Sahu *et al. EURASIP Journal on Advances in Signal Processing* (2018) 2018:66
https://doi.org/10.1186/s13634-018-0586-0

EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Communication efficient distributed weighted non-linear least squares estimation

Anit Kumar Sahu[1], Dusan Jakovetic[2*], Dragana Bajovic[3] and Soummya Kar[1]

Abstract

The paper addresses design and analysis of communication-efficient distributed algorithms for solving weighted non-linear least squares problems in multi-agent networks. *Communication efficiency* is highly relevant in modern applications like cyber-physical systems and the Internet of things, where a significant portion of the involved devices have energy constraints in terms of limited battery power. Furthermore, *non-linear models* arise frequently in such systems, e.g., with power grid state estimation. In this paper, we develop and analyze a non-linear communication-efficient distributed algorithm dubbed $\mathcal{CREDO}-\mathcal{NL}$ (non-linear $\mathcal{CREDO}$). $\mathcal{CREDO}-\mathcal{NL}$ generalizes the recently proposed linear method $\mathcal{CREDO}$ (Communication efficient REcursive Distributed estimatOr) to non-linear models. We establish for a broad class of non-linear least squares problems and generic underlying multi-agent network topologies $\mathcal{CREDO}-\mathcal{NL}$'s strong consistency. Furthermore, we demonstrate communication efficiency of the method, both theoretically and by simulation examples. For the former, we rigorously prove that $\mathcal{CREDO}-\mathcal{NL}$ achieves significantly faster mean squared error rates in terms of the elapsed communication cost over existing alternatives. For the latter, the considered simulation experiments show communication savings by at least an order of magnitude.

Keywords: Distributed estimation, Stochastic approximation, Statistical inference, Non-linear least squares

1 Introduction

We consider distributed non-linear least squares estimation in networked systems. The networked system considered consists of heterogeneous networked entities or agents where the inter-agent collaboration conforms to a pre-assigned possibly sparse communication graph. The agents acquire their local, noisy, non-linear observations about the unknown phenomenon (unknown static vector parameter $\boldsymbol{\theta}$) in a streaming fashion over discrete time instances t. The goal for each agent is to continuously generate an estimate of $\boldsymbol{\theta}$ over time instances t in a recursive fashion, where the estimate update of an agent involves simultaneous assimilation of the newly acquired local observations, and the received information through messages with agents in its immediate neighborhood. The assumed setup is highly relevant in several emerging applications in the context of cyber-physical systems (CPS) and the Internet of things (IoT), like state estimation in smart grid, predictive maintenance, and production monitoring in industrial manufacturing systems. For example, with continuous state estimation of a smart grid, the acquired measurements (voltages, angles) are in general non-linear functions of the unknown state; further, the measurements are inherently distributed across different physical locations (elements of the system), and they arrive continuously over time with a prescribed sampling rate. Furthermore, the scale (network size) of the distributed system (e.g., a large scale micro-grid) and near real-time requirements on the estimation results make distributed, fusion center-free processing a desirable choice.

An important aspect of distributed estimation algorithms in the context of the applications described above is communication efficiency, i.e., achieving good estimation performance with minimal communication cost. Real-world applications such as large-scale deployment of CPS or IoT typically involve entities or agents with limited on board energy resources. In addition to the limited on board power, the energy requirement per unit communication is usually significantly higher than the energy

*Correspondence: djakovet@uns.ac.rs
[2]University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics, 21000 Novi Sad, Serbia
Full list of author information is available at the end of the article

requirement per unit computation [48]. Hence, communication efficiency is a highly desirable trait in such systems. Moreover, for large-scale systems which require continuous system monitoring, it is crucial to reduce the communication cost as much as possible without compromising on the performance of the inference task at hand, which then ensure longer lifetime of such systems.

In this paper, we propose and analyze a communication efficient distributed estimator for non-linear observation models that we refer to as $\mathcal{CREDO}-\mathcal{NL}$. The estimator $\mathcal{CREDO}-\mathcal{NL}$ generalizes the recently proposed linear distributed estimator $\mathcal{CREDO}$, see [37, 38], that is designed and works for linear measurement (observation) models only. Specific contributions of the paper are as follows.

We propose the non-linear distributed estimator $\mathcal{CREDO}-\mathcal{NL}$ that works for a broad class of non-linear observation models and where the model information in terms of the node i's sensing function and noise statistic is only available at the individual agent i itself. With the proposed algorithm, each agent communicates probabilistically sparsely over time. More precisely, the probability which determines whether a node communicates at time t decays sub-linearly to zero with t, which then makes the communication cost scale sub-linear with time t.

Despite dropping communications and the presence of non-linearities in the sensing model, we show that the proposed algorithm achieves the optimal $O(1/t)$ rate of the mean square error (MSE) decay[1]. The achievability of the optimal MSE decay in terms of time t translates into significant improvements in the rate at which MSE scales with respect to the per-agent average communication cost $\mathcal{C}_t$ up to time t, namely from $O(1/\mathcal{C}_t)$ with existing methods, e.g., [15, 16, 31, 34–36, 40], to $O\left(1/\mathcal{C}_t^{2-\zeta}\right)$ with the proposed method, where $\zeta > 0$ is arbitrarily small. We also establish strong consistency of the estimate sequence at each agent, showing that each agent's local estimator converges almost surely to the true parameter $\boldsymbol{\theta}$. Simulation examples confirm significant communication savings of $\mathcal{CREDO}-\mathcal{NL}$ over existing alternatives, by at least an order of magnitude.

We now briefly review the literature on distributed inference and motivate our algorithm $\mathcal{CREDO}-\mathcal{NL}$. Distributed inference algorithms can be broadly divided into two classes based on the presence of a fusion center. The first class assumes presence of a fusion center, e.g., [11, 23, 26, 27, 47]. The fusion center assigns sub-tasks to the individual agents and subsequently fuses the information from different agents. However, when the data samples are geographically distributed across the individual agents and are streamed in time, fusion center-based solutions are impractical.

The second class of distributed inference methods is fusion center-free. These works typically assume that the agents are interconnected over a generic network, and each agent acquires its local measurements in a streaming fashion. These estimators are iterative (recursive), where at each iteration (time instance), each agent assimilates its new measurement and exchanges messages with its immediate neighbors, see, e.g., [2, 4–6, 14, 20, 22, 24, 25, 28–31, 34–36, 39, 43, 46]. Most related to our work are references that consider distributed estimation under non-linear observation models, as we do here, or distributed convex stochastic optimization, e.g., [15, 16, 31, 34–36, 40]. However, among these works, the best achieved *MSE communication rate* is $O(1/\mathcal{C}_t)$. In contrast, we establish here a strictly faster MSE communication rate equal to $O\left(1/\mathcal{C}_t^{2-\zeta}\right)$ ($\zeta > 0$ is arbitrarily small). Finally, it is worth noting that there exist a few distributed algorithms (without fusion node) that are also designed to achieve communication efficiency, e.g., [13, 21, 44–46]. In [46], a data censoring method is employed to save in terms of computation and communication costs. However, the communication savings in [46] is a constant proportion with respect to a vanilla method which uses all allowable communications at all times. In [21], the communication savings come at a cost of extra computations. References [13, 44, 45] also consider a different setup than we do here, namely they study distributed optimization (with no fusion center) where the data is available a priori (i.e., it is not streamed). In terms of the strategy to save communications, references [13, 21, 44, 45] consider, respectively, deterministically increasingly sparse communication, adaptive communication scheme, and selective activation of agents. These strategies are different from ours that utilizes a randomized, increasing, "sparsification" of communications.

Consensus+innovations methods, see, e.g., [16, 17, 19, 20]), are a sub-class of distributed recursive algorithms (the second class of algorithms mentioned above) that process data in a streaming fashion. With consensus+innovation methods, each node updates its estimate at each iteration two-fold: by weight-averaging its solution estimate (consensus) with the neighbors' solution estimates and by assimilating its newly acquired data sample (innovation). Therein, the consensus and innovation weights are usually time-varying and are carefully designed towards achieving optimal asymptotic performance, measured, e.g., through asymptotic covariance of the estimate sequence. Within the class of *consensus+innovations* distributed estimation algorithms (see, e.g., [18, 20]), the design of communication efficient methods has been addressed in [37], see also [38], for linear observation models, wherein a mixed time-scale stochastic approximation method dubbed $\mathcal{CREDO}$ has been proposed. We extend here $\mathcal{CREDO}$ to non-linear observation models. Technically speaking, establishing convergence and asymptotic rates of convergence for $\mathcal{CREDO}-\mathcal{NL}$

involves establishing guarantees for existence of stochastic Lyapunov functions for the estimate sequence. The update of the estimate sequence in $\mathcal{CREDO}-\mathcal{NL}$ involves a gain matrix which is in turn a function of the estimate itself. Moreover, in addition to the gain matrix being a function of the estimate, the sensing functions exhibit localized behavior in terms of smoothness and global observability in the proposed algorithm. Hence, the setup considered in this paper requires technical tools different from $\mathcal{CREDO}$, which we develop in this paper.

The rest of the paper is organized as follows. Section 2 describes the problem that we consider and gives the needed preliminaries on conventional (centralized) and distributed recursive estimation. Section 3 presents the novel $\mathcal{CREDO}-\mathcal{NL}$ algorithm that we propose, while Section 4 states our main results on the algorithm's performance. Section 5 presents the simulations experiments, and finally, we conclude in Section 7. Proofs of the main results are relegated to Appendix A.

2 Model and preliminaries

2.1 Sensing and network models

Let $\boldsymbol{\theta} \in \Theta$, where $\Theta \subset \mathbb{R}^M$ (the properties of it to be specified shortly) be an M-dimensional parameter that is to be estimated by a network of N agents. Every agent n at time index t makes a noisy observation $\mathbf{y}_n(t)$, a noisy function of $\boldsymbol{\theta}$. Formally, the observation model for the n-th agent is given by,

$$\mathbf{y}_n(t) = \mathbf{f}_n(\boldsymbol{\theta}) + \gamma_n(t), \tag{1}$$

where $\mathbf{f}_n : \mathbb{R}^M \mapsto \mathbb{R}^{M_n}$ is a non-linear sensing function, where $M_n \ll M$, $\{\mathbf{y}_n(t)\} \in \mathbb{R}^{M_n}$ is the observation sequence for the n-th agent and $\{\gamma_n(t)\}$ is a zero mean temporally independent and identically distributed (i.i.d.) noise sequence at the n-th agent with nonsingular covariance $\mathbf{R}_n$, where $\mathbf{R}_n \in \mathbb{R}^{M_n \times M_n}$. The noise processes are independent across different agents. We state an assumption on the noise processes before proceeding further. Throughout, we denote by $\|\cdot\|$ the $\mathcal{L}_2$-norm of its vector or matrix argument and by $\mathbb{E}[\,.\,]$ the expectation operator.

Assumption 1 *There exists $\epsilon_1 > 0$, such that, for all n, $\mathbb{E}\left[\|\gamma_n(t)\|^{2+\epsilon_1}\right] < \infty$.*

We remark that the main results of the paper (Theorems 4.1 and 4.2) continue to hold even if $\epsilon_1 = 0$[2]. The above assumption encompasses a general class of noise distributions in the setup.

The heterogeneity of the setup is exhibited in terms of the agent dependent sensing functions and the noise covariances at the agents. Each agent is interested in reconstructing the true underlying parameter $\boldsymbol{\theta}$. We assume an agent is aware only of its local observation model, i.e, the non-linear sensing function $\mathbf{f}_n(\cdot)$ and the associated noise covariance $\mathbf{R}_n$, and hence, it has no information about the observation matrix and noise processes of other agents.

The agents are interconnected through a communication network that we shall assume throughout the paper is modeled as an *undirected* simple connected graph $G = (V, E)$, with $V = [1 \cdots N]$ and E denoting the set of agents (nodes) and communication links, see [3]. (With the proposed $\mathcal{CREDO}-\mathcal{NL}$ method, the available links in E will be activated selectively across algorithm iterations in a probabilistic fashion, as it will be detailed in Section 3). The neighborhood of node n in graph G is

$$\Omega_n = \{l \in V \,|\, (n,l) \in E\}. \tag{2}$$

The node n has degree $d_n = |\Omega_n|$. The structure of the graph is described by the $N \times N$ adjacency matrix, $\mathbf{A} = \mathbf{A}^\top = [\mathbf{A}_{nl}]$, $\mathbf{A}_{nl} = 1$, if $(n,l) \in E$, $\mathbf{A}_{nl} = 0$, otherwise. Let $\mathbf{D} = \text{diag}\,(d_1 \cdots d_N)$. The graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is positive semidefinite, with eigenvalues ordered as $0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L})$. The eigenvector of $\mathbf{L}$ corresponding to $\lambda_1(\mathbf{L})$ is $(1/\sqrt{N})\mathbf{1}_N$. (Here, $\mathbf{1}_N$ is the N-dimensional vector with all entries equal to one.) The multiplicity of its zero eigenvalue equals the number of connected components of the network; for a connected graph, $\lambda_2(\mathbf{L}) > 0$. This second eigenvalue is the algebraic connectivity or the Fiedler value of the network (see [7] for instance).

Example: distributed static phase estimation in smart grids

Many applications within cyber physical systems and the Internet of things can be modeled as non-linear distributed estimation problems of type (1). Such class of models arises, e.g., with state estimation in power systems; therein, a phasorial representation of voltages and currents is usually utilized, wherein non-linearity in general emerges from power-flow equations [1, 33]. Here, we focus on the specific problem within the class, namely distributed static phase estimation in smart grids. We describe the model briefly and refer to, e.g., [12, 19] for more details. Here, graph G corresponds to a power grid network of $n = 1, ..., N$ generators and loads (here, a single generator or a single load is a node in the graph), while the edge set E corresponds to the set of transmission lines or interconnections. (For simplicity, even though not necessary, we assume that the physical interconnection network matches the inter-node communication network.) Assume that G is connected. The state of a node n is described by $(\mathcal{V}_n, \phi_n)$, where $\mathcal{V}_n$ is the voltage magnitude and ϕ_n is the phase angle. As commonly assumed, e.g., [12], we let the voltages $\mathcal{V}_n$ be known constants; on the other hand, angles ϕ_n are unknown ant are to be estimated. Following a standard approximation path, the real

power flow across the transmission line between nodes n and l can be expressed as, e.g., [12]:

$$\mathcal{P}_{nl}(\phi) = \mathcal{V}_n \mathcal{V}_l b_{nl} \sin(\phi_{nl}), \quad (3)$$

where ϕ is the vector that collects the unknown phase angles ϕ_n across all nodes, b_{nl} is line (n,l)'s admittance, and $\phi_{nl} = \phi_n - \phi_l$. Denote by $E_m \subset E$ the set of lines equipped with power flow measuring devices. The power flow measurement at line (n,l) is then given by:

$$y_{nl}(t) = \mathcal{P}_{nl}(\phi) + \gamma_{nl}(t) = \mathcal{V}_n \mathcal{V}_l b_{nl} \sin(\phi_{nl}) + \gamma_{nl}(t), \quad (4)$$

where $\{\gamma_{nl}(t)\}$ is the zero mean i.i.d. measurement noise with finite moment $\mathbb{E}[|\gamma_{nl}(t)|^{2+\epsilon_1}]$, for some $\epsilon_1 > 0$. Assume that each measurement $y_{nl}(t)$ is assigned to one of its incident nodes n or l. Further, let Ω'_n denote the set of all indexes l such that measurements $y_{nl}(t)$ are available at node n. Then, it becomes clear that the angle estimation problem is a special case of model (1), with the measurement vectors $\mathbf{y}_n(t) = [\, y_{nl}(t),\ l \in \Omega'_n]^\top$, $n = 1, ..., N$, noise vectors $\gamma_n(t) = [\, \gamma_{nl}(t),\ l \in \Omega'_n]^\top$, $n = 1, ..., N$, and sensing functions $\mathbf{f}_n(\boldsymbol{\phi}) = [\, \mathcal{V}_n \mathcal{V}_l b_{nl} \sin(\phi_{nl}),\ l \in \Omega'_n]^\top$. $n = 1, ..., N$. It can be shown that under reasonable assumptions on phase angle ranges (that correspond to the admissible parameter set Θ) and the smart grid network and admittances structure, the assumptions we make on the sensing model are satisfied,[3] and hence, $\mathcal{CREDO} - \mathcal{NL}$ can be effectively applied; we refer to [12, 19] for details.

2.2 Preliminaries: centralized batch and recursive weighted non-linear least squares estimation

In this subsection, we go over the preliminaries of centralized and distributed weighted non-linear least squares estimation.

Consider a networked setup with a hypothetical fusion center which has access to the samples collected at all nodes at all times. In such a setting, in lieu of the sensing model as described in (1), one of the classical algorithms that finds extensive use is the weighted non-linear least squares (WNLS) (see, for example, [15]). The applicability of WNLS to fairly generic setups which are characterized by the absence of noise statistics makes it particularly appealing in practice. We discuss properties of the WNLS estimator before proceeding further. Define the cost function $\mathcal{Q}_t$ as follows:

$$\mathcal{Q}_t(\mathbf{z}) = \sum_{s=0}^{t} \sum_{n=1}^{N} (\mathbf{y}_n(s) - \mathbf{f}_n(\mathbf{z}))^\top \mathbf{R}_n^{-1} (\mathbf{y}_n(s) - \mathbf{f}_n(\mathbf{z})). \quad (5)$$

The hypothetical fusion center in such a setting generates the estimate sequence $\{\widehat{\boldsymbol{\theta}}_t\}$ in the following way:

$$\widehat{\boldsymbol{\theta}}_t \in \operatorname{argmin}_{\mathbf{z}\in\Theta} \mathcal{Q}_t(\mathbf{z}). \quad (6)$$

The consistency and the asymptotic behavior of the estimate sequence $\{\widehat{\boldsymbol{\theta}}_t\}$ have been analyzed in the literature under the following weak assumptions:

Assumption 2 *The set Θ is compact convex subset of $\mathbb{R}^M$ with non-empty interior int(Θ) and the true (but unknown) parameter $\boldsymbol{\theta} \in$ int(Θ).*

Assumption 3 *The sensing model is globally observable, i.e., any pair $\boldsymbol{\theta}, \acute{\boldsymbol{\theta}}$ of possible parameter instances in Θ satisfies*

$$\sum_{n=1}^{N} \left\| \mathbf{f}_n(\boldsymbol{\theta}) - \mathbf{f}_n\left(\acute{\boldsymbol{\theta}}\right) \right\|^2 = 0 \quad (7)$$

if and only if $\boldsymbol{\theta} = \acute{\boldsymbol{\theta}}$.

Assumption 4 *The sensing function $\mathbf{f}_n(.)$ for each n is continuously differentiable in the interior int(Θ) of the set Θ. For each $\boldsymbol{\theta}$ in the set Θ, the (normalized) gain matrix $\Gamma_{\boldsymbol{\theta}}$ defined by*

$$\Gamma_{\boldsymbol{\theta}} = \frac{1}{N} \sum_{n=1}^{N} \nabla \mathbf{f}_n(\boldsymbol{\theta}) \mathbf{R}_n^{-1} \nabla \mathbf{f}_n^\top(\boldsymbol{\theta}), \quad (8)$$

is invertible, where $\nabla \mathbf{f}_n(\cdot) \in \mathbb{R}^{M \times M_n}$ denotes the gradient of $\mathbf{f}_n(\cdot)$.

Smoothness conditions on the sensing functions, such as the one imposed by Assumption 3, are common in statistical estimation with non-linear observations models. Note that the matrix $\Gamma_{\boldsymbol{\theta}}$ is well defined at the true value of the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta} \in$ int(Θ) and the continuous differentiability of the sensing functions holds for all $\boldsymbol{\theta} \in$ int(Θ).

The asymptotic properties of the WNLS estimator in terms of consistency and asymptotic normality are characterized by the following classical result:

Proposition 1 ([15]) *Let the parameter set Θ be compact and the sensing function $f_n(\cdot)$ be continuous on Θ for each n. Let $\mathcal{G}_t$ be an increasing sequence of σ-algebras such that $\mathcal{G}_t = \sigma\left(\left\{\left\{\mathbf{y}_n(s)\right\}_{s=0}^{t-1}\right\}_{n=1}^{N}\right)$. Further, denote by $\boldsymbol{\theta}$ the true parameter to be estimated. Then, a WNLS estimator of $\boldsymbol{\theta}$ exists, i.e., there exists an $\{\mathcal{G}_t\}$-adapted process $\left\{\widehat{\boldsymbol{\theta}}_t\right\}$ such that*

$$\widehat{\boldsymbol{\theta}}_t \in \operatorname{argmin}_{\mathbf{z}\in\Theta} \mathcal{Q}_t(\mathbf{z}),\ \forall t. \quad (9)$$

Moreover, if the model is globally observable, i.e., Assumption 3 holds, the WNLS estimate sequence $\left\{\widehat{\boldsymbol{\theta}}_t\right\}$ is consistent, i.e.,

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \widehat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}\right) = 1, \quad (10)$$

where $\mathbb{P}_{\boldsymbol{\theta}}(\cdot)$ denotes the probability operator. Additionally, if Assumption 4 holds, the parameter estimate sequence is asymptotically normal, i.e.,

$$\sqrt{t+1}\left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right) \stackrel{D}{\Longrightarrow} \mathcal{N}\left(0, \Sigma_c\right), \tag{11}$$

where

$$\Sigma_c = \left(N\Gamma_{\boldsymbol{\theta}}\right)^{-1}, \tag{12}$$

$\Gamma_{\boldsymbol{\theta}}$ is as given by (8) and $\stackrel{\mathcal{D}}{\Longrightarrow}$ refers to convergence in distribution (weak convergence).

The centralized WNLS estimator above suffers from significant communication overhead due to the inherent access to data samples across all agents at all times. Moreover, the minimization in (6) requires batch processing due to the non-sequential nature of the minimization. Recursive centralized estimators utilizing stochastic approximation type approaches have been proposed in [9, 10, 32, 41, 42], which mitigate the batch processing through the development of sequential albeit centralized estimators. However, such recursive estimators still suffer from the enormous communication overhead as the fusion center requires access to the data samples across all agents at all times and the global model information in terms of the sensing functions and the noise statistics across agents.

2.3 Preliminaries: distributed WNLS

Sequential distributed recursive schemes conforming to the *consensus + innovations* (see for example, [19] and Eq. (16) ahead) type update, where the agents' knowledge of the model is limited to themselves have been proposed in [16, 40]. In [16], so as to achieve the optimal asymptotic covariance, the global model information is made available through a carefully constructed gain matrix update, which adds additional computation complexity and communication cost. In contrast with [16, 40] introduces the trade off in terms of sub-optimality of the asymptotic covariance while using local model information at individual agents for evaluating the gain matrix and thus saving communication cost. However, both the aforementioned algorithms in [16, 40] have the number of communication scales linearly with the number of per-node sampled observations $\{\mathbf{y}_n(t)\}$. This paper builds upon the ideas of sequential distributed recursive schemes catering to non-linear observation models as proposed in [16, 40] to construct a communication efficient scheme without compromising on the performance in terms of the mean square error. That is, we aim to achieve the order optimal MSE decay rate of $\Theta(1/t)$ (see, e.g., [9]) in terms of the number of per-node processed samples, while reducing the $\Theta(t)$ communication cost which is a characteristic of previous approaches.

Before proceeding further, we briefly summarize the estimator in [40] which is referred to as the *benchmark* estimator henceforth. The overall update rule at an agent n corresponds to

$$\begin{aligned}\widehat{\mathbf{x}}_n(t+1) = \mathbf{x}_n(t) &- \underbrace{\widehat{\beta}_t \sum_{l\in\Omega_n}\left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{neighborhood consensus}} \\ &- \underbrace{\widehat{\alpha}_t \left(\nabla \mathbf{f}_n(\mathbf{x}_n(t))\right) \mathbf{R}_n^{-1}\left(\mathbf{f}_n(\mathbf{x}_n(t)) - \mathbf{y}_n(t)\right)}_{\text{local innovation}}\end{aligned} \tag{13}$$

and

$$\mathbf{x}_n(t+1) = \mathcal{P}_{\Theta}[\widehat{\mathbf{x}}_n(t+1)], \tag{14}$$

where Ω_n is the communication neighborhood of agent n (determined by the Laplacian $\mathbf{L}$); $\nabla f_n(\cdot)$ is the gradient of $\mathbf{f}_n$; $\mathcal{P}_{\Theta}[\cdot]$ the projection operator corresponding to projecting on Θ; and $\{\beta_t\}$ and $\{\alpha_t\}$ are consensus and innovation weight sequences given by

$$\widehat{\beta}_t = \frac{\widehat{\beta}_0}{(t+1)^{\delta_1}}, \ \widehat{\alpha}_t = \frac{\widehat{\alpha}_0}{t+1}, \tag{15}$$

where $\widehat{\alpha}_0, \widehat{\beta}_0 > 0, 0 < \delta_1 < 1/2 - 1/(2+\epsilon_1)$ and ϵ_1 was defined in Assumption 1. From the asymptotic normality in Theorem 2 in [40], it can be inferred that the MSE decays as $O(1/t)$.

Communication efficiency

The communication cost $\mathcal{C}_t$ is defined as the expected number of per-node communications up to iteration t. Formally, the communication cost $\mathcal{C}_t$ is given by

$$\mathcal{C}_t = \mathbb{E}\left[\sum_{s=0}^{t-1} \mathbb{I}_{\{\text{agent } n \text{ transmits at } s\}}\right], \tag{16}$$

where agent n is arbitrary (the expectation in (16) does not depend on n) and $\mathbb{I}_A$ represents the indicator of event A. The communication cost $\mathcal{C}_t$ for both the centralized WNLS estimator (where all agents transmit their samples $\mathbf{y}_n(t)$ to the fusion center at all times t) and the distributed estimators in [16, 40] is $\mathcal{C}_t = \Theta(t)$, where we note that the iteration count t is equivalent to the number of per node samples collected till time t. Technically speaking, the MSE decays as $O\left(\frac{1}{\mathcal{C}_t}\right)$.

3 $\mathcal{CREDO} - \mathcal{NL}$: a communication efficient distributed WNLS estimator

In this section, we present the $\mathcal{CREDO} - \mathcal{NL}$ estimator. $\mathcal{CREDO} - \mathcal{NL}$ is based on a carefully chosen protocol which aids in making the communications increasingly probabilistically sparse. Intuitively speaking, the communication protocol exploits the idea that with a gradual information accumulation at the agents through communications, an agent is able to accumulate sufficient information about the parameter of interest which then allows

it to drop communications increasingly often. Technically speaking, for each node n, at every time t, we introduce a binary random variable $\psi_{n,t}$, where

$$\psi_{n,t} = \begin{cases} \rho_t & \text{with probability } \zeta_t \\ 0 & \text{else,} \end{cases} \tag{17}$$

where $\psi_{n,t}$'s are independent both across time and the nodes, i.e., across t and n, respectively as well are independent from nodes' observations in (1). The random variable $\psi_{n,t}$ abstracts out the decision of the node n at time t whether to participate in the neighborhood information exchange or not. We specifically take ρ_t and ζ_t of the form

$$\rho_t = \frac{\rho_0}{(t+1)^{\epsilon/2}}, \zeta_t = \frac{\zeta_0}{(t+1)^{(1/2-\epsilon/2)}}, \tag{18}$$

where $0 < \epsilon < 1$. Furthermore, define β_t to be

$$\beta_t = (\rho_t \zeta_t)^2 = \frac{\beta_0}{(t+1)}, \ \beta_0 > 0. \tag{19}$$

With the above development in place, we define the random time-varying Laplacian $\mathbf{L}(t)$, where $\mathbf{L}(t) \in \mathbb{R}^{N\times N}$ which abstracts the inter-node information exchange as follows:

$$\mathbf{L}_{i,j}(t) = \begin{cases} -\psi_{i,t}\psi_{j,t} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ \sum_{l\neq i} \psi_{i,t}\psi_{l,t} & i = j. \end{cases} \tag{20}$$

The communication protocol (17)–(20) assumes that the neighboring nodes communicate only when the corresponding communication link is bi-directional. How bi-directional communication links can be enforced in practice is discussed next. Let us first assume that there exists a dedicated reliable bi-directional communication link between any two neighboring nodes. Consider a link between nodes n and l at time t. If $\psi_{n,t} = 1$, node n participates in communication, and it turns on both its transmitting and receiving antennas. If $\psi_{n,t} = 0$, it switches off both its transmitting and receiving antennas. Suppose that $\psi_{n,t} = 1$, and consider two scenarios: (1) $\psi_{l,t} = 0$ and (2) $\Psi_{l,t} = 1$. Consider first the former case. Since node n listens the dedicated channel to node l and node l does not transmit, node n verifies that it does not receive the respective message from node l (e.g., within a prescribed time window), and hence, it does not incorporate node l's estimate in its update. Also, as $\Psi_{l,t} = 0$, node l does not include the estimate by node n, by algorithm construction. Next, consider the case $\Psi_{l,t} = 1$. In this case, node n listens the channel and receives the message by node l, and thus, it incorporates node l's estimate in its update. Completely symmetrically, node l listens the channel from node n to node l, receives the respective message, and includes node n's estimate in its update. Overall, the preceding discussion explains how the symmetric communication protocol can be established. A very similar consideration can be derived if the links are unreliable but still symmetric, in the sense that if the link from n to l is strong enough to support communication, then so is the link from l to n. Finally, if the physical links can fail in an asymmetric fashion, then the proposed algorithm (see ahead (26)–(28) cannot be implemented in its direct form. More precisely, asymmetric failing links yield the Laplacian matrices $\mathbf{L}(t)$ become non-symmetric. The algorithm (26)–(28) and the corresponding analysis need to change in such scenario. This lies outside the scope of this paper, but it corresponds to an interesting future research direction.

With the protocol described in (17)–(20), both the weight assigned to the links and the probability of the existence of a link decay over time. We next consider the first moment, the second moment, and the variance of the Laplacian entries for $\{i,j\} \in E$:

$$\begin{aligned} \mathbb{E}\left[\mathbf{L}_{i,j}(t)\right] &= -\left(\rho_t\zeta_t\right)^2 = -\beta_t = -\frac{\beta_0}{(t+1)} \\ \mathbb{E}\left[\mathbf{L}^2_{i,j}(t)\right] &= \left(\rho_t^2\zeta_t\right)^2 = \frac{\rho_0^2\beta_0}{(t+1)^{1+\epsilon}} \end{aligned} \tag{21}$$

$$Var\left(\mathbf{L}_{i,j}(t)\right) = \frac{\rho_0^2\beta_0}{(t+1)^{1+\epsilon}} - \frac{\beta_0^2}{(t+1)^2}. \tag{22}$$

For future reference, we also introduce the mean Laplacian matrix $\{\mathbf{L}(t)\}$ as $\overline{\mathbf{L}}(t) = \mathbb{E}\left[\mathbf{L}(t)\right]$, and $\widetilde{\mathbf{L}}(t) = \mathbf{L}(t) - \overline{\mathbf{L}}(t)$. Thus, it holds that $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(t)\right\|^2\right] \leq 2N^3\mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right] \leq \frac{2N^3\beta_0\rho_0^2}{(t+1)^{1+\epsilon}}, \tag{23}$$

where $\|\cdot\|$ denotes the L_2 norm. Inequality (23) can be obtained as follows. First, we have that $\left\|\widetilde{\mathbf{L}}(t)\right\| \leq \left\|\widetilde{\mathbf{L}}(t)\right\|_F$, where $\|.\|_F$ denotes the Frobenius norm. Also, note that

$$\begin{aligned} \left\|\widetilde{\mathbf{L}}(t)\right\|_F^2 &= \sum_{i,j=1}^{N}\left|\widetilde{\mathbf{L}}_{i,j}(t)\right|^2 = \sum_{i=1}^{N}\left(\sum_{j\neq i}\left|\widetilde{\mathbf{L}}_{i,j}(t)\right|^2 + \left|\widetilde{\mathbf{L}}_{i,i}(t)\right|^2\right) \\ &= \sum_{i=1}^{N}\left(\sum_{j\neq i}\left|\widetilde{\mathbf{L}}_{i,j}(t)\right|^2 + \left|\sum_{j\neq i}\widetilde{\mathbf{L}}_{i,j}(t)\right|^2\right) \\ &\leq \sum_{i=1}^{N}\left(\sum_{j\neq i}\left|\widetilde{\mathbf{L}}_{i,j}(t)\right|^2 + N\sum_{j\neq i}\left|\widetilde{\mathbf{L}}_{i,j}(t)\right|^2\right) \\ &\leq 2N\sum_{i=1}^{N}\sum_{j\neq i}\left|\widetilde{\mathbf{L}}_{i,j}(t)\right|^2. \end{aligned}$$

Taking expectation and using (17), inequality (23) follows.

Next, we also have that, $\overline{\mathbf{L}}(t) = \beta_t\overline{\mathbf{L}}$, where

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l\neq i} L_{i,l} & i = j. \end{cases} \tag{24}$$

We next give an assumption on the connectivity of the inter-agent communication graph.

Assumption 5 *The inter-agent communication graph is connected on average, i.e.,* $\lambda_2(\overline{\mathbf{L}}) > 0$*, which implies* $\lambda_2(\overline{\mathbf{L}}(t)) > 0$*, where* $\overline{\mathbf{L}}(t)$ *denotes the mean of the Laplacian matrix* $\mathbf{L}(t)$ *and* $\lambda_2(\cdot)$ *denotes the second smallest eigenvalue.*

Assumption 5 ensures consistent information flow among the agent nodes. Technically speaking, the communication graph modeled here as a random undirected graph need not be connected at all times. It is to be noted that Assumption 3 ensures that $\overline{\mathbf{L}}(t)$ is connected at all times as $\overline{\mathbf{L}}(t) = \beta_t \overline{\mathbf{L}}$. We now state additional assumption on the smoothness of the sensing functions for the distributed setup.

Assumption 6 *For each n, the sensing function* $\mathbf{f}_n(\cdot)$ *is Lipschitz continuous on* Θ*, i.e., for each agent n, there exists a constant* $k_n > 0$ *such that*

$$\|\mathbf{f}_n(\boldsymbol{\theta}) - \mathbf{f}_n(\boldsymbol{\theta})\| \leq k_n \|\boldsymbol{\theta} - \boldsymbol{\theta}\|, \tag{25}$$

for all $\boldsymbol{\theta}, \boldsymbol{\theta} \in \Theta$.

With the communication protocol established, we propose an update, where every node n generates an estimate sequence $\{\mathbf{x}_n(t)\}$, where $\mathbf{x}_n(t) \in \mathbb{R}^M$ in the following way:

$$\widehat{\mathbf{x}}_n(t+1) = \mathbf{x}_n(t) \underbrace{- \beta_t \sum_{l \in \Omega_n} \psi_{n,t}\psi_{l,t} \left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{neighborhood consensus}} \underbrace{- \alpha_t \left(\nabla \mathbf{f}_n(\mathbf{x}_n(t))\right) \mathbf{R}_n^{-1} \left(\mathbf{f}_n(\mathbf{x}_n(t)) - \mathbf{y}_n(t)\right)}_{\text{local innovation}} \tag{26}$$

and

$$\mathbf{x}_n(t+1) = \mathcal{P}_\Theta[\widehat{\mathbf{x}}_n(t+1)], \tag{27}$$

where Ω_n denotes the neighborhood of node n with respect to the network represented by $\overline{\mathbf{L}}$, α_t is the innovation gain sequence which is given by $\alpha_t = \alpha_0/(t+1)$, $\alpha_0 > 0$, and $\mathcal{P}_\Theta[\cdot]$ the projection operator corresponding to projecting on Θ. The random variable $\psi_{n,t}$ determines the activation state of a node n. By activation, we mean, if $\psi_{n,t} \neq 0$, then node n can send and receive information in its neighborhood at time t. However, when $\psi_{n,t} = 0$, node n neither transmits nor receives information. The link between node n and node l gets assigned a weight of ρ_t^2 if and only if $\psi_{n,t} \neq 0$ and $\psi_{l,t} \neq 0$.

The update in (26) can be written in a compact manner as follows:

$$\widehat{\mathbf{x}}(t+1) = \mathbf{x}(t) - \left(\mathbf{L}(t) \otimes \mathbf{I}_M\right) \mathbf{x}(t) + \alpha_t \mathbf{G}(\mathbf{x}(t)) \mathbf{R}^{-1} \left(\mathbf{y}(t) - \mathbf{f}(\mathbf{x}(t))\right). \tag{28}$$

Here, $\otimes$ denotes the Kronecker product, I_M denotes the $M \times M$ identity matrix, and:

$$\begin{aligned}
\mathbf{x}(t)^\top &= \left[\mathbf{x}_1(t)^\top \cdots \mathbf{x}_N(t)^\top\right] \mathbf{y}(t)^\top = \left[y_1(t)^\top \cdots y_N(t)^\top\right] \\
\widehat{\mathbf{x}}(t)^\top &= \left[\widehat{\mathbf{x}}_1(t)^\top \cdots \widehat{\mathbf{x}}_N(t)^\top\right] \\
\mathbf{f}(\mathbf{x}(t)) &= \left[\mathbf{f}_1(\mathbf{x}_1(t))^\top \cdots \mathbf{f}_N(\mathbf{x}_N(t))^\top\right]^\top \\
\mathbf{R}^{-1} &= \operatorname{diag}\left[\mathbf{R}_1^{-1}, \cdots, \mathbf{R}_N^{-1}\right] \\
\mathbf{G}(\mathbf{x}(t)) &= \operatorname{diag}\left[\nabla \mathbf{f}_1(\mathbf{x}_1(t)), \cdots, \nabla \mathbf{f}_N(\mathbf{x}_N(t))\right].
\end{aligned}$$

Remark 1 *The Laplacian sequence that plays a role in the analysis in this paper, takes the form* $L(t) = \beta_t \overline{L} + \widetilde{L}(t)$*, where* $\widetilde{L}(t)$ *the residual Laplacian sequence does not scale with* β_t *owing to the fact that the communication rate is chosen adaptively and thus makes the Laplacian matrix sequence not identically distributed.*

We refer to the parameter estimate update in (26) and the projection in (27) in conjunction with the randomized communication protocol as the $\mathcal{CREDO} - \mathcal{NL}$ algorithm. We propose a condition on the sensing functions (standard in the literature of general recursive procedures) that guarantees the existence of stochastic Lyapunov functions and, hence, the convergence of the distributed estimation procedure.

Assumption 7 *The following aggregate strict monotonicity condition holds: there exists a constant* $c_1 > 0$ *such that for each pair* $\boldsymbol{\theta}, \acute{\boldsymbol{\theta}}$ *in* Θ *we have that*

$$\sum_{n=1}^{N} \left(\boldsymbol{\theta} - \acute{\boldsymbol{\theta}}\right)^\top \left(\nabla f_n(\boldsymbol{\theta})\right) \mathbf{R}_n^{-1} \left(f_n(\boldsymbol{\theta}) - f_n(\acute{\boldsymbol{\theta}})\right) \geq c_1 \left\|\boldsymbol{\theta} - \acute{\boldsymbol{\theta}}\right\|^2. \tag{29}$$

The instrumental step in analyzing the convergence of the proposed algorithm is ensuring the existence of appropriate stochastic Lyapunov functions (see, for example [16–20]) which is in turn guaranteed by Assumption 7.

Remark 2 *It is to be noted that the Assumptions 6–7 are only sufficient conditions. Moreover, the assumptions which play a key role in establishing the main results, i.e., Assumptions 2, 1, 6, and 7 are required to hold only in the parameter set* Θ *instead of the entire space* $\mathbb{R}^M$*, which makes our algorithm to apply to very general non-linear sensing functions.*

We consider a specific example to give more intuition about the assumptions in this paper. If the $\mathbf{f}_n(\cdot)$'s are linear, i.e., $\mathbf{f}_n(\boldsymbol{\theta}) = \mathbf{F}_n \boldsymbol{\theta}$, where $\mathbf{F}_n$ is the sensing matrix with dimensions $M_n \times M$, Assumption 3 becomes

equivalent to $\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{R}_n^{-1} \mathbf{F}_n$ being full rank.[4] Under this context, the monotonicity condition in Assumption 7 is trivially satisfied by the positive definiteness of the matrix $\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{R}_n^{-1} \mathbf{F}_n$. We formalize an assumption on the innovation gain sequence $\{\alpha_t\}$ before proceeding further.

Assumption 8 *We require that α_0 satisfies*

$$\alpha_0 c_1 > 1, \tag{30}$$

where c_1 is defined in Assumption 7 and α_0 is the innovation gain at $t = 0$.

The communication cost per node for the proposed algorithm is given by $\mathcal{C}_t = \sum_{s=0}^{t-1} \zeta_s = \Theta\left(t^{(1+\epsilon)/2}\right)$, which in turn is strictly sub-linear as $\epsilon < 1$.

4 Main results

In this section, we present the main results of the proposed algorithm $\mathcal{CREDO} - \mathcal{NL}$, while the proofs of the main results are relegated to Section 7. The first result concerns with the consistency of the estimate sequence $\{\mathbf{x}_n(t)\}$.

Theorem 4.1 *Let Assumptions 1–3 and 5–8 hold. Consider the sequence $\{\mathbf{x}_n(t)\}$ generated by algorithm (26)–(27) at each agent n, with the parameters set to $\rho_t = \frac{\rho_0}{(t+1)^{\epsilon/2}}$, $\zeta_t = \frac{\zeta_0}{(t+1)^{(1/2-\epsilon/2)}}$, and $\alpha_t = \alpha_0/(t+1)$, where $\rho_0, \zeta_0, \alpha_0$ are arbitrary positive numbers. Then, for each n, we have*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1. \tag{31}$$

Theorem 4.1 verifies that the estimate sequence generated by $\mathcal{CREDO} - \mathcal{NL}$ at any agent n is strongly consistent, i.e., $\mathbf{x}_n(t) \to \boldsymbol{\theta}$ almost surely (a.s.) as $t \to \infty$. While Assumption 4 is needed for asymptotic normality results as in Proposition 1, it is not necessary for Theorem 4.1 (nor Theorem 4.2 ahead) to hold.

We now state a main result of this paper which establishes the MSE communication rate for the proposed algorithm $\mathcal{CREDO} - \mathcal{NL}$.

Theorem 4.2 *Let the hypothesis of Theorem 4.1 hold. Then, we have, for each n,*

$$\mathbb{E}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = O\left(\frac{1}{t}\right). \tag{32}$$

Furthermore, for each n, we have:

$$\mathbb{E}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = O\left(\mathcal{C}_t^{-\frac{2}{\epsilon+1}}\right), \tag{33}$$

where $0 < \epsilon < 1$ and is as defined in (18).

We make several remarks on Theorems 4.1 and 4.2.

Remark 3 *Note that ϵ in Theorem 4.2 can be taken to be arbitrarily small. Hence, $\mathcal{CREDO} - \mathcal{NL}$ achieves MSE communication rate arbitrarily close to $1/\mathcal{C}_t^2$. This is a significant improvement over existing non-linear distributed consensus + innovations estimation methods, e.g., [18, 20]. They have $O(t)$ communication cost up to time t and a MSE iteration-wise rate of $O(1/t)$, hence achieving $O(1/\mathcal{C}_t)$ MSE communication rates. $\mathcal{CREDO} - \mathcal{NL}$ achieves the order-optimal $O(1/t)$ MSE iteration-wise rate with a reduced communication cost, thus significantly improving the MSE communication rate.*

Remark 4 *Observe that $\mathcal{CREDO} - \mathcal{NL}$ algorithm, with $\beta_t = \beta_0 (t+1)^{-1}$ has communication cost of $\mathcal{C}_t = \Theta\left(t^{0.5(1+\epsilon)}\right)$. From this, we can see that MSE as a function of $\mathcal{C}_t$ is given by MSE $= O\left(\mathcal{C}_t^{-2/(1+\epsilon)}\right)$.*

Of course, with β_t that decays faster than $1/t$, communication cost reduces further. However, it can be shown that in this case the algorithm no longer produces good estimates. Namely, from standard arguments in stochastic approximation, it can be shown that for $\beta_t = \beta_0 (t+1)^{-1-\delta}$, with $\delta > 0$, $\mathcal{CREDO} - \mathcal{NL}$'s estimate sequence may not converge to $\boldsymbol{\theta}$.

Remark 5 *The $\mathcal{CREDO} - \mathcal{NL}$ algorithm builds on our prior work in [37, 38, 40], but establishing Theorems 4.1–4.2 incurs several technical challenges with respect to our past work. Namely, from a technical standpoint, the $\mathcal{CIWNLS}$ algorithm in [40] incurs the challenge of non-linear observation models. On the other hand, $\mathcal{CREDO}$ in [37, 38] incurs the challenge of increasingly sparse communications. Differently from $\mathcal{CREDO}$ and $\mathcal{CIWNLS}$, this paper simultaneously accounts for both of these challenges. This makes mean square and asymptotic normality analysis more challenging. As a consequence of this difference, while for $\mathcal{CIWNLS}$ and $\mathcal{CREDO}$ we establish both MSE iteration-wise convergence rate analysis and asymptotic normality, here we establish only the MSE (iteration-wise and communication-wise) convergence rate results. Next, $\mathcal{CREDO} - \mathcal{NL}$ is a single time scale stochastic approximation-type algorithm, while both $\mathcal{CIWNLS}$ and $\mathcal{CREDO}$ are two time scale algorithms. Further, the consensus potentials in $\mathcal{CIWNLS}$ and in $\mathcal{CREDO} - \mathcal{NL}$ are the same only on average, i.e., up to the first moment. The difference in higher order moments corresponds to different analyses, namely, the randomized communication protocol that incurs with $\mathcal{CREDO} - \mathcal{NL}$, an increased upper bound of the iteration-wise estimate of MSE. A careful analysis in this paper shows that the additional terms in the MSE bounds with $\mathcal{CREDO} - \mathcal{NL}$ decay faster with time t than $1/t$, and hence, the MSE iteration-wise rate remains order-optimal and equal to $1/t$ (see the proof of Theorems 4.1 and 4.2 in Appendix A.) Finally, we point out*

that the differences of Theorem 4.1 with respect to works [37, 38] *mainly arise from the fact that we consider here nonlinear observation models. Due to this difference, several terms that appear in MSE upper bounds are bounded in a technically different way—see the proof of Lemma A1 in Appendix A. Therein, we need to use the arguments like the non-expansiveness property of projections and Lipschitz continuity of functions* $\mathbf{f}_n$*, none of which is explicitly used in* [37, 38].

5 Simulation experiments

This section corroborates our theoretical findings through simulation examples and demonstrates the communication efficiency of $\mathcal{CREDO} - \mathcal{NL}$.

Specifically, we compare the proposed communication efficient distributed estimator, $\mathcal{CREDO}$, with the benchmark distributed recursive estimator in (13) and the diffusion algorithm as in [43][5], which both utilize all inter-neighbor communications at all times, i.e., they have a linear communication cost. The example demonstrates that the proposed communication efficient estimator has a similar MSE iteration-wise rate as the two benchmark estimators. The simulation also shows that the proposed estimator improves the MSE *communication rate* with respect to the two benchmarks.

We generate a random geometric network of 10 agents, shown in Fig. 1.

The relative degree[6] of the graph is equal to 0.4. The graph was generated as a connected instance of the geometric graph model with radius $r = \sqrt{\ln(N)/N}$. To be specific, the first step involves generating 10 points in a unit square grid and the nodes are connected with a link if the distance between them is less than $\sqrt{\ln(N)/N}$. We repeat the procedure until we get a connected graph instance. We choose the parameter set Θ to be $\Theta = \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^7 \in \mathbb{R}^7$. This choice of Θ conforms with Assumption 2. The sensing functions are chosen to be certain trigonometric functions as described below. The underlying parameter is set as $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_6]$ and thus $\boldsymbol{\theta} \in \mathbb{R}^7$. The sensing functions at the agents are taken to be, $\mathbf{f}_1(\boldsymbol{\theta}) = \sin(\theta_1 + \theta_2 + \theta_3), \mathbf{f}_2(\boldsymbol{\theta}) = \sin(\theta_3 + \theta_2 + \theta_4), \mathbf{f}_3(\boldsymbol{\theta}) = \sin(\theta_3 + \theta_4 + \theta_5), \mathbf{f}_4(\boldsymbol{\theta}) = \sin(\theta_4 + \theta_5 + \theta_6), \mathbf{f}_5(\boldsymbol{\theta}) = \sin(\theta_6 + \theta_5 + \theta_7), \mathbf{f}_6(\boldsymbol{\theta}) = \sin(\theta_6 + \theta_7 + \theta_1), \mathbf{f}_7(\boldsymbol{\theta}) = \sin(\theta_1 + \theta_2 + \theta_7), \mathbf{f}_8(\boldsymbol{\theta}) = \sin(\theta_1 + \theta_2 + \theta_4), \mathbf{f}_9(\boldsymbol{\theta}) = \sin(\theta_2 + \theta_3 + \theta_6)$ and $\mathbf{f}_{10}(\boldsymbol{\theta}) = \sin(\theta_3 + \theta_4 + \theta_6)$. Thus, it is to be noted that each node makes a scalar observation at time t. The noises $\gamma_n(t)$ are Gaussian and are i.i.d. both in time and across nodes and have the covariance matrix equal to $0.25 \times \mathbf{I}_{10}$. The local sensing functions render the parameter $\boldsymbol{\theta}$ locally unobservable, but the parameter $\boldsymbol{\theta}$ is globally observable as, under the parameter set Θ considered in this setup, $\sin(\cdot)$ is one-to-one and the set of linear combinations of the $\boldsymbol{\theta}$ components corresponding to the arguments of the $\sin(\cdot)$'s constitute a full-rank system for $\boldsymbol{\theta}$. Hence, the global observability requirement specified by Assumption 3 is satisfied. The unknown but deterministic value of the parameter is taken to be $\boldsymbol{\theta} = [\pi/6, -\pi/7, \pi/12, -\pi/5, \pi/16, 7\pi/36, \pi/10]$. Under the model considered here in terms of the sensing functions as specified above and the parameter set $\Theta = \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^7$, it can be easily verified that the model conforms to the conditions specified in Assumptions 3–7. The projection operator $\mathcal{P}_\Theta$ onto the set Θ defined in (14) is given by,

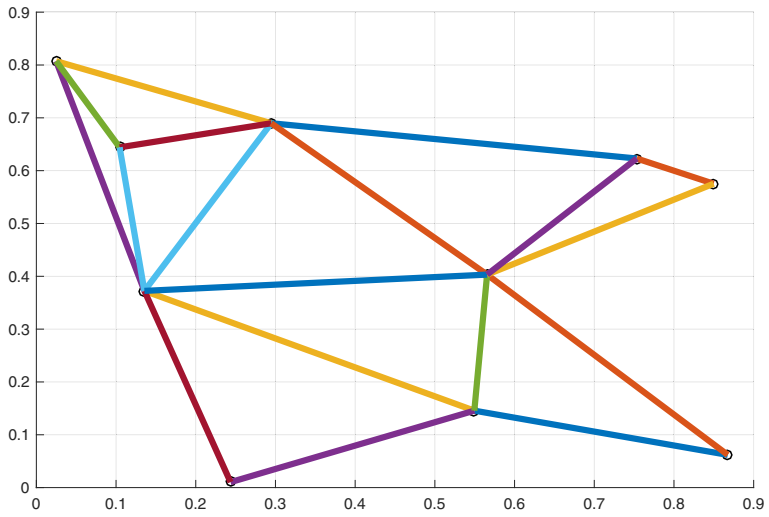

Fig. 1 Network deployment of 10 agents

$$[\mathbf{x}_n(t)]_i = \begin{cases} \frac{\pi}{4} & [\widehat{\mathbf{x}}_n(t)]_i \geq \frac{\pi}{4} \\ [\widehat{\mathbf{x}}_n(t)]_i & \frac{-\pi}{4} < [\widehat{\mathbf{x}}_n(t)]_i < \frac{\pi}{4} \\ \frac{-\pi}{4} & [\widehat{\mathbf{x}}_n(t)]_i < \frac{-\pi}{4}, \end{cases} \tag{34}$$

for all $i = 1, \cdots, M$.

The parameters of the two benchmarks and of the proposed estimator are as follows. The benchmark estimator in (13) has the consensus weight set to $0.48(t+1)^{-1}$. For the proposed estimator, we set $\rho_t = 0.45(t+1)^{-0.01}$ and $\zeta_t = (t+1)^{-0.49}$. The step size sequence for the benchmark estimator proposed in [43] is set to $\mu_t = (0.3(t+20))^{-1}$.

It is to be noted that the Laplacian matrix considered for the benchmark estimator and the expected Laplacian matrix for the proposed estimator, $\mathcal{CREDO}-\mathcal{NL}$ are equal, i.e., $\overline{\mathbf{L}} = \mathbf{L}$. The innovation weight is set to $\alpha_t = (0.3(t+20))^{-1}$. It is to be noted that with the time shifted innovation potential, the theoretical results in this paper continue to hold. As a performance metric, we use the relative MSE estimate averaged across nodes:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\|\mathbf{x}_n(t)-\boldsymbol{\theta}\|^2}{\|\mathbf{x}_n(0)-\boldsymbol{\theta}\|^2},$$

further averaged across 100 independent runs of the estimators. In the above equation, $\mathbf{x}_n(0)$ refers to the initial estimates at each node, which is set as $\mathbf{x}_n(0) = 0$. Figure 2 plots the relative MSE decay in terms of the number of iterations or the number of samples. It can be seen that the MSE decay of the two benchmark estimators and the MSE decay of the proposed estimator $\mathcal{CREDO}-\mathcal{NL}$ are very similar with respect to the iteration count. Figure 3 plots the MSE decay of the three estimators in terms of the communication cost per node. It can be seen for example that, at a relative MSE level of 10^{-1}, the proposed estimator requires 20 and 18 times less communications as compared to the estimator in (13) and the algorithm in [43]. One can also notice a faster MSE decay in terms of the communication cost for $\mathcal{CREDO}-\mathcal{NL}$ as compared to the benchmark (13), thus confirming our theory.

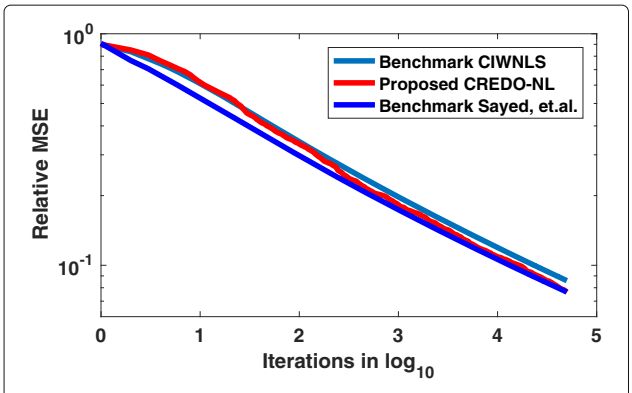

Fig. 2 Comparison of the proposed and benchmark estimators in terms of relative MSE: Number of Iterations. The light blue line represents the $\mathcal{CIWNLS}$ algorithm, the dark blue line represents the diffusion based algorithm proposed in [43], and the red line represents the proposed estimator

6 Discussion

In the context of existing work on non-linear distributed methods, e.g., [15, 16, 31, 34–36, 40], the current paper contributes by developing a method with a strictly faster communication rate of $O(1/\mathcal{C}_t^{2-\zeta})(\zeta > 0$ arbitrarily small) with respect to existing $O(1/\mathcal{C}_t)$ rates. Further, with respect to existing works that develop methods designed to achieve communication efficiency, e.g., [13, 21, 44–46], we develop here a different scheme with *randomized increasingly sparse communications*. Finally, this paper is a continuation of works [37, 38] but, in contrast with [37, 38], it considers non-linear observation models. This requires novel analysis techniques as detailed in Section 1. It would be interesting to apply the proposed method on real data sets, e.g., in the context of IoT or power systems applications, in addition to synthetic data tests considered here.

7 Conclusions

In this paper, we have proposed $\mathcal{CREDO}-\mathcal{NL}$—a communication-efficient distributed estimation scheme for non-linear observation models. We established strong consistency of the estimate sequence at each agent and characterized the MSE decay in terms of the per-agent communication cost $\mathcal{C}_t$. $\mathcal{CREDO}-\mathcal{NL}$ achieves the MSE decay rate $O\left(\mathcal{C}_t^{-2+\zeta}\right)$, where $\zeta > 0$ and ζ is arbitrarily small. Future research directions include extending the proposed algorithm to a mixed-time scale stochastic approximation type algorithm, so as to achieve an asymptotic covariance independent of the network, as well as to extend the presented ideas to distributed stochastic optimization.

Endnotes

[1] From now on, in order to better distinguish the MSE rate of decay with respect to the number of iterations t and with respect to the number of per-node communications, we will refer to the former as the *MSE iteration-wise rate* and to the latter as the *MSE communication rate*.

[2] The stronger requirement imposed here, with ϵ_1 being strictly positive, is only required for the benchmark estimator in Eqs. (13)-(14) ahead to be defined properly; the reason for this requirement is the two time scale nature

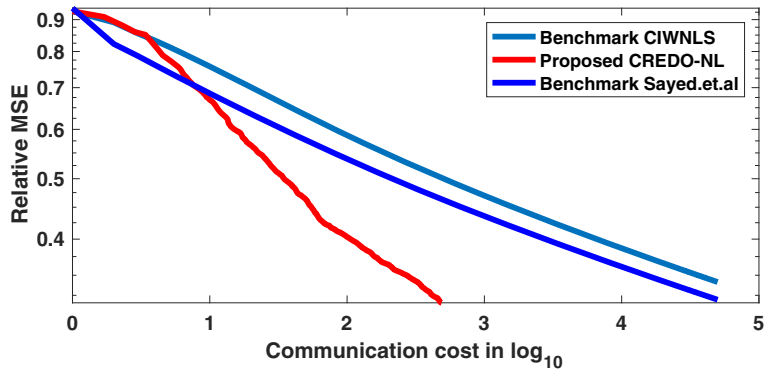

Fig. 3 Comparison of the proposed and benchmark estimators in terms of relative MSE: Communication Cost Per Node. The light blue line represents the $\mathcal{CIWNLS}$ algorithm, the dark blue line represents the diffusion-based algorithm proposed in [43], and the red line represents the proposed estimator

of the benchmark estimator (13)-(14). As the proposed $\mathcal{CREDO}-\mathcal{NL}$ estimator is single time scale, ϵ_1 can be taken to be zero, and the main results (Theorems 1 and 2 ahead) continue to hold.

[3] To see this, note that the dependence of the measurements on the state is through sinusoidal functions (see Eq. (4)), which are everywhere differentiable and thus the gradient of $\mathbf{f}_n(\cdot)$ within the domain Θ exists everywhere. Moreover, as the derivatives of $\sin(\cdot)$ and $\cos(\cdot)$ are bounded, the norm of gradient of $\mathbf{f}_n(\cdot)$ is bounded. Finally, regarding Assumption 3, it can be shown that the assumption is satisfied if (1) graph G is connected; (2) the set of admissible phase angle values, i.e., the parameter constraint set Θ, is chosen appropriately; (3) the real power flow between nodes n and l is non-zero if and only if there exists a physical transmission line connecting the nodes; and (4) voltage magnitude $\mathcal{V}_n \neq 0$, for all nodes n. Please see Proposition 27 in [19].

[4] To see why this is true, consider for simplicity the case $\mathbf{R}_n = I$, for all n. Then, there holds: $\sum_{n=1}^{N} \|\mathbf{f}_n(\boldsymbol{\theta}) - \mathbf{f}_n(\boldsymbol{\theta}')\|^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{F}_n\right)(\boldsymbol{\theta} - \boldsymbol{\theta}')$. Now, the statement of Assumption 3 becomes the following: the matrices $\mathbf{F}_n$, $n = 1, ..., N$, are such that there holds: $(\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \left(\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{F}_n\right)(\boldsymbol{\theta} - \boldsymbol{\theta}') = 0$ if and only if $\boldsymbol{\theta} - \boldsymbol{\theta}'$. But this is equivalent to requiring that $\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{F}_n$ is full rank.

[5] Applied to our setting and in our notation, the diffusion method as in [43] takes the following form:

$$\mathbf{x}'_n(t+1) = \mathbf{x}_n(t) - \mu_t \left(\nabla \mathbf{f}_n(\mathbf{x}_n(t))\right) \mathbf{R}_n^{-1} \left(\mathbf{f}_n(\mathbf{x}_n(t)) - \mathbf{y}_n(t)\right)$$
$$\mathbf{x}_n(t+1) = \sum_{l \in \Omega_n \cup \{n\}} a_{ln}\, \mathbf{x}'_l(t+1).$$

Here, $\mathbf{x}_n(t)$ is the solution estimate at agent n, $\mathbf{x}'_n(t)$ is an auxiliary sequence at agent n, μ_t is the step-size, and the a_{ln}'s are combination weights that constitute together a $N \times N$ column-stochastic matrix.

[6] Relative degree is the ratio of the number of links in the graph to the number of possible links in the graph.

Appendix A: Proof of Main Results

We present the proofs of main results in this section.

Proof of Theorem 4.1 We start the proof with the following useful Lemma. □

Lemma 1 *For each n, the process* $\{\mathbf{x}_n(t)\}$ *satisfies*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\sup_{t \geq 0} \|\mathbf{x}(t)\| < \infty\right) = 1. \tag{35}$$

Proof Consider (14). Since the projection is onto a compact convex set, it is non-expansive. It follows that the inequality

$$\|\mathbf{x}_n(t+1) - \boldsymbol{\theta}\| \leq \|\widehat{\mathbf{x}}_n(t+1) - \boldsymbol{\theta}\| \tag{36}$$

holds for all n and t. We first note that,

$$\mathbf{L}(t) = \beta_t \overline{\mathbf{L}} + \widetilde{\mathbf{L}}(t), \tag{37}$$

where $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$ and $\mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right] = \frac{\rho_0^2 \beta_0}{(t+1)^{1+\epsilon}} - \frac{\beta_0^2}{(t+1)^2}$, for $\{i,j\} \in E, i \neq j$.

Define, $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{1}_N \otimes \boldsymbol{\theta}$ and $V(t) = \|\mathbf{z}(t)\|^2$. (Here, $\mathbf{1}_N$ is the all-ones N by 1 vector.) Note that $\mathbf{z}(t)$ corresponds to the estimation error vector at time t; its squared norm $V(t)$ will first serve us as a Lyapunov function to establish the almost sure boundedness of $\mathbf{x}(t)$ as in

Lemma A1. Let $\{\mathcal{F}_t\}$ be the natural filtration generated by the random observations and the random Laplacians i.e.,

$$\mathcal{F}_t = \sigma\left(\left\{\left\{\mathbf{y}_n(s)\right\}_{n=1}^{N}, \{\mathbf{L}(s)\}\right\}_{s=0}^{t-1}\right). \tag{38}$$

Now, consider the update rules (26)–(28). By algebraic manipulations, conditional independence, and utilizing (36), we have that,

$$\begin{aligned}
&\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] \leq V(t) + \beta_t^2 \mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)^2\mathbf{z}(t)\\
&+\alpha_t^2\mathbb{E}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\right]\\
&-2\beta_t\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{z}(t)\\
&-2\alpha_t\mathbf{z}^\top(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&+2\alpha_t\beta_t\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&+\alpha_t^2\left\|\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)^\top\mathbf{G}^\top\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\right\|^2\\
&+\mathbf{z}^\top(t)\mathbb{E}\left[\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)^2\right]\mathbf{z}(t).
\end{aligned} \tag{39}$$

Consider the orthogonal decomposition

$$\mathbf{z} = \mathbf{z}_C + \mathbf{z}_{C\perp}, \tag{40}$$

where $\mathbf{z}_C$ denotes the projection of $\mathbf{z}$ to the consensus subspace $\mathcal{C} = \left\{\mathbf{z}\in\mathbb{R}^{MN} | \mathbf{z} = \mathbf{1}_N\otimes\mathbf{a}, \text{for some } \mathbf{a}\in\mathbb{R}^M\right\}$. The following inequalities hold for all $t \geq t_1$, where t_1 is a sufficiently large positive integer:

$$\begin{aligned}
&\mathbf{z}^\top(t)\mathbb{E}\left[\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)^2\right]\mathbf{z}(t) \overset{(q0)}{\leq} \frac{c_5\left\|\mathbf{z}_{\mathcal{C}^\perp}(t)\right\|^2}{(t+1)^{1+\epsilon}}\\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)^2\mathbf{z}(t) \overset{(q1)}{\leq} \lambda_N^2(\overline{\mathbf{L}})||\mathbf{z}_{C\perp}(t)||^2;\\
&\mathbf{z}^\top(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right) \overset{(q2)}{\geq} c_1||\mathbf{z}(t)||^2\geq 0;\\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{z}(t) \overset{(q3)}{\geq} \lambda_2(\overline{\mathbf{L}})\left\|\mathbf{z}_{C\perp}(t)\right\|^2;\\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&\quad\overset{(q4)}{\leq} c_2\left\|\mathbf{z}(t)\right\|^2.
\end{aligned} \tag{41}$$

Here, we recall that $\lambda_N(\overline{\mathbf{L}})$ is the largest eigenvalue of matrix $\overline{\mathbf{L}}$. Further, c_1 is defined in Assumption 7, and c_2, c_5 are appropriately chosen positive constants. Here, $\mathbf{z}_{C\perp}(t) = \mathbf{z}(t) - \mathbf{z}_C(t)$, where $\mathbf{z}_C(t)$ is the projection of $\mathbf{z}(t)$ on the consensus subspace $\mathcal{C}$. Inequality $(q0)$ holds because, as noted above, there holds that $\mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(t)\right] \leq \frac{\rho_0^2\beta_0}{(t+1)^{1+\epsilon}}$, for $\{i,j\}\in E, i\neq j$. Specifically, constant c_5 can be taken to equal $2N^3\rho_0^2\beta_0$. Next, inequalities $(q1)$ and $(q3)$ follow from the properties of the Laplacian. Inequality $(q2)$ follows from Assumption 7, and $(q4)$ follows from Assumption 6 since we have that $\|\nabla\mathbf{f}_n(\mathbf{x}_n(t))\|$ is uniformly bounded from above by k_n for all n, and hence, we have that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$. (Recall quantity $\mathbf{G}\left(\mathbf{x}(t)\right)$ defined before Remark 3.1.) That is, c_2 can be taken as $(\max_{n=1,\cdots,N} k_n)^2\ (\max_{n=1,\cdots,N}\|\mathbf{R}_n^{-1}\|)\ \|\overline{\mathbf{L}}\|$. We also have

$$\mathbb{E}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\right] \leq c_4, \tag{42}$$

for some constant $c_4 > 0$. In (42), we use the fact that the noise process under consideration has finite covariance. We also use the fact that, almost surely, $\|\mathbf{G}\left(\mathbf{x}(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$, which in turn follows from Assumption 6. In particular, c_4 may be taken as $(\max_{n=1,\cdots,N} k_n)^2\ (\max_{n=1,\cdots,N}\|\mathbf{R}_n^{-1}\|)^2\ (\max_{n=1,\cdots,N}\|\mathbf{R}_n\|)^2$. We further have that,

$$\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2 \leq c_3\left\|\mathbf{z}(t)\right\|^2, \tag{43}$$

where $c_3 > 0$ is a constant. It is to be noted that (43) follows from the Lipschitz continuity in Assumption 6 and the result that $\|\mathbf{G}\left(\mathbf{x}(t)\right)\| \leq \max_{n=1,\cdots,N} k_n$. That is, c_3 may be taken as $(\max_{n=1,\cdots,N} k_n)^4\ (\max_{n=1,\cdots,N}\|\mathbf{R}_n^{-1}\|)^2$. Applying the bounds (41)–(43) in (39), we obtain, after some algebraic manipulations,

$$\begin{aligned}
\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right] &\leq (1+c_8\alpha_t^2)V(t)\\
&\quad - c_9\left(\beta_t - \frac{c_5}{(t+1)^{\tau_1+\epsilon}}\right)\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2 + c_6\alpha_t^2,
\end{aligned} \tag{44}$$

where c_6, c_8, c_9 are appropriately chosen positive constants, and c_5 is as in (41). In particular, c_6 may be taken as $c_6 = c_4$; c_8 may be taken as $\beta_0^2\left(\lambda_N(\overline{\mathbf{L}})\right)^2/\alpha_0^2 + 2\beta_0\sqrt{c_3} + c_3$, and c_9 may be taken as $2\,\lambda_2(\overline{\mathbf{L}})$.

As $\frac{c_5}{(t+1)^{1+\epsilon}}$ goes to zero faster than β_t, $\exists t_2$ such that $\forall t \geq t_2$, $\beta_t \geq \frac{c_5}{(t+1)^{1+\epsilon}}$. By the above construction we obtain $\forall t \geq t_2$,

$$\mathbb{E}[\,V(t+1)|\mathcal{F}_t] \leq (1+c_8\alpha_t^2)V(t) + \widehat{\alpha}_t^2, \tag{45}$$

where $\widehat{\alpha}(t) = \sqrt{c_6}\alpha_t$. The product $\prod_{s=t}^{\infty}(1+\alpha_s^2)$ exists for all t. Now, let $\{W(t)\}$ be such that

$$W(t) = \left(\prod_{s=t}^{\infty}(1+c_8\alpha_s^2)\right)V(t) + \sum_{s=t}^{\infty}\widehat{\alpha}_s^2,\ \forall t\geq t_2. \tag{46}$$

By (46), it can be shown that $\{W(t)\}$ satisfies,

$$\mathbb{E}[\,W(t+1)|\mathcal{F}_t] \leq W(t). \tag{47}$$

Hence, $\{W(t)\}$ is a non-negative supermartingale and converges a.s. to a bounded random variable W^* as $t\to\infty$. It then follows from (46) that $V(t)\to W^*$ as $t\to\infty$. Thus, we conclude that the desired claim holds. □

The following Lemma will play a key role in establishing the convergence of the estimate sequence.

Lemma 2 (Lemma 4.1 in [18]) *Consider the scalar time-varying linear system*

$$u(t+1) \leq (1-r_1(t))u(t) + r_2(t), \tag{48}$$

where $\{r_1(t)\}$ *is a sequence, such that*

$$\frac{a_1}{(t+1)^{\delta_1}} \leq r_1(t) \leq 1 \tag{49}$$

with $a_1 > 0, 0 \leq \delta_1 < 1$, *whereas the sequence* $\{r_2(t)\}$ *is given by*

$$r_2(t) \leq \frac{a_2}{(t+1)^{\delta_2}} \tag{50}$$

with $a_2 > 0, \delta_2 \geq 0$. *Then, if* $u(0) \geq 0$ *and* $\delta_1 < \delta_2$, *we have*

$$\lim_{t\to\infty} (t+1)^{\delta_0} u(t) = 0, \tag{51}$$

for all $0 \leq \delta_0 < \delta_2 - \delta_1$. *Also, if* $\delta_1 = \delta_2$, *then the sequence* $\{u(t)\}$ *stays bounded, i.e.* $\sup_{t\geq 0} \|u(t)\| < \infty$.

We now prove the almost sure convergence of the estimate sequence to the true parameter. Following similar steps as in the proof of Lemma 1, for t large enough

$$\begin{aligned}\mathbb{E}[\, V(t+1)|\mathcal{F}_t] &\leq \left(1 - 2c_1\alpha_t + c_7\alpha_t^2\right) V(t) + c_6\alpha_t^2 \\ &\leq V(t) + c_6\alpha_t^2,\end{aligned} \tag{52}$$

as for t large enough, $-2c_1\alpha_t + c_7\alpha_t^2 < 0$. Here, constant c_6 is as in (44), and c_7 is appropriately chosen positive constant that may be taken as $\beta_0^2\,(\lambda_N(\bar{\mathbf{L}}))^2/\alpha_0^2 + 2\beta_0\sqrt{c_3} + c_3$. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}V_1(t) &= V(t) + c_6 \sum_{s=t}^{\infty} \alpha_s^2 \\ &= V(t) + c_6\alpha_0^2 \sum_{s=t}^{\infty} (t+1)^{-2}.\end{aligned} \tag{53}$$

Since $\left\{(t+1)^{-2}\right\}$ is summable, the process $\{V_1(t)\}$ is bounded from above. Moreover, it also follows that $\{V_1(t)\}_{t\geq t_1}$ is a supermartingale and hence converges a.s. to a finite random variable. By definition from (53), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (52), we have that,

$$\mathbb{E}[\, V(t+1)] \leq (1 - c_1\alpha_t)\, \mathbb{E}[\, V(t)] + c_6\alpha_0^2\,(t+1)^{-2}, \tag{54}$$

for t large enough. The sequence $\{V(t)\}$ then falls under the purview of Lemma 3 ahead, and we have $\mathbb{E}[V(t)] \to 0$ as $t \to \infty$. Finally, by Fatou's Lemma, where we use the non-negativity of the sequence $\{V(t)\}$, we conclude that

$$0 \leq \mathbb{E}[\, V^*] \leq \liminf_{t\to\infty} \mathbb{E}[\, V(t)] = 0, \tag{55}$$

which thus implies that $V^* = 0$ a.s. Hence, $\|\mathbf{z}(t)\| \to 0$ a.s. as $t \to \infty$, and the desired assertion follows.

We will use the following approximation result (Lemma 3) and the generalized convergence criterion (Lemma 4) for the proof of Theorem 2. Lemma 3 is an extension of Lemma 5 in [18]. Lemma 4 is Lemma 10 in [8].

Lemma 3 *Let* $\{b_t\}$ *be a scalar sequence satisfying*

$$b_{t+1} \leq \left(1 - \frac{c}{t+1}\right) b_t + d(t+1)^{-2}, \tag{56}$$

where $d > 0$ *and* $c > 1$. *Then, we have,*

$$\limsup_{t\to\infty}\ (t+1)\, b_t < \infty. \tag{57}$$

Lemma 4 *Let* $\{J(t)\}$ *be an* $\mathbb{R}$*-valued* $\{\mathcal{F}_{t+1}\}$*-adapted process such that* $\mathbb{E}\,[J(t)|\mathcal{F}_t] = 0$ *a.s. for each* $t \geq 1$. *Then the sum* $\sum_{t\geq 0} J(t)$ *exists and is finite a.s. on the set where* $\sum_{t\geq 0} \mathbb{E}\left[J(t)^2|\mathcal{F}_t\right]$ *is finite.*

Proof of Theorem 4.2 Consider inequality (54), and recall that, by Assumption 8, we have that $\alpha_0\, c_1 > 1$. We can now see that the sequence $\{V(t)\}$ then falls under the purview of Lemma 3, and we have

$$\begin{aligned}&\limsup_{t\to\infty} (t+1)\mathbb{E}[\, V(t+1)] < \infty \\ &\Rightarrow \mathbb{E}[\, V(t)] = O\left(\frac{1}{t}\right).\end{aligned} \tag{58}$$

Inequality (58) now clearly implies that, for each agent n, there holds:

$$\mathbb{E}[\, \|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\,] = O\left(\frac{1}{t}\right). \tag{59}$$

The communication cost $\mathcal{C}_t$ for the proposed $\mathcal{CREDO} - \mathcal{NL}$ algorithm is given by $\mathcal{C}_t = \Theta\left(t^{\frac{\epsilon+1}{2}}\right)$, and thus the assertion follows in conjunction with (59). □

Abbreviations

CPS: Cyber-physical systems; $\mathcal{CREDO}$: Communication efficient REcursive Distributed estimatOr; $\mathcal{CREDO} - \mathcal{NL}$: $\mathcal{CREDO}$-non-linear; i.i.d.: Independent identically distributed; IoT: Internet of things

Funding

This work is supported by the I-BiDaaS project, funded by the European Commission under Grant Agreement No. 780787. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. The work of D. Jakovetic is also supported in part by the Serbian Ministry of Education, Science, and Technological Development, grant 174030. The work is also partially supported by the National Science Foundation under grant CCF-1513936.

Availability of data and materials

The data used in this paper is synthetic and is generated as described in Section 5 of the paper. Please contact authors for data requests.

Authors' contributions

AKS lead the writing of Sections 2–5 and Appendix, he also lead carrying out theoretical analysis, and he carried out numerical experiments in Section 5. He also contributed in writing Sections 1, 6, and 7. DJ lead the writing of Sections 1, 6, and 7. He also contributed in writing Sections 2–5 and Appendix and in developing the code for carrying out numerical results in Section 5. DB contributed in writing Sections 1–4. SK contributed in writing Sections 1–3 and Appendix. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA. [2]University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics, 21000 Novi Sad, Serbia. [3]University of Novi Sad, Faculty of Technical Sciences, Department of Power, Electronic and Communication Engineering, 21000 Novi Sad, Serbia.

Received: 23 March 2018 Accepted: 24 September 2018
Published online: 19 October 2018

References

1. A. Abur, A. G. Exposito, *Power System State Estimation: Theory and Implementation*. (Marcel Dekker, New York, 2004)
2. D. Bajović, J. M. F. Moura, J. Xavier, B. Sinopoli, Distributed inference over directed networks: performance limits and optimal design. IEEE Trans. Sig. Process. **64**(13), 3308–3323 (2016)
3. B. Bollobas, *Modern Graph Theory*. (Springer Verlag, New York, 1998)
4. P. Braca, S. Marano, V. Matta, Enforcing consensus while monitoring the environment in wireless sensor networks. IEEE Trans. Sig. Process. **56**(7), 3375–3380 (2008)
5. F. Cattivelli, A. H. Sayed, Diffusion LMS strategies for distributed estimation. IEEE Trans. Sig. Process. **58**(3), 1035–1048 (2010)
6. J. Chen, C. Richard, A. H. Sayed, Multitask diffusion adaptation over networks. IEEE Trans. Sig. Process. **62**(16), 4129–4144 (2014)
7. F. R. K. Chung, *Spectral graph theory, vol. 92*. (American Mathematical Soc., Providence, 1997)
8. L. E. Dubins, D. A. Freedman, A sharper form of the Borel-Cantelli lemma and the strong law. Ann. Math. Stat. **36**(3), 800–807 (1965)
9. V. Fabian, On asymptotically efficient recursive estimation. Ann. Stat. **6**(4), 854–866 (1978)
10. R. Z. Has'minskij, in *Proc. Prague Symp. Asymptotic Statist.* Sequential estimation and recursive asymptotically optimal procedures of estimation and observation control, vol. 1 (Charles Univ., Prague, 1974), pp. 157–178
11. C. Heinze, B. McWilliams, N. Meinshausen, in *19th International Conference on Artificial Intelligence and Statistics*. Dual-loco: distributing statistical estimation using random projections, (Cadiz, 2016), pp. 875–883
12. M. D. Ilic', J. Zaborszky, *Dynamics and Control of Large Electric Power Systems*. (Wiley, New York, 2000)
13. D. Jakovetic, D. Bajovic, N. Krejic, N. Krklec Jerinkic, Distributed gradient methods with variable number of working nodes. IEEE Trans. Sig. Process. **64**(15), 4080–4095 (2016)
14. D. Jakovetic, J. Xavier, J. M. F. Moura, Cooperative convex optimization in networked systems: augmented Lagrangian algorithms with directed gossip communication. IEEE Trans. Sig. Process. **59**(8), 3889–3902 (2011)
15. R. I. Jennrich, Asymptotic properties of non-linear least squares estimators. Ann. Math. Stat. **40**(2), 633–643 (1969)
16. S. Kar, J. M. F. Moura, Asymptotically efficient distributed estimation with exponential family statistics. IEEE Trans. Inf. Theory. **60**(8), 4811–4831 (2014)
17. S. Kar, J. M. F. Moura, H. V. Poor, Distributed linear parameter estimation: asymptotically efficient adaptive strategies. SIAM J. Control Optim. **51**(3), 2200–2229 (2013)
18. S. Kar, J. M. F. Moura, H. V. Poor, QD-Learning: A Collaborative Distributed Strategy for Multi-Agent Reinforcement Learning Through Consensus + Innovations. IEEE Trans. Signal Process. **61**(7), 1848–1862 (2013)
19. S. Kar, J. M. F. Moura, K. Ramanan, Distributed parameter estimation in sensor networks: nonlinear observation models and imperfect communication. IEEE Trans. Inf. Theory. **58**(6), 3575–3605 (2012)
20. S. Kar, J. M. F. Moura, Convergence rate analysis of distributed gossip (linear parameter) estimation: fundamental limits and tradeoffs. IEEE J. Sel. Top. Sig. Process. **5**(4), 674–690 (2011)
21. G. Lan, S. Lee, Y. Zhou, Communication-efficient algorithms for decentralized and stochastic optimization. arXiv preprint arXiv:1701.03961 (2017)
22. J. Li, A. H. Sayed, Modeling bee swarming behavior through diffusion adaptation with asymmetric information sharing. EURASIP J. Adv. Sig. Process. **18**(1), 2012
23. Q. Liu, A. T. Ihler, in *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*. Distributed estimation, information loss and exponential families (MIT Press, Cambridge, 2014), pp. 1098–1106
24. C. G. Lopes, A. H. Sayed, Diffusion least-mean squares over adaptive networks: formulation and performance analysis. IEEE Trans. Sig. Process. **56**(7), 3122–3136 (2008)
25. P. D. Lorenzo, A. H. Sayed, Sparse distributed learning based on diffusion adaptation. IEEE Trans. Sig. Process. **61**(6), 1419–1433 (2013)
26. C. Ma, M. Takáč, Partitioning data on features or samples in communication-efficient distributed optimization? arXiv preprint arXiv:1510.06688 (2015)
27. C. Ma, V. Smith, M. Jaggi, M. Jordan, P. Richtarik, M. Takac, in *ICML'15 Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. Adding vs. averaging in distributed primal-dual optimization, (Lille, 2015), pp. 1973–1982
28. G. Mateos, G. B. Giannakis, Distributed recursive least-squares: stability and performance analysis. IEEE Trans. Sig. Process. **60**(7), 3740–3754 (2012)
29. G. Mateos, I. Schizas, G. B. Giannakis, Performance analysis of the consensus-based distributed LMS algorithm. EURASIP J. Adv. Sig. Process. **68**, 2009 (2009)
30. A. Nedić, A. Olshevsky, C. Uribe, in *2015 American Control Conference (ACC)*. Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs (IEEE, Chicago, 2015). https://doi.org/10.1109/ACC.2015.7172262
31. A. Nedic, A. Ozdaglar, Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control. **54**(1), 48–61 (2009)
32. J. Pfanzagl, in *Proceedings of the Prague Symposium on Asymptotic Statistics*, ed. by J. Hajek. Asymptotic optimum estimation and test procedures, vol. 1 (Charles University, Prague, 1974), pp. 201–272
33. A. Primadianto, C. N. Lu, A review on distribution system state estimation. IEEE Trans. Power Syst. **32**(5), 3875–3883 (2017)
34. S. S. Ram, A. Nedic, V. V. Veeravalli, Incremental stochastic subgradient algorithms for convex optimization. SIAM J. Optim. **20**(2), 691–717 (2009)
35. S. S. Ram, A. Nedić, V. V. Veeravalli, Distributed stochastic subgradient projection algorithms for convex optimization. J. Optim. Theory Appl. **147**(3), 516–545 (2010)
36. S. S. Ram, V. V. Veeravalli, A. Nedic, Distributed and recursive parameter estimation in parametrized linear state-space models. IEEE Trans. Autom. Control. **55**(2), 488–492 (2010)
37. A. K. Sahu, D. Jakovetic, S. Kar, Communication optimality trade-offs for distributed estimation. arXiv preprint arXiv:1801.04050 (2018)
38. A. K. Sahu, D. Jakovetic, S. Kar, in *International Symposium on Information Theory, ISIT*. CREDO: A communication-efficient distributed estimation algorithm, (Vail, 2018)
39. A. K. Sahu, S. Kar, Distributed sequential detection for Gaussian shift-in-mean hypothesis testing. IEEE Trans. Sig. Process. **64**(1), 89–103 (2016)
40. A. K. Sahu, S. Kar, J. M. F. Moura, H. V. Poor, Distributed constrained recursive nonlinear least-squares estimation: algorithms and asymptotics. IEEE Trans. Sig. Inf. Process. Over Networks. **2**(4), 426–441 (2016)
41. D. J. Sakrison, Efficient recursive estimation; application to estimating the parameters of a covariance function. Int. J. Eng. Sci. **3**(4), 461–483 (1965)
42. C. J. Stone, Adaptive maximum likelihood estimators of a location parameter. Ann. Stat. **3**(2), 267–284 (1975)
43. Z. J. Towfic, J. Chen, A. H. Sayed, Excess-risk of distributed stochastic learners. IEEE Trans. Inf. Theory. **62**(10), 5753–5785 (2016)
44. K. Tsianos, S. Lawlor, M. G. Rabbat, in *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*. Communication/computation tradeoffs in consensus-based distributed optimization (Curran Associates Inc., Lake Tahoe, 2012), pp. 1943–1951
45. K. I. Tsianos, S. F. Lawlor, J. Y. Yu, M. G. Rabbat, in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. Networked optimization with adaptive communication (IEEE, Austin, 2013), pp. 579–582

46. Z. Wang, Z. Yu, Q. Ling, D. Berberidis, G. B. Giannakis, Decentralized RLS with data-adaptive censoring for regressions over large-scale networks. IEEE Trans. Signal Proc. **66**(6) (2018)
47. Y. Zhang, J. Duchi, M. Wainwright, in *Proceedings of the 26th Annual Conference on Learning Theory, PMLR. Vol. 30*. Divide and conquer kernel ridge regression, (Princeton, 2013), pp. 592–617
48. F. Zhao, L. J. Guibas, L. Guibas. Wireless Sensor Networks: An Information Processing Approach (Morgan Kaufmann, San Francisco, 2004)