalues of $\mathbf{L}$ by $0=\lambda_1<\lambda_2\leq\cdots\leq\lambda_N$. Let the graph be regular, for simplicity, with degree d. Using Theorem 2, we have that the asymptotic covariance matrix equals:

$$\mathbf{S}=a^2\int\limits_0^\infty e^{\boldsymbol{\Sigma}v}\mathbf{S}_0e^{\boldsymbol{\Sigma}v}dv.$$

Here, $\mathbf{S}_0 = \left(h^2\sigma_{\text{obs}}^2 + \frac{b^2}{a^2}d\sigma^2\right)\mathbf{I}$; also, recall $\sigma^2 = \int\limits_{-\infty}^{\infty} |\Psi(w)|^2 f(w)dw$, the effective communication noise per link. We assume that $f(w)$ has a zero mean and variance $\sigma_{\text{comm}}^2 = \int\limits_{-\infty}^{\infty} w^2 f(w)dw$ that is finite. Also,

$$\mathbf{\Sigma} = \frac{1}{2}\mathbf{I} - a\left(h^2\mathbf{I} + \frac{b}{a}\varphi'(0)\mathbf{L}\right),$$

where φ is given in (6). For the nonlinearity considered here, we have that

$$\sigma^2 = 2\int\limits_0^{+B} w^2 f(w)dw + B^2\left(1 - 2\int_0^{+B} f(w)dw\right),$$

$$\varphi'(0) = 2\int\limits_0^{+B} f(w)dw.$$

Denote by $\sigma_B^2 = \frac{1}{N}\operatorname{Tr}(\mathbf{S})$ the average per-agent asymptotic variance. Analogously to (76)-(86) in [17], for $a > \frac{1}{2h^2}$ we get:

$$\sigma_B^2 = \frac{a^2h^2\sigma_{\text{obs}}^2 + b^2 d\sigma^2}{N\left(2ah^2 - 1\right)} + \frac{a^2h^2\sigma_{\text{obs}}^2 + b^2 d\sigma^2}{N}\sum_{i=2}^{N}\frac{1}{2b\lambda_i\varphi'(0) + \left(2ah^2 - 1\right)}$$

We next analyze the values of σ_B^2 as $B \to 0$ and $B \to +\infty$.
For $B \to 0$, we have that $\sigma^2 \to 0$, $\varphi'(0) \to 0$ and

$$\sigma_B^2 \to \frac{a^2h^2\sigma_{\text{obs}}^2}{2ah^2 - 1} =: \sigma_0^2.$$

That is, when $B \to 0$, we effectively have the case that each agent is working in isolation, hence not seeing the effect of the communication noise.
For $B \to +\infty$, we have that $\varphi'(0) \to 1$, $\sigma^2 \to \sigma_{\text{comm}}^2$ and

$$\sigma_B^2 \to \frac{a^2h^2\sigma_{\text{obs}}^2 + b^2 d\sigma_{\text{comm}}^2}{N\left(2ah^2 - 1\right)} + \frac{a^2h^2\sigma_{\text{obs}}^2 + b^2 d\sigma_{\text{comm}}^2}{N}\sum_{i=2}^{N}\frac{1}{2b\lambda_i + \left(2ah^2 - 1\right)} =: \sigma_\infty^2.$$

This is the asymptotic variance of the linear $\mathcal{LU}$ scheme in [17]. Note that, for any set of values of system parameters and any $a > \frac{1}{2h^2}$ and $b > 0$, there holds that

$$\sigma_\infty^2 > \sigma_0^2 \tag{29}$$

for a sufficiently large σ_{comm}^2.
Assume from now on that (29) holds. It can be shown that there exists an optimal B, i.e., there exists B^* such that $B^* \in (0, +\infty)$ and $\inf\limits_{B\in(0,+\infty)} \sigma_B^2 = \sigma_{B^*}^2$ (see Supplementary material D).
Note that the above analysis generalizes also to the case when

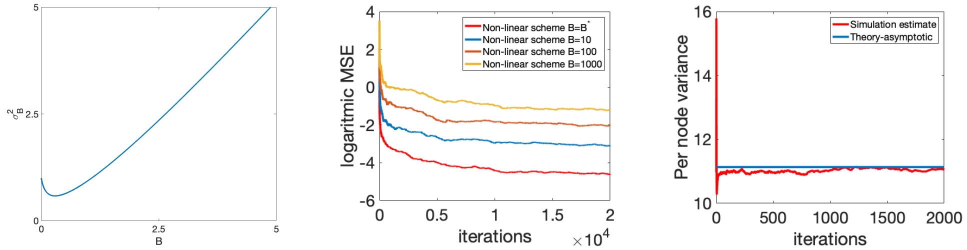
$$\sigma_{\text{comm}}^2 = \int\limits_{-\infty}^{+\infty} w^2 f(w)dw = +\infty,$$

i.e., when the noise variance is $+\infty$. In this case, we have that $\sigma_\infty^2 = +\infty$, for the linear scheme and $\sigma_0^2 = \frac{a^2h^2\sigma_{\text{obs}}^2}{2ah^2-1}$ for the isolation scheme. It can be shown that $\inf\limits_{B\in(0,+\infty)} \sigma_B^2$ is achieved at some $B^* \in (0, +\infty)$ (see Supplementary material D).

In order to demonstrate the results above, we minimize σ_B^2 and calculate B^* for a specific numerical example (see Figure 1a). We consider a sensor (agents) network with $N = 8$ agents, where the underlying topology is given by a regular graph with degree $d = 3$. We set innovation and consensus constants as $a = b = 1$, the observation parameter $h = 1$, and the true parameter $\theta^* = 1$. The observation noise for each sensor's measurements is standard normal, and the communication noise for each communication link has the following pdf

$$f(w) = \frac{\beta - 1}{2\,(1 + |w|)^{\beta}}, \tag{30}$$

with $\beta = 2.05$. (This pdf's distribution has the infinite variance.) Figure 1b shows performance of the nonlinear consensus+innovations estimator (3) in terms of the estimated per-sensor mean squared error (MSE) across iterations, for the optimal B^* and for some sub-optimal choices of B, obtained through a Monte Carlo simulation. We can see that the scheme with B^* performs better than for the considered sub-optimal choices of B. Figure 1c shows that Monte Carlo estimate of the per-agent asymptotic variance, i.e., $\hat{S} = \frac{1}{N}\|\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*\|^2\, t$ matches well the corresponding theoretical value as per Theorem 3.4.



(a) (b) (c)

Figure 1: (a) Per-agent asymptotic variance σ_B^2 versus B for the nonlinear consensus+ innovations estimator and the NL2 nonlinearity. (b) Monte Carlo-estimated per-sensor MSE error on logarithmic scale for the nonlinear consensus+innovations estimator with the NL2 nonlinearity for different choices of B. (c) Monte Carlo estimate of the per-agent asymptotic variance, and the corresponding theoretical value as per Theorem 2.

Example 2: We consider the same network and sensing models as in Example 1 and the heavy-tail communication noise distribution in (30). Furthermore, we assume that $\Psi(w) = \text{sign}(w)$ (the NL3 nonlinearity). For the $\mathcal{LU}$ scheme, it can be shown that (see Supplementary material E):

$$\sigma^2 = \sigma^2_{\text{comm}} = \frac{2}{(\beta-3)(\beta-2)},$$
$$\varphi'(0) = 1.$$

It can be shown here that the average per-agent asymptotic variance $\sigma_{\text{L}}^2 = \frac{1}{N}\,\text{Tr}(\mathbf{S})$ for the $\mathcal{LU}$ scheme is equal to

$$\sigma_{\text{L}}^2 = \begin{cases} \infty & , \quad 2 < \beta \leq 3, \\ \frac{a^2h^2\sigma^2_{\text{obs}} + b^2 d\sigma^2}{N(2ah^2-1)} + \frac{a^2h^2\sigma^2_{\text{obs}} + b^2 d\sigma^2}{N}\sum_{i=2}^{N}\frac{1}{2b\lambda_i + (2ah^2-1)} & , \quad \beta > 3. \end{cases} \tag{31}$$

For $\beta > 3$, quantity σ_{L}^2 can be written as

$$\sigma_{\text{L}}^2 = A_{\text{L}} + B_{\text{L}}\frac{1}{(\beta-3)(\beta-2)}, \tag{32}$$

where

$$A_{\mathrm{L}} = \frac{a^2h^2\sigma^2_{\mathrm{obs}}}{N(2ah^2-1)} + \frac{a^2h^2\sigma^2_{\mathrm{obs}}}{N}\sum_{i=2}^{N}\frac{1}{2b\lambda_i + (2ah^2-1)},$$
$$B_{\mathrm{L}} = 2\left(\frac{b^2d}{N(2ah^2-1)} + \frac{b^2d}{N}\sum_{i=2}^{N}\frac{1}{2b\lambda_i + (2ah^2-1)}\right).$$

We next consider the nonlinear consensus+innovations scheme with the nonlinearity $\Psi(w) = \operatorname{sign} w$. We have that

$$\sigma^2 = 1,$$
$$\varphi(a) = 2\int_0^a f(w)dw,$$

which means that $\varphi'(a) = 2f(a)$ and $\varphi'(0) = 2f(0) = (\beta - 1)$. Hence, we have that the average per-agent asymptotic variance for the nonlinear scheme $\sigma^2_{\mathrm{NL}} = \frac{1}{N}\operatorname{Tr}(S)$ is given by:

$$\sigma^2_{\mathrm{NL}} = \frac{a^2h^2\sigma^2_{\mathrm{obs}} + b^2d\sigma^2}{N\left(2ah^2-1\right)} + \frac{a^2h^2\sigma^2_{\mathrm{obs}} + b^2d\sigma^2}{N}\sum_{i=2}^{N}\frac{1}{4b\lambda_i f(0) + (2ah^2-1)}, \tag{33}$$

which can be written in the form

$$\sigma^2_{\mathrm{NL}} = A_{\mathrm{NL}} + B_{\mathrm{NL}}\frac{P_{N-2}(\beta)}{\prod\limits_{i=2}^{N}(\beta-\beta_i)}, \tag{34}$$

where

$$A_{\mathrm{NL}} = \frac{a^2h^2\sigma^2_{\mathrm{obs}} + b^2d}{N\left(2ah^2-1\right)},$$
$$B_{\mathrm{NL}} = \frac{a^2h^2\sigma^2_{\mathrm{obs}} + b^2d}{N\prod\limits_{i=2}^{N}2b\lambda_i},$$
$$P_{N-2}(\beta) = \sum_{i=2}^{N}\prod_{\substack{j=2\\ j\neq i}}^{N}2b\lambda_j(\beta-\beta_j).$$
$$\beta_i = 1 - \frac{2ah^2-1}{2b\lambda_i}, \quad i = 2, ..., N.$$

We next compare the average per-agent asymptotic variances for the linear consensus+innovations scheme and the nonlinear consensus+innovations scheme. From (31) it is obvious that $\sigma^2_{\mathrm{NL}} < \sigma^2_{\mathrm{L}}$ for $\beta \in (2,3]$. For $\beta > 3$, if $A_{\mathrm{L}} \gg A_{\mathrm{NL}}$ (see Supplementary material E), the linear scheme is worse than the nonlinear scheme for all $\beta > 3$. It is obvious that σ^2_{L} decreases on interval $(3,\infty)$ and σ^2_{NL} decreases on the interval (β_m,∞), where $\beta_m = \max\limits_{i=2,...,N}\beta_i < 1$ is closest β_i to 1. Function $\sigma^2_{\mathrm{L}} = \sigma^2_{\mathrm{L}}(\beta)$ has an asymptote at $\beta = 3$, and function $\sigma^2_{\mathrm{NL}} = \sigma^2_{\mathrm{NL}}(\beta)$ at $\beta = \beta_m$, where $\beta_m < 3$, also, A_{L} and A_{NL} are horizontal asymptotes for σ^2_{L} and σ^2_{NL}, respectively. Therefore, if A_{L} is much larger than A_{NL}, σ^2_{L} is above σ^2_{NL} for all $\beta > 3$. Moreover, if $A_{\mathrm{L}} < A_{\mathrm{NL}}$ there exists $\beta^* > 3$ such that the average per-agent asymptotic variance is still better for the nonlinear than for the linear scheme for $\beta \in (2,\beta^*]$. Defining $k = \frac{\sigma^2_{\mathrm{L}}}{\sigma^2_{\mathrm{NL}}}$, it is possible to show that $k \to \infty$ as $\beta \to 3$, and $k \to \frac{A_{\mathrm{L}}}{A_{\mathrm{NL}}}$ as $\beta \to \infty$. Therefore, if $A_{\mathrm{L}} < A_{\mathrm{NL}}$, there exists β^* such that $\sigma^2_{\mathrm{NL}} < \sigma^2_{\mathrm{L}}$ for all $\beta \in (2,\beta^*)$. In other words, there exists a threshold value $\beta^* > 3$, such that the nonlinear scheme outperforms the linear scheme for the "heavy-tail regime" $\beta \in (2, \beta^*)$, and the linear scheme performs better for $\beta > \beta^*$. To summarize, in Example 2, depending on sensing and network parameters, it holds that either the nonlinear scheme outperforms the linear one for all β, or there exists a threshold value β^* such that the nonlinear scheme is better than the linear one for $\beta \in (2, \beta^*)$. Figure 2 shows the ratio $k = \frac{\sigma_{\mathrm{L}}^2}{\sigma_{\mathrm{NL}}^2}$ versus β for the same sensing and network parameters as in Example 1. As it can be seen, there exists a threshold β^*, that here approximately equals $\beta^* = 3.9$, such that $k > 1$ for $\beta \in (2, \beta^*)$. On the other hand, for $\beta > \beta^*$, the ratio becomes smaller than one, which means that for the given numerical parameters, the linear scheme performs better for $\beta > \beta^*$. This is in accordance with the analysis that we provided above.

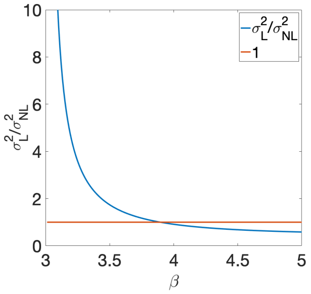


Figure 2: Ratio $k = \frac{\sigma_{\mathrm{L}}^2}{\sigma_{\mathrm{NL}}^2}$ versus β for Example 2.

4.2 Simulation examples

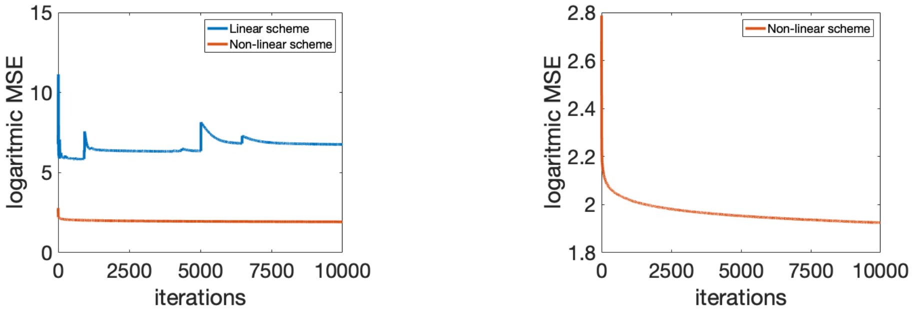
In this section, we illustrate the performance of the proposed nonlinear consensus+innovations estimator for two different choices of the non-linear operator Ψ. For both nonlinearity choices, our method is compared with the corresponding linear consensus+innovations estimator $\mathcal{LU}$ in [17], when the communication noise has probability distribution function given by (30).

We consider a sensor network with $N = 40$ agents. The underlying topology is an instance of a random geometric graph. We use the same initialization $\mathbf{x}^0 = \mathbf{0}$ and same step sizes $\alpha_t = \frac{1}{t+1}, a = 1, b = 1$, for both the linear and the nonlinear estimators. Also, we assume that the observation noise is normally distributed, i.e., $n_i^t \sim \mathcal{N}(0, 1)$, for each t, for each i. The true parameter $\boldsymbol{\theta}^* \in \mathbb{R}^{10}$ is generated randomly, where the entries of θ^* are drawn mutually independently from the uniform distribution on [-10,10]. The observation vectors $\mathbf{h}_i \in \mathbb{R}^{10}$ are also generated at random, for which the condition 4 of Assumption 3 is true. We use the communication noise pdf in (30) with $\beta = 2.05$. Note that, in this case, the communication noise has an infinite variance.

Figure 4 compares the linear $\mathcal{LU}$ estimator in [17] with the nonlinear estimator (3) with $\Psi(w)$ given in (28) for $B = 5$. Figure 3 shows the comparison between $\mathcal{LU}$ and [17] with $\Psi(w) = \mathrm{sign}(w)$. Both Figures show the iteration counter t at the x-axis and a Monte-Carlo estimate of the average mean square error (MSE) across agents on the y-axis. We can see that, as predicted by our theory, the nonlinear estimator, for both nonlinearity choices, persistently decreases MSE along iterations, despite the fact that the communication noise has an infinite variance. At the same time, $\mathcal{LU}$ fails to produce a useful estimation result.

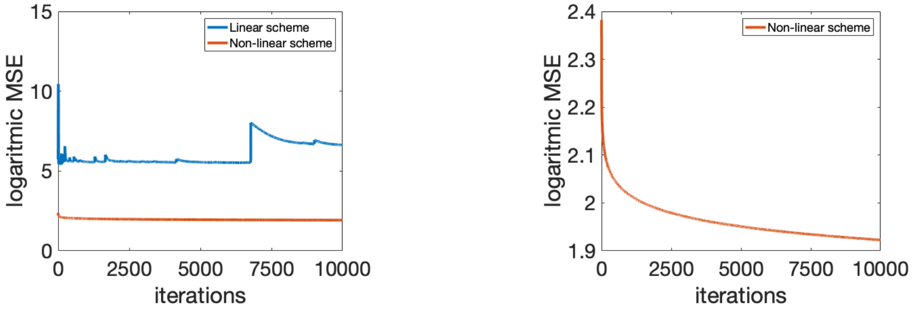
5 Conclusion

We studied consensus+innovations distributed estimation in the presence of impulsive, heavy-tail communication noise. To combat the impulsive communication noise, we introduce for the first time a general nonlinearity in the *consensus update* for consensus+innovations distributed estimation. We establish almost sure convergence of the nonlinear consensus+innovations estimator to the true parameter, prove its asymptotic normality, and explicitly evaluate the corresponding asymptotic variance. We compare the proposed



(a) (b)

Figure 3: Monte-Carlo average per-agent MSE estimate versus iteration counter on logarithmic scale for the proposed nonlinear estimator (3) with the nonlinearity in (28) for $B = 5$ and the linear $\mathcal{LU}$ scheme in [17].



(a) (b)

Figure 4: Monte-Carlo average per-agent MSE estimate versus iteration counter on logarithmic scale for the proposed nonlinear estimator (3) with the nonlinearity $\Psi(w) = \mathrm{sign}(w)$ and the linear $\mathcal{LU}$ scheme in [17].

nonlinear estimator with conventional consensus+innovation estimators that utilize linear consensus update. Analytical and numerical examples demonstrate significant gains of introducing consensus nonlinearity in low SNR (high communication noise) regimes. Most notably, we demonstrate that, when the communication noise has infinite variance, the proposed nonlinear consensus+innovations estimator is strongly consistent (converges almost surely), while the corresponding linear counterpart provides a sequence of estimators with infinite variance.

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] W. Ben-Ameur, P. Bianchi, and J. Jakubowicz, *Robust distributed consensus using total variation*, IEEE Transactions on Automatic Control, 61 (2016), pp. 1550–1564.

[5] M. Cao, A. Morse, and B. Anderson, *Reaching a consensus in a dynamically changing environment: A graphical approach*, SIAM J. Control and Optimization, 47 (2008), pp. 575–600.

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

[7] 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.

[8] 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.

[9] 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).

[10] F. Fagnani and S. Zampieri, *Average consensus with packet drop communication*, in Proceedings of the 45th IEEE Conference on Decision and Control, 2006, pp. 1007–1012.

[11] M. Huang and J. Manton, *Coordination and consensus of networked agents with noisy measurements: Stochastic algorithms and asymptotic behavior*, SIAM J. Control and Optimization, 48 (2009), pp. 134–161.

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

[13] 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.

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

[16] 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.

[17] 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.

[18] U. A. Khan, S. Kar, and J. M. F. Moura, *Distributed average consensus: Beyond the realm of linearity*, in 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, 2009, pp. 1337–1342.

[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*, 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] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar, *Robust estimation via robust gradient estimation*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82 (2020), pp. 601–627.

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

[32] B. Selim, M. S. Alam, V. Carvalho, G. Kaddoum, and B. L. Agba, *Noma-based iot networks: Impulsive noise effects and mitigation*, IEEE Communications Magazine, 58 (2020), pp. 69–75.

[33] S. Stankovic, M. Beko, and M. Stankovic, *A robust consensus seeking algorithm*, in IEEE EUROCON 2019-18th International Conference on Smart Technologies, 2019, pp. 1–6.

[34] S. Sundaram and B. Gharesifard, *Consensus-based distributed optimization with malicious nodes*, in 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015, pp. 244–249.

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

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

[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.

Supplementary material

A. Some results on Stochastic approximation

We make use of the following standard stochastic approximation result, see [28], see also [17].

Theorem 3. *Let* $\{\mathbf{x}^t \in \mathbb{R}^l\}_{t\geq 0}$ *be a random sequence:*

$$\mathbf{x}^{t+1} = \mathbf{x}^t + \alpha_t[\mathbf{r}(\mathbf{x}^t) + \boldsymbol{\gamma}(t+1, \mathbf{x}^t, \omega)], \tag{35}$$

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* $(\Omega, \mathcal{F}, \mathcal{P})$*, and* $\omega \in \Omega$ *is a canonical element. Let the following sets of assumptions hold:*

B1: *The function* $\boldsymbol{\gamma}(t, \cdot, \cdot) : \mathbb{R}^l \times \Omega \to \mathbb{R}^l$ *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* $t = 0, 1, ...,$ *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*

$$\mathbf{r}(\mathbf{x}) = \mathbf{B}(\mathbf{x} - \mathbf{x}^*) + \boldsymbol{\delta}(\mathbf{x}), \tag{36}$$

where

$$\lim_{\mathbf{x}\to\mathbf{x}^*} \frac{||\boldsymbol{\delta}(\mathbf{x})||}{||\mathbf{x} - \mathbf{x}^*||} = 0. \tag{37}$$

(Note, in particular, if $\boldsymbol{\delta}(\mathbf{x}) \equiv 0$ *then* (37) *is satisfied.)*

C2: *The weight sequence* $\{\alpha_t\}_{t\geq 0}$ *is of form*

$$\alpha_t = \frac{a}{t+1}, \forall t \geq 0, \tag{38}$$

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* (38) *and* (36)*, 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*

$$\lim_{R \to \infty} \sup_{||\mathbf{x} - \mathbf{x}^*|| < \epsilon} \sup_{t \geq 0} \int\limits_{||\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)|| > R} ||\boldsymbol{\gamma}(t+1, \mathbf{x}, \omega)||^2 dP = 0$$

Let Assumptions B1–B5 hold for $\{\mathbf{x}(t)\}_{t \geq 0}$ *in* (35). *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 Assumptions B1–B5, Assumptions C1–C5 are also satisfied. In particular, as* $t \to \infty$

$$\sqrt{t}(\mathbf{x}^t - \mathbf{x}^*) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{S}), \tag{39}$$

where $\Rightarrow$ *denotes convergence in distribution (weak convergence). Also, asymptotic variance,* $\mathbf{S}$, *in* (39) *is*

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

B. Analysis of $\mathcal{LU}$ in [17] under heavy-tail noise

We show that (3) with the identity nonlinearity Ψ, i.e., the $\mathcal{LU}$ algorithm in [17], generates a sequence of estimates $\{\mathbf{x}^t\}$, $t = 1, 2, ...$, such that $E[||\mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*||^2] = \infty$, for any $t = 1, 2, ...$ under assumptions 1-4, and assuming that communication noise has infinite variance, i.e., $\int a^2 d\Phi(a) = +\infty$.
First, note that, for $\Psi(a) \equiv a$, algorithm (3) can be compactly written as:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \left(\frac{b}{a}\left(\mathbf{L} \otimes \mathbf{I}_M\right)\right)\mathbf{x}^t + \alpha_t \mathbf{H}^\top\left(\mathbf{z}^t - \mathbf{H}x^t\right) + \alpha_t \frac{b}{a}\hat{\boldsymbol{\xi}}^t,$$

for $t = 0, 1, ...,$ where $\hat{\boldsymbol{\xi}}^t = \begin{bmatrix} \vdots \\ \sum\limits_{j \in \Omega_i} \boldsymbol{\xi}_{ij}^t \\ \vdots \end{bmatrix}$. In other words, random vector $\hat{\boldsymbol{\xi}}^t \in \mathbb{R}^{MN}$ stacks one on top of another the communication noise $\sum\limits_{j \in \Omega_i} \boldsymbol{\xi}_{ij}^t$ injected at each agent $i = 1, 2, ..., N$.
Denote by $\mathbf{e}^t = \mathbf{x}^t - \mathbf{1}_N \otimes \boldsymbol{\theta}^*$. It is easy to show that for $t = 0, 1, ...,$ we have that:

$$\mathbf{e}^{t+1} = \underbrace{\left(\mathbf{I} - \alpha_t\left(\frac{b}{a}\left(\mathbf{L} \otimes \mathbf{I}_M\right) + \mathbf{H}^\top\mathbf{H}\right)\right)}_{\mathbf{F}_t}\mathbf{e}^t + \alpha_t \underbrace{\left(\mathbf{H}^t\mathbf{H}\hat{\boldsymbol{\xi}}^t\right)}_{\boldsymbol{\mu}^t}. \tag{40}$$

Note that $\mathbf{e}^t$ is measurable and is zero-mean, for all t. It can be shown (see [17]) that matrix $\mathbf{F}_t$ is positive definite, for t large enough. From (40), we have:

$$||\mathbf{e}^{t+1}||^2 = \left(\mathbf{e}^t\right)^\top \mathbf{F}_t^\top \mathbf{F}_t \mathbf{e}^t + \alpha_t \left(\mathbf{e}^t\right)^\top \mathbf{F}_t^\top \boldsymbol{\mu}^t + \alpha_t ||\boldsymbol{\mu}^t||^2 \tag{41}$$

Taking expectation, using the fact that $\mathbf{F}_t^\top \mathbf{F}_t$ is positive definite, and using independence of $\mathbf{e}^t$ and $\boldsymbol{\mu}^t$, we get:

$$\mathbb{E}\left[||\mathbf{e}^{t+1}||^2\right] \geq \alpha \mathbb{E}\left[||\boldsymbol{\mu}^t||^2\right] = +\infty,$$

for any $t = 0, 1, ...,$ because

$$\mathbb{E}\left[||\boldsymbol{\mu}^t||^2\right] \geq \frac{b^2}{a^2} \mathbb{E}\left[||\hat{\boldsymbol{\xi}}^t||^2\right] \geq b^2 \int a^2 d\Psi(a) = +\infty.$$

C. The proof of extensions in Remark 1

Define function $\varphi_{ij,\ell} : \mathbb{R} \to \mathbb{R}$, as follows:

$$\varphi_{ij,\ell}(a) = \int \Psi_{ij,\ell}(a + w) d\Phi_{ij,\ell}(w), \tag{42}$$

where $\Phi_{ij,\ell}$ is the marginal distribution of random variable $[\boldsymbol{\xi}_{ij}^t]_\ell$.
We now provide missing arguments to establish the proof of extensions in Remark 1. All arguments follow straight forwardly by repeating steps in the proof of Theorem 1 with the following modification.
The map $\mathbf{L}_{\boldsymbol{\varphi}} : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ gets replaced by

$$\hat{\mathbf{L}}_{\boldsymbol{\varphi}}(\mathbf{x}) = \mathbb{E}\begin{bmatrix} \vdots \\ \sum_{j \in \Omega_i} \boldsymbol{\Psi}_{ij}(\mathbf{x}_i - \mathbf{x}_j + \boldsymbol{\xi}_{ij}^t) \\ \vdots \end{bmatrix},$$

for any $\mathbf{x} \in \mathbb{R}^{MN}$, where for all $(i, j) \in E$, function $\boldsymbol{\Psi}_{ij} : \mathbb{R}^{MN} \to \mathbb{R}^{MN}$ is given with

$$\boldsymbol{\Psi}_{ij}(\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_{MN}) = [\Psi_{ij,1}(\mathbf{y}_1), \Psi_{ij,2}(\mathbf{y}_2), ..., \Psi_{ij,MN}(\mathbf{y}_{MN})]^\top,$$

for $\mathbf{y} \in \mathbb{R}^{MN}$. The key is to show that (12) continues to hold, where now

$$\mathbf{r}(\mathbf{x}) = -\mathbf{H}^\top \mathbf{H}(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*) - \frac{b}{a}\hat{\mathbf{L}}_\varphi(\mathbf{x}).$$

That is, we must show that $(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*)^\top \hat{\mathbf{L}}_\varphi(\mathbf{x}) \geq 0$, for any $\mathbf{x} \in \mathbb{R}^{MN}$, and $(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*)^\top \hat{\mathbf{L}}_\varphi(\mathbf{x}) = 0$ if and only if $\mathbf{x}$ is of the form $\mathbf{x} = \mathbf{1}_N \otimes \mathbf{m}$, for some $\mathbf{m} \in \mathbb{R}^M$. We have:

$$(\mathbf{x} - \mathbf{1}_N \otimes \boldsymbol{\theta}^*)^\top \hat{\mathbf{L}}_\varphi(\mathbf{x}) = \sum_{\substack{\{i,j\} \in E \\ i<j}} \underbrace{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbb{E}\left[\boldsymbol{\Psi}_{ij}\left(\mathbf{x}_i - \mathbf{x}_j + \boldsymbol{\xi}_{ij}^t\right)\right]}_{r_{ij}}.$$

Let $\mathbf{y} = \mathbf{x}_i - \mathbf{x}_j$ and $\boldsymbol{\zeta} = \boldsymbol{\xi}_{ij}^t$ for simplicity. We have:

$$\begin{aligned} r_{ij} &= \mathbf{y}^\top \int \boldsymbol{\Psi}_{ij}(\mathbf{y} + \boldsymbol{\zeta}) d\boldsymbol{\Phi}_{ij}(\boldsymbol{\zeta}) = \sum_{\ell=1}^{M} \int \mathbf{y}_\ell \Psi_{ij,\ell}(\mathbf{y}_\ell + \boldsymbol{\zeta}_\ell) d\boldsymbol{\Phi}_{ij}(\boldsymbol{\zeta}) \\ &= \sum_{\ell=1}^{M} \int \mathbf{y}_\ell \Psi(\mathbf{y}_\ell + \boldsymbol{\zeta}_\ell) d\Phi_{ij,\ell}(\boldsymbol{\zeta}_\ell) = \sum_{\ell=1}^{M} \mathbf{y}_\ell \varphi_{ij,\ell}(\mathbf{y}_\ell), \end{aligned} \tag{43}$$

where $\varphi_{ij,\ell}$ is defined in (42). Note that, Φ_{ij} is a symmetric (multivariate) distribution and hence function $\varphi_{ij,\ell}$ obeys all properties given in Lemma 1. Hence, $\mathbf{y}_\ell$ and $\Psi_{ij,\ell}(\mathbf{y}_\ell)$ in (43) have equal sign, and the proof proceeds analogously to the proof of Theorem 1.

D. Derivations for Example 1

For the nonlinearity NL2 given with (28), we have that

$$\sigma^2 = \int\limits_{-\infty}^{\infty} |\Psi(w)|^2 f(w)dw = \int\limits_{-B}^{+B} w^2 f(w)dw + \int\limits_{\mathbb{R}\setminus[-B,+B]} B^2 f(w)dw$$
$$= 2\int\limits_{0}^{+B} w^2 f(w)dw + B^2 \left(1 - 2\int_0^{+B} f(w)dw\right),$$

where in last equality we use that $f(w)$ is symmetric. Recalling that

$$\varphi(a) = \int\limits_{-\infty}^{\infty} \Psi(a+w)f(w)dw,$$

we have that

$$\varphi'(0) = \int\limits_{-\infty}^{\infty} \Psi'(w)f(w)dw = \int\limits_{-B}^{+B} f(w)dw = 2\int\limits_{0}^{+B} f(w)dw.$$

We next calculate $\frac{\partial}{\partial B}\sigma_B^2$ for $B \to 0^+$, and we show that, $\frac{\partial}{\partial B}\sigma_B^2 < 0$ for B sufficiently close to 0^+, i.e., for $B \in (0, B_{min})$, for some positive scalar B_{min}. This means, together with (29) and with continuity of σ_B^2 with respect to B, that there exists $B^* \in (0, +\infty)$ such that $\inf\limits_{B\in(0,+\infty)} \sigma_B^2 = \sigma_{B^*}^2$, i.e., that the infimum is not at 0^+ nor at $+\infty$. This in turn shows that, under condition (29), the nonlinear scheme with $B = B^*$ outperforms the linear scheme (with $B = +\infty$) and the isolation scheme (with $B = 0$).
It remains to show that $\frac{\partial}{\partial B}\sigma_B^2 < 0$ for B sufficiently small, i.e., for $B \in (0, B_{min})$ for some $B_{min} > 0$. Rewrite σ_B^2 as:

$$\sigma_B^2 = c_0 + c_1\rho_1(B) + \sum_{i=2}^{N} \frac{c_2}{c_{4,i}\rho_2(B) + c_5} + c_3 \sum_{i=2}^{N} \frac{\rho_1(B)}{c_{4,i}\rho_2(B) + c_5}, \tag{44}$$

with

$$\begin{aligned}
\rho_1(B) &= \sigma^2(B),\\
\rho_2(B) &= \varphi'(0)(B),\\
c_0 &= \frac{a^2h^2\sigma_{\text{obs}}^2}{N\left(2ah^2 - 1\right)} > 0,\\
c_1 &= \frac{b^2 d}{N\left(2ah^2 - 1\right)} > 0,\\
c_2 &= \frac{a^2h^2\sigma_{\text{obs}}^2}{N} > 0,\\
c_3 &= \frac{b^2 d}{N} > 0,\\
c_{4,i} &= 2b\lambda_i > 0, \quad i = 2, \cdots, N,\\
c_5 &= 2ah^2 - 1 > 0.
\end{aligned} \tag{45}$$

Notice that $\rho_1(B) \to 0$ and $\rho_2(B) \to 0$ as $B \to 0^+$. Recall that we assumed that f is strictly positive in the vicinity of zero. Taking the derivative of ρ_1 with respect to B we have that

$$\begin{aligned}\frac{\partial}{\partial B}\rho_1(B) &= \frac{\partial}{\partial B}\left(2\int\limits_0^{+B} w^2 f(w)dw + B^2\left(1-2\int\limits_0^{+B} f(w)dw\right)\right)\\ &= 2B^2 f(B) + 2B\left(1-2\int\limits_0^{+B} f(w)dw\right) + B^2\left(-2f(B)\right),\end{aligned}$$

and hence $\frac{\partial}{\partial B}\rho_1(B) \to 0$ as $B \to 0^+$. From

$$\frac{\partial}{\partial B}\rho_2(B) = \frac{\partial}{\partial B}\left(2\int\limits_0^{+B} w^2 f(w)dw\right) = 2f(B).$$

we have that $\frac{\partial}{\partial B}\rho_2(B) \to 2f(0)$ as $B \to 0^+$.
We can see form (44) that

$$\begin{aligned}\frac{\partial}{\partial B}\sigma_B^2 = \frac{\partial}{\partial B}\rho_1(B) &+ \sum_{i=2}^{N}\frac{-c_2}{\left(c_{4,i}\rho_2(B)+c_5\right)^2}c_{4,i}\frac{\partial}{\partial B}\rho_2(B)\\ &+ c_3\sum_{i=2}^{N}\frac{\frac{\partial}{\partial B}\rho_1(B)(c_{4,i}\rho_2(B)+c_5) - c_{4,i}\frac{\partial}{\partial B}\rho_2(B)\rho_1(B)}{\left(c_{4,i}\rho_2(B)+c_5\right)^2}.\end{aligned}$$

Therefore, we have that

$$\frac{\partial}{\partial B}\sigma_B^2 \to -\sum_{i=2}^{N} 2c_2 c_{4,i} f(0)\frac{1}{c_5^2} < 0,$$

as $B \to 0^+$. Hence, due to continuity of the function $\frac{\partial}{\partial B}\sigma_B^2$, we have that $\frac{\partial}{\partial B}\sigma_B^2 < 0$ for B small enough, i.e. $B \in (0, B_{min})$.

E. Derivations for Example 2

For the linear consensus+innovations $\mathcal{LU}$ scheme, regardless of communication noise, we have that

$$\varphi(a) = \int\limits_{-\infty}^{\infty}(a+w)f(w)dw = a + \int\limits_{-\infty}^{\infty} wf(w)dw = a.$$

Using that pdf of communication noise is given with (30), for $\mathcal{LU}$, assuming $\beta > 3$, we have that:

$$\begin{aligned}\sigma^2 = \sigma^2_{\text{comm}} &= \int\limits_{-\infty}^{\infty} w^2 f(w)dw = 2\int\limits_0^{\infty}\frac{cw^2}{(1+w)^\beta}dw\\ &= 2c\int\limits_1^{\infty}\frac{(u-1)^2}{u^\beta}du = 2c\int\limits_1^{\infty} u^{2-\beta} - 2u^{1-\beta} + u^{-\beta}du \qquad (46)\\ &= 2c\left(\frac{u^{3-\beta}}{3-\beta} - 2\frac{u^{2-\beta}}{2-\beta} + \frac{u^{1-\beta}}{1-\beta}\right)\Bigg|_1^{\infty}\\ &= 2c(\frac{1}{\beta-3} - \frac{2}{\beta-2} + \frac{1}{\beta-1}) = \frac{2}{(\beta-3)(\beta-2)}.\end{aligned}$$

The integral in (46) does not converge for $\beta \leq 3$.
For the nonlinear consensus+innovations scheme in (3) with the nonlinearity $\Psi(w) = \mathrm{sign}(w)$ we have that

$$\begin{aligned}
\sigma^2 &= \int\limits_{-\infty}^{\infty} \mathrm{sign}^2(w) f(w) dw = 1, \\
\varphi(a) &= \int\limits_{-\infty}^{\infty} \Psi(a+w) f(w) dw = \int\limits_{-\infty}^{\infty} \mathrm{sign}(a+w) f(w) dw \\
&= \int\limits_{-\infty}^{\infty} \mathrm{sign}(u) f(u-a) du = -\int\limits_{-\infty}^{0} f(u-a) du + \int\limits_{0}^{\infty} f(u-a) du \\
&= -\int\limits_{-\infty}^{-a} f(w) dw + \int\limits_{-a}^{\infty} f(w) dw = 2\int\limits_{0}^{a} f(w) dw.
\end{aligned}$$

Considering (33), we have that for all $i = 2, \cdots, N$,

$$\begin{aligned}
4b\lambda_i f(0) + 2ah^2 - 1 &= 2b\lambda_i(\beta - 1) + \underbrace{(2ah^2 - 1)}_{>0} \\
&= 2b\lambda_i(\beta - \beta_i),
\end{aligned}$$

where $\beta_i < 1$ for all $i = 2, \cdots, N$.