Communication Efficient Distributed Weighted Non-Linear Least Squares Estimation

Anit Kumar Sahu, Dusan Jakovetic, Dragana Bajovic, and Soummya Kar

Abstract The paper addresses design and analysis of communication efficient distributed algorithms for solving weighted non-linear least square problems in multi-agent networks. *Communication efficiency* is highly relevant in modern applications like cyber-physical systems and 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, like, 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

Anit Kumar Sahu
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
E-mail: anits@andrew.cmu.edu
Dusan Jakovetic
Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad
21000 Novi Sad, Serbia
E-mail: djakovet@uns.ac.rs
Dragana Bajovic
Faculty of Technical Sciences, University of Novi Sad
21000 Novi Sad, Serbia
E-mail: dbajovic@uns.ac.rs

Soummya Kar
Department of Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, USA
E-mail: soummyak@andrew.cmu.edu

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 nonlinear 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 internet of things (IoT), like state estimation in smart grid, predictive maintenance and production monitoring in industrial manufacturing systems, and so on. 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 for communication is an order or two more than computation. 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, consensus + innovations-type, distributed estimator for non-linear observation models that we refer to as $\mathcal{CREDO}-\mathcal{NL}$. $\mathcal{CREDO}-\mathcal{NL}$ generalizes the recently proposed linear distributed estimator $\mathcal{CREDO}$ 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-linearly 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 $\Theta(1/t)$ rate of the mean square error (MSE) decay. 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 $\Theta(1/\mathcal{C}_t)$ with existing methods, e.g., [16, 20, 30, 33–35, 39], to $\Theta(1/\mathcal{C}_t^{2-\zeta})$ with the proposed method, where $\zeta > 0$ 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. [12, 24, 26, 27, 45]. The fusion center assigns subtasks 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., [1, 3–5, 7, 14, 19, 23, 25, 28–31, 33–35, 38, 44]. 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., [16, 20, 30, 33–35, 39]. However, among these works, the best achieved MSE rate of decay in terms of per-agent communication cost is $\Theta(1/t)$. In contrast, we establish here a strictly faster communication rate equal to $\Theta(1/\mathcal{C}_t^{2-\zeta})$ ($\zeta > 0$ 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., [15, 22, 42–44]. In [44], a data censoring method is employed to save in terms of computation and communication costs. However, the communication savings in [44] is a constant proportion with respect to a vanilla method which uses all allowable communications at all times. In [22], the communication savings come at a cost of extra computations. References [15, 42, 43] 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 [15, 22, 42, 43] 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.

Within the class of *consensus+innovations* distributed estimation algorithms (see, e.g., [19, 21]), the design of communication efficient methods has been addressed in [37], see also [36], 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 $\|\cdot\|$ as the $\mathcal{L}_2$-norm.

Assumption M1. There exists $\epsilon_1 > 0$, such that, for all n, $\mathbb{E}_{\boldsymbol{\theta}}\left[\|\gamma_n(t)\|^{2+\epsilon_1}\right] < \infty$.

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 [2]. (With the proposed $\mathcal{CRED}\mathcal{O} - \mathcal{NL}$ 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 $(1/\sqrt{N})\mathbf{1}_N$. 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 [6] for instance).

Example: Distributed Static Phase Estimation in Smart Grids

Many applications within cyber physical systems and internet of things can be modeled as non-linear distributed estimation problems of type (1). As an example, we briefly discuss here distributed static phase estimation in smart grids, while we refer to, e.g., [13, 18] 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., [13], 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., [13]:

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

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:

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

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 noise angle ranges (that correspond to the admissible parameter set $\boldsymbol{\Theta}$) and the smart grid network and admittances structure, the assumptions we make on the sensing model are satisfied, and hence $\mathcal{CRLEDO} - \mathcal{NL}$ can be effectively applied; we refer to [13, 18] 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 nonlinear least squares (WNLS) (see, for example, [16]). 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:

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

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

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

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 M2. The set Θ is compact convex subset of $\mathbb{R}^M$ with non-empty interior $\operatorname{int}(\Theta)$ and the true (but unknown) parameter $\boldsymbol{\theta} \in \operatorname{int}(\Theta)$.

Assumption M3. The sensing model is globally observable, i.e., any pair $\boldsymbol{\theta}, \acute{\boldsymbol{\theta}}$ of possible parameter instances in Θ satisfies

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

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

Assumption M4. The sensing function $\mathbf{f}_n(.)$ for each n is continuously differentiable in the interior $\operatorname{int}(\Theta)$ of the set Θ. For each $\boldsymbol{\theta}$ in the set Θ, the (normalized) gain matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ defined by

(8)

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

Smoothness conditions on the sensing functions, such as the one imposed by assumption M3 is common in statistical estimation with non-linear observations models. Note that the matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ is well defined at the true value of the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta} \in \text{int}(\Theta)$ and the continuous differentiability of the sensing functions holds for all $\boldsymbol{\theta} \in \text{int}(\Theta)$.
The asymptotic properties of the WNLS estimator in terms of consistency and asymptotic normality are characterized by the following classical result:

Proposition 1 ([16]) *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 $\mathcal{G}_t = \sigma\left(\{\{\mathbf{y}_n(s)\}_{s=0}^{t-1}\}_{n=1}^{N}\right)$. 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 $\{\widehat{\boldsymbol{\theta}}_t\}$ such that*

$$\widehat{\boldsymbol{\theta}}_t \in argmin_{\mathbf{z} \in \Theta} \mathcal{Q}_t(\mathbf{z}),\ \forall t. \tag{9}$$

Moreover, if the model is globally observable, i.e., Assumption M3 holds, the WNLS estimate sequence $\{\widehat{\boldsymbol{\theta}}_t\}$ is consistent, i.e.,

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

Additionally, if Assumption M4 holds, the parameter estimate sequence is asymptotically normal, i.e.,

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

where

$$\boldsymbol{\Sigma}_c = \left(N \boldsymbol{\Gamma}_{\boldsymbol{\theta}}\right)^{-1}, \tag{12}$$

$\boldsymbol{\Gamma}_{\boldsymbol{\theta}}$ is as given by (8) *and $\overset{\mathcal{D}}{\Longrightarrow}$ 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 [10,11,32,40,41], 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, [18] and equation (16) ahead) type update, where the agents' knowledge of the model is limited to themselves have been proposed in [20, 39]. In [20], 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 [20], [39] 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 [20,39] have the number of communication scales linearly with the number of per-node sampled observations $\{\mathbf{y}_n(t)\}$. This paper builds on the ideas of sequential distributed recursive schemes catering to non-linear observation models as proposed in [20, 39] 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 $\Theta(1/t)$ in terms of the number of per-node processed samples, which reducing the $\Theta(t)$ communication cost which is a characteristic of previous approaches.

Before proceeding further, we briefly summarize the estimator in [39] which is referred to as the *benchmark* estimator henceforth. The overall update rule at an agent n corresponds to

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

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

$$\beta_t = \frac{\beta_0}{(t+1)^{\delta_1}}, \alpha_t = \frac{\alpha_0}{t+1}, \tag{15}$$

where $\alpha_0, \beta_0 > 0, 0 < \delta_1 < 1/2 - 1/(2+\epsilon_1)$ and ϵ_1 was defined in Assumption M1. From the asymptotic normality in Theorem 4.2 in [39] it can be inferred that the MSE decays as $\Theta(1/t)$.

Communication Efficiency

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

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

where agent i is arbitrary (the expectation in (16) does not depend on i) 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 [20,39] 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 $\Theta\left(\frac{1}{\mathcal{C}_t}\right)$.

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

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

where $\psi_{n,t}$'s are independent both across time and the nodes, i.e., across t and n respectively. 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 $0 < \epsilon < \tau_1$ and $0 < \tau_1 \leq 1$. 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 above communication protocol allows two nodes to communicate only when the link is established in a bi-directional fashion and hence avoids directed graphs. The design of the communication protocol as depicted in (17)-(20) not only decays the weight assigned to the links over time but also decays the probability of the existence of a link. The communication protocol depicted above closely replicates

real world networked setups constituting of entities with finite power and decreasing quality of communication over time owing to the communications being power hungry. We have, for $\{i,j\} \in E$:

$$\begin{aligned}\mathbb{E}\left[\mathbf{L}_{i,j}(t)\right] &= -\left(\rho_t \zeta_t\right)^2 = -\beta_t = -\frac{c_3}{(t+1)} \\ \mathbb{E}\left[\mathbf{L}_{i,j}^2(t)\right] &= \left(\rho_t^2 \zeta_t\right)^2 = \frac{c_4}{(t+1)^{1+\epsilon}}.\end{aligned} \tag{21}$$

Thus, we have that, the variance of $\mathbf{L}_{i,j}(t)$ is given by,

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

Define, the mean of the random time-varying Laplacian sequence $\{\mathbf{L}(t)\}$ as $\overline{\mathbf{L}}(t) = \mathbb{E}\left[\mathbf{L}(t)\right]$ and $\widetilde{\mathbf{L}}(t) = \mathbf{L}(t) - \overline{\mathbf{L}}(t)$. Note that, $\mathbb{E}\left[\widetilde{\mathbf{L}}(t)\right] = \mathbf{0}$, and

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(t)\right\|^2\right] \leq N^2 \mathbb{E}\left[\widetilde{\mathbf{L}}_{i,j}^2(t)\right] = \frac{N^2 \beta_0 \rho_0^2}{(t+1)^{\tau_1+\epsilon}} - \frac{N^2 a^2}{(t+1)^{2\tau_1}}, \tag{23}$$

where $\|\cdot\|$ denotes the L_2 norm. The above equation follows from equivalence of the L_2 and Frobenius norms.
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 formalize an assumption on the connectivity of the inter-agent communication graph before proceeding further.

Assumption M5. 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 M3 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 M3 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 M6. 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

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

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:

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

Here, $\otimes$ denotes the Kronecker product, and:

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

Remark 31. *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 $\mathcal{CRED}\mathcal{O} - \mathcal{NL}$ 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 M7. 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

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

The instrumental step in analyzing the convergence of the proposed algorithm is ensuring the existence of appropriate stochastic Lyapunov functions (see, for example [17–20]) which is in turn guaranteed by Assumption M7.

Remark 32. *It is to be noted that the assumptions M6-M7 are only sufficient conditions. Moreover, the assumptions which play a key role in establishing the main results, i.e., Assumptions M2, M1, M6, and M7 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 nonlinear 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 M3 becomes equivalent to $\sum_{n=1}^{N} \mathbf{F}_n^\top \mathbf{R}_n^{-1}\mathbf{F}_n$ being full rank. Under this context, the monotonicity condition in Assumption M7 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 M8. We require that α_0 satisfies

$$\alpha_0 \mu \geq 1, \tag{30}$$

where μ is defined in Assumption M7 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 A. The first result concerns with the consistency of the estimate sequence $\{\mathbf{x}_n(t)\}$.

Theorem 41. *Let assumptions M1-M8 hold. Consider the sequence $\{\mathbf{x}_n(t)\}$ generated by (26) at each agent n. 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 41 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 $t \to \infty$.

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

Theorem 42. *Let the hypothesis of Theorem 41 hold. Then, we have,*

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

Furthermore, when $\beta_t = \frac{\beta_0}{t+1}$, we have:

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

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

We make several comments on Theorem 42. First, note that ϵ in Theorem 42 can be taken to be arbitrarily small. Hence, $\mathcal{CREDO}-\mathcal{NL}$ achieves MSE 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 $\Theta(t)$ communication cost up to time t and a MSE rate of $\Theta(1/t)$, hence achieving $\Theta(1/\mathcal{C}_t)$ MSE communication rates. $\mathcal{CREDO}-\mathcal{NL}$ achieves the order-optimal $\Theta(1/t)$ MSE rate with a reduced communication cost, thus significantly improving the MSE communication rate.

Next, 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 $\mathcal{C}_t$ in the case of $\tau_1 = 1$ is given by MSE $= \Theta(\mathcal{C}_t^{-2/(1+\epsilon)})$. Of course, with a further increase of τ_1 beyond unity, 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 $\boldsymbol{\theta}$.

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) which utilizes all inter-neighbor communications at all times, i.e., has a linear communication cost. The example demonstrates that the proposed communication-efficient estimator matches the MSE rate of the benchmark estimator. The simulation also shows that the proposed estimator improves the MSE *communication rate* with respect to the benchmark.
We generate a random geometric network of 10 agents, shown in Figure 1. The relative degree[1] of the graph is given by 0.4. The graph was generated as a connected graph using 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 $\Theta = \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5 \in \mathbb{R}^5$. This choice of Θ conforms with Assumption M2. The sensing functions are chosen to be certain trigonometric functions as described below. The underlying parameter is set as $\boldsymbol{\theta} = [\theta_1,\ \theta_2,\ \theta_3,\ \theta_4,\ \theta_5]$ and thus $\boldsymbol{\theta} \in \mathbb{R}^5$. The sensing functions at the agents are taken to be, $\mathbf{f}_1(\boldsymbol{\theta}) = \sin(\theta_1+\theta_2), \mathbf{f}_2(\boldsymbol{\theta}) = \sin(\theta_3+\theta_2), \mathbf{f}_3(\boldsymbol{\theta}) = \sin(\theta_3+\theta_4), \mathbf{f}_4(\boldsymbol{\theta}) = \sin(\theta_4+\theta_5), \mathbf{f}_5(\boldsymbol{\theta}) = \sin(\theta_1+\theta_5), \mathbf{f}_6(\boldsymbol{\theta}) = \sin(\theta_1+\theta_3), \mathbf{f}_7(\boldsymbol{\theta}) = \sin(\theta_4+\theta_2), \mathbf{f}_8(\boldsymbol{\theta}) = \sin(\theta_3+\theta_5), \mathbf{f}_9(\boldsymbol{\theta}) = \sin(\theta_1+\theta_4)$ and $\mathbf{f}_{10}(\boldsymbol{\theta}) = \sin(\theta_1+\theta_5)$. 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

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

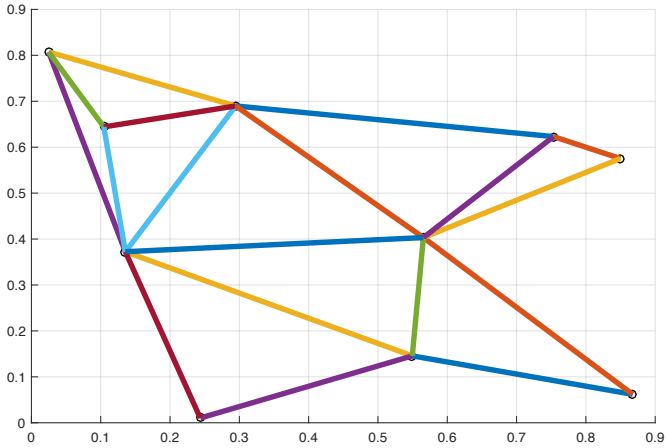

Fig. 1: Network Deployment of 10 agents

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 M3 is satisfied. The unknown but deterministic value of the parameter is taken to be $\boldsymbol{\theta} = [\pi/6,\ -\pi/7,\ \pi/12,\ -\pi/5,\ \pi/16]$. Under the model considered here in terms of the sensing functions as specified above and the parameter set $\Theta = \left[-\frac{\pi}{4}, \frac{\pi}{4}\right]^5$ it can be easily verified that the model conforms to the conditions specified in Assumptions M3-M7. The projection operator $\mathcal{P}_\Theta$ onto the set Θ defined in (14) is given by,

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

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

The parameters of the benchmark and the proposed estimator are as follows. The benchmark estimator's consensus weight is set to $0.48(t+1)^{-1}$. For the proposed estimator, we set $\rho_t = 0.45(t+1)^{-0.01}$ and $\zeta_t = (t+1)^{-0.49}$. It is to be noted that the Laplacian matrix considered for the benchmark estimator and the expected Laplacian matrix for the proposed estimator, $\mathcal{CRED}\mathcal{O} - \mathcal{NL}$ 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 benchmark estimator in (13) and the MSE decay of the proposed estimator $\mathcal{CREDO} - \mathcal{NL}$ practically match with respect to the iteration count. Figure 3 plots the MSE decay of both the estimators in terms of the communication cost per node. It can be seen that at a relative MSE level of 10^{-1}, the proposed estimator requires 20x less communications as compared to the benchmark estimator. At lower relative MSE levels, for instance, at a relative MSE level of 0.03 the communication cost savings are in the order of 100x. One can also notice a faster MSE decay in terms of the communication cost for $\mathcal{CREDO} - \mathcal{NL}$ as compared to the benchmark, thus confirming our theory.

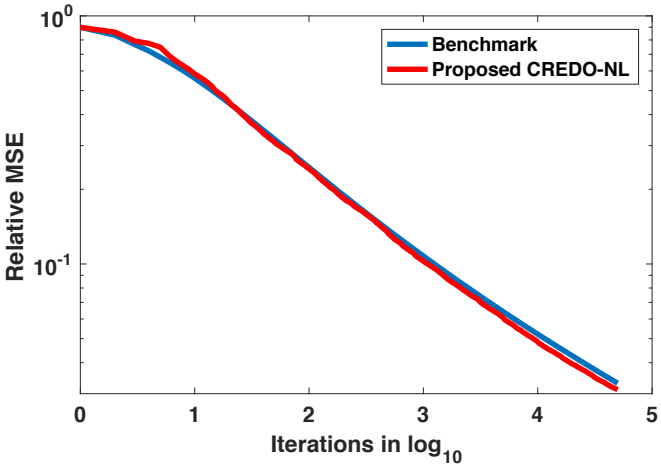

Fig. 2: *Comparison of the proposed and benchmark estimators in terms of relative MSE: Number of Iterations.* The blue line represents the benchmark, while the red line represents the proposed estimator.

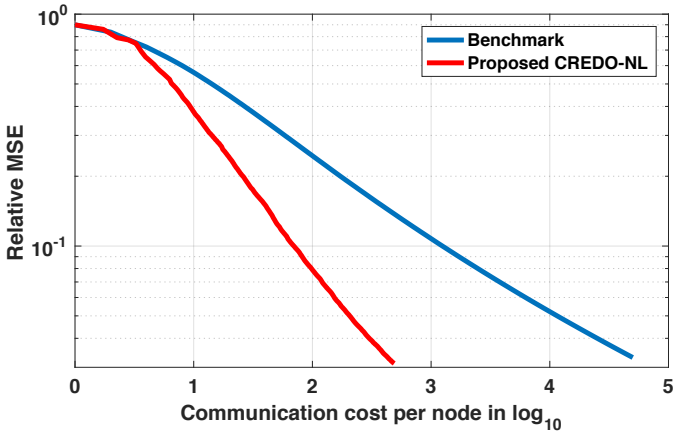

Fig. 3: *Comparison of the proposed and benchmark estimators in terms of relative MSE: Communication Cost Per Node.* The blue line represents the benchmark, while the red line represents the proposed estimator.

6 Discussion

In the context of existing work on non-linear distributed methods, e.g., [16, 20, 30, 33–35, 39]. current paper contributes by developing a method with a strictly faster communication rate of $\Theta(1/\mathcal{C}_t^{2-\zeta})(\zeta > 0$ arbitrarily small) with respect to existing $\Theta(1/\mathcal{C}_t)$ rates. Further, with respect to existing works that develop methods designed to achieve communication efficiency, e.g., [15, 22, 42–44], we develop here a different scheme with *randomized increasingly sparse communications*. Finally, this paper is a continuation of works [36, 37] but, in contrast with [36, 37], 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 Conclusion

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 $\Theta\left(\mathcal{C}_t^{-2+\zeta}\right)$, 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.

8 Abbreviations

Throughout the paper, we use the following abbreviations: IoT: Internet of Things; CPS: cyber-physical systems; i.i.d.: independent identically distributed; $\mathcal{CREDO}$: Communication-efficient recursive distributed estimator; $\mathcal{CREDO} - \mathcal{NL}$: $\mathcal{CREDO}$-non-linear.

9 Declarations

9.1 ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

9.2 CONSENT FOR PUBLICATION

Not applicable.

9.3 AVAILABILITY OF DATA AND MATERIAL

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.

9.4 COMPETING INTERESTS

The authors declare that they have no competing interests.

9.5 FUNDING

The work of D. Jakovetic and D. Bajovic is supported in part by the EU Horizon 2020 project I-BiDaaS, project number 780787. 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.

9.6 AUTHORS' CONTRIBUTIONS

Anit Kumar Sahu lead writing of Sections 2–5 and Section A; 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. Dusan Jakovetic lead writing of Sections 1, 6, and 7. He also contributed in writing Sections 2–5 and A and to developing the code for carrying out numerical results in Section 5. Dragana Bajovic contributed in writing Sections 1–4. Soummya Kar contributed in writing Sections 1–3 and Section A.

9.7 **ACKNOWLEDGEMENTS**

There are no further acknowledgements.

A Proof of Main Results

We present the proofs of main results in this section.

Proof of Theorem 41.

Lemma A1. *For each n, the process $\{\mathbf{x}_n(t)\}$ satisfies*

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

Proof. Since the projection is onto a 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 $\mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right]=\frac{c_4}{(t+1)^{1+\epsilon}}-\frac{c_3^2}{(t+1)^2}$.
Define, $\mathbf{z}(t)=\mathbf{x}(t)-\mathbf{1}_N\otimes\boldsymbol{\theta}^*$ and $V(t)=\|\mathbf{z}(t)\|^2$. By conditional independence, we have that,

$$\begin{aligned}
&\mathbb{E}\left[V(t+1)|\mathcal{F}_t\right]\leq V(t)+\beta_t^2\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)^2\mathbf{z}(t)\\
&+\alpha_t^2\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{y}(t)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\right\|^2\right]\\
&-2\beta_t\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{z}(t)\\
&-2\alpha_t\mathbf{z}^\top(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&+2\alpha_t\beta_t\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&+\alpha_t^2\left\|\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)^\top\mathbf{G}^\top\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\right\|^2+\mathbf{z}^\top(t)\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)^2\right]\mathbf{z}(t)
\end{aligned} \tag{38}$$

where the filtration $\{\mathcal{F}_t\}$ may be taken to be the natural filtration generated by the random observations, the random Laplacians i.e.,

$$\mathcal{F}_t=\sigma\left(\left\{\{\mathbf{y}_n(s)\}_{n=1}^N,\{\mathbf{L}(s)\}\right\}_{s=0}^{t-1}\right), \tag{39}$$

which is the σ-algebra induced by the observation processes. We use the following inequalities $\forall t\geq t_1$,

$$\begin{aligned}
&\mathbf{z}^\top(t)\mathbb{E}_{\boldsymbol{\theta}^*}\left[\left(\widetilde{\mathbf{L}}(t)\otimes\mathbf{I}_M\right)^2\right]\mathbf{z}(t)\leq\frac{c_5\left\|\mathbf{z}_{\mathcal{C}^\perp}\right\|^2}{(t+1)^{1+\epsilon}}\\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)^2\mathbf{z}(t)\overset{(q1)}{\leq}\lambda_N^2(\overline{\mathbf{L}})||\mathbf{z}_{C^\perp}(t)||^2;\\
&\mathbf{z}^\top(t)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\geq c_1||\mathbf{z}(t)||^2\overset{(q2)}{\geq}0;\\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{z}(t)\overset{(q3)}{\geq}\lambda_2(\overline{\mathbf{L}})\left\|\mathbf{z}_{C^\perp}(t)\right\|^2;\\
&\mathbf{z}^\top(t)\left(\overline{\mathbf{L}}\otimes\mathbf{I}_M\right)\mathbf{G}\left(\mathbf{x}(t)\right)\mathbf{R}^{-1}\left(\mathbf{f}\left(\mathbf{x}(t)\right)-\mathbf{f}\left(\mathbf{1}_N\otimes\boldsymbol{\theta}\right)\right)\\
&\overset{(q4)}{\leq}c_2\left\|\mathbf{z}(t)\right\|^2,
\end{aligned} \tag{40}$$

for c_1 as defined in Assumption M7 and a positive constant c_2. Inequalities $(q1)$ and $(q4)$ follow from the properties of the Laplacian. Inequality $(q2)$ follows from Assumption M7 and

$(q4)$ follows from Assumption M6 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 $\|\mathbf{G}(\mathbf{x}(t))\| \leq \max_{n=1,\cdots,N} k_n$. We also have

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\mathbf{G}(\mathbf{x}(t))\,\mathbf{R}^{-1}(\mathbf{y}(t)-\mathbf{f}(\mathbf{1}_N\otimes\boldsymbol{\theta}))\right\|^2\right] \leq c_4, \tag{41}$$

for some constant $c_4 > 0$. In (41), we use the fact that the noise process under consideration has finite covariance. We also use the fact that $\|\mathbf{G}(\mathbf{x}(t))\| \leq \max_{n=1,\cdots,N} k_n$, which in turn follows from Assumption M5. We further have that,

$$\left\|\mathbf{G}(\mathbf{x}(t))\,\mathbf{R}^{-1}(\mathbf{f}(\mathbf{x}(t))-\mathbf{f}(\mathbf{1}_N\otimes\boldsymbol{\theta}))\right\|^2 \leq c_3\|\mathbf{z}(t)\|^2, \tag{42}$$

where $c_3 > 0$ is a constant. It is to be noted that (42) follows from the Lipschitz continuity in Assumption M5 and the result that $\|\mathbf{G}(\mathbf{x}(t))\| \leq \max_{n=1,\cdots,N} k_n$. Using the inequalities derived in (40), we have,

$$\begin{aligned}\mathbb{E}[V(t+1)|\mathcal{F}_t] &\leq (1+c_8\alpha^2(t))V(t)\\ &- c_9\left(\beta_t - \frac{c_5}{(t+1)^{\tau_1+\epsilon}}\right)\|\mathbf{z}_{\mathcal{C}^\perp}\|^2 + c_6\alpha^2(t).\end{aligned} \tag{43}$$

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 $\forall t \geq t_2$,

$$\mathbb{E}_{\theta^*}[V(t+1)|\mathcal{F}_t] \leq (1+\alpha^2(t))V(t) + \widehat{\alpha}_t^2, \tag{44}$$

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

$$W(t) = \left(\prod_{s=t}^{\infty}(1+\alpha_s^2)\right)V_2(t) + \sum_{s=t}^{\infty}\widehat{\alpha}_s^2,\ \forall t \geq t_2. \tag{45}$$

By (45), it can be shown that $\{W(t)\}$ satisfies,

$$\mathbb{E}_{\theta^*}[W(t+1)|\mathcal{F}_t] \leq W(t). \tag{46}$$

Hence, $\{W(t)\}$ is a non-negative super martingale and converges a.s. to a bounded random variable W^* as $t \to \infty$. It then follows from (45) that $V(t) \to W^*$ as $t \to \infty$. Thus, we conclude that the sequences $\{\mathbf{x}_n(t)\}$ are bounded for all n. □

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

Lemma A2 (Lemma 4.1 in [21])**.** *Consider the scalar time-varying linear system*

$$u(t+1) \leq (1-r_1(t))u(t) + r_2(t), \tag{47}$$

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

$$\frac{a_1}{(t+1)^{\delta_1}} \leq r_1(t) \leq 1 \tag{48}$$

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

$$r_2(t) \leq \frac{a_2}{(t+1)^{\delta_2}} \tag{49}$$

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

$$\lim_{t\to\infty}(t+1)^{\delta_0}u(t) = 0, \tag{50}$$

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 as in the proof of Lemma A1, for t large enough

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] &\leq \left(1 - 2c_1\alpha_t + c_7\alpha_t^2\right) V(t) + c_6\alpha_t^2 \\ &\leq V(t) + c_6\alpha_t^2, \end{aligned} \tag{51}$$

as for t large enough, $-2c_1\alpha_t + c_7\alpha_t^2 < 0$. Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}V_1(t) &= V(t) + c_6 \sum_{s=t}^{\infty} \alpha_s^2 \\ &= V(t) + c_8 \sum_{s=t}^{\infty} (t+1)^{-2}, \end{aligned} \tag{52}$$

for appropriately chosen positive constant c_8.Since, $\{(t+1)^{-2}\}$ 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 (52), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (51), we have that,

$$\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] \leq (1 - c_1\alpha_t)\, \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_9(t+1)^{-2}, \tag{53}$$

for $t \geq t_1$. The sequence $\{V(t)\}$ then falls under the purview of Lemma A2, and we have $\mathbb{E}_{\boldsymbol{\theta}}[V(t)] \to 0$ as $t \to \infty$. Finally, by Fatou's Lemma, where we use the non-negativity of the sequence $\{V(t)\}$, we conclude that

$$0 \leq \mathbb{E}_{\boldsymbol{\theta}}[V^*] \leq \liminf_{t \to \infty} \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = 0, \tag{54}$$

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

We will use the following approximation result (Lemma A3) and the generalized convergence criterion (Lemma A4) for the proof of Theorem 41.

Lemma A3 (Lemma 4.3 in [9])**.** *Let $\{b_t\}$ be a scalar sequence satisfying*

$$b_{t+1} \leq \left(1 - \frac{c}{t+1}\right) b_t + d_t(t+1)^{-\tau}, \tag{55}$$

where $c > \tau, \tau > 0$, and the sequence d_t is summable. Then, we have,

$$\limsup_{t \to \infty}\ (t+1)^{\tau} b_t < \infty. \tag{56}$$

Lemma A4 (Lemma 10 in [8])**.** *Let $\{J(t)\}$ be an $\mathbb{R}$-valued $\{\mathcal{F}_{t+1}\}$-adapted process such that $\mathbb{E}\left[J(t)|\mathcal{F}_t\right] = 0$ 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 42. Proceeding as in proof of Theorem 41, we have, for t large enough

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)|\mathcal{F}_t] &\leq \left(1 - 2c_1\alpha_t + c_7\alpha_t^2\right) V(t) + c_6\alpha_t^2 \\ &\leq V(t) + c_6\alpha_t^2, \end{aligned} \tag{57}$$

Now, consider the $\{\mathcal{F}_t\}$-adapted process $\{V_1(t)\}$ defined as follows

$$\begin{aligned}V_1(t) &= V(t) + c_6 \sum_{s=t}^{\infty} \alpha_s^2 \\ &= V(t) + c_8 \sum_{s=t}^{\infty} (t+1)^{-2}, \end{aligned} \tag{58}$$

for appropriately chosen positive constant c_8.Since, $\{(t+1)^{-2}\}$ 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 (52), we also have that $\{V(t)\}$ converges to a non-negative finite random variable V^*. Finally, from (57), we have that,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] \leq (1-c_1\alpha_t)\,\mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_8(t+1)^{-2}\\ &\Rightarrow E_{\boldsymbol{\theta}}[V(t+1)] \leq (1-c_1\alpha_t)\,\mathbb{E}_{\boldsymbol{\theta}}[V(t)] + c_{10}\alpha_t(t+1)^{-1}\end{aligned} \tag{59}$$

for $t \geq t_1$. The summability of $\{\alpha_t\}$ in conjunction with assumption M8 ensures that the sequence $\{V(t)\}$ then falls under the purview of Lemma A3, and we have

$$\begin{aligned}&\limsup_{t\to\infty}(t+1)\mathbb{E}_{\boldsymbol{\theta}}[V(t+1)] < \infty\\ &\Rightarrow \mathbb{E}_{\boldsymbol{\theta}}[V(t)] = O\left(\frac{1}{t}\right).\end{aligned} \tag{60}$$

Furthermore, from (58), we also have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}[V_1(t)] \leq \mathbb{E}_{\boldsymbol{\theta}}[V(t)] + \frac{c_6\pi^2}{6}\\ &\Rightarrow \mathbb{E}_{\boldsymbol{\theta}}[\|\mathbf{x}_n(t)-\boldsymbol{\theta}\|^2] = O\left(\frac{1}{t}\right).\end{aligned} \tag{61}$$

It is to be noted that 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 (61). □

References

1. Dragana Bajović, José M. F Moura, João Xavier, and Bruno Sinopoli. Distributed inference over directed networks: Performance limits and optimal design. *arXiv preprint arXiv:1504.07526*, 2015.
2. B. Bollobas. *Modern Graph Theory*. Springer Verlag, New York, NY, 1998.
3. Paolo Braca, Stefano Marano, and Vincenzo Matta. Enforcing consensus while monitoring the environment in wireless sensor networks. *IEEE Transactions on Signal Processing*, 56(7):3375–3380, 2008.
4. Federico S Cattivelli and Ali H Sayed. Diffusion lms strategies for distributed estimation. *IEEE Transactions on Signal Processing*, 58(3):1035–1048, 2010.
5. Jie Chen, Cédric Richard, and Ali H Sayed. Multitask diffusion adaptation over networks. *IEEE Transactions on Signal Processing*, 62(16):4129–4144, 2014.
6. Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
7. Paolo Di Lorenzo and Ali H Sayed. Sparse distributed learning based on diffusion adaptation. *IEEE Transactions on signal processing*, 61(6):1419–1433, 2013.
8. Lester E Dubins and David A Freedman. A sharper form of the Borel-Cantelli lemma and the strong law. *The Annals of Mathematical Statistics*, pages 800–807, 1965.
9. V. Fabian. Stochastic approximation of minima with improved asymptotic speed. *The Annals of Mathematical Statistics*, 37(1):191–200, Feb 1967.
10. V. Fabian. On asymptotically efficient recursive estimation. *The Annals of Statistics*, 6(4):854–866, Jul. 1978.
11. R.Z. Has'minskij. Sequential estimation and recursive asymptotically optimal procedures of estimation and observation control. In *Proc. Prague Symp. Asymptotic Statist.*, volume 1, pages 157–178, Charles Univ., Prague, 1974.
12. Christina Heinze, Brian McWilliams, and Nicolai Meinshausen. Dual-loco: Distributing statistical estimation using random projections. In *Artificial Intelligence and Statistics*, pages 875–883, 2016.
13. M. D. Ilic' and J. Zaborszky. *Dynamics and Control of Large Electric Power Systems*. Wiley, 2000.

14. D. Jakovetic, J. Xavier, and J. M. F. Moura. Cooperative convex optimization in networked systems: Augmented Lagrangian algorithms with directed gossip communication. *IEEE Transactions on Signal Processing*, 59(8):3889–3902, August 2011.
15. Dusan Jakovetic, Dragana Bajovic, Natasa Krejic, and Natasa Krklec Jerinkic. Distributed gradient methods with variable number of working nodes. *IEEE Trans. Signal Processing*, 64(15):4080–4095, 2016.
16. Robert I Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.
17. S. Kar, J. M. F. Moura, and H. V. Poor. Distributed linear parameter estimation: asymptotically efficient adaptive strategies. *SIAM J. on Control Optim.*, 51(3):2200 – 2229, May 2013.
18. S. Kar, J. M. F. Moura, and K. Ramanan. Distributed parameter estimation in sensor networks: nonlinear observation models and imperfect communication. *IEEE Transactions on Information Theory*, 58(6):3575 – 3605, June 2012.
19. Soummya Kar and José M. F Moura. Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):674–690, 2011.
20. Soummya Kar and José M. F Moura. Asymptotically efficient distributed estimation with exponential family statistics. *IEEE Transactions on Information Theory*, 60(8):4811–4831, 2014.
21. Soummya Kar, José M. F Moura, and H Vincent Poor. Distributed linear parameter estimation: Asymptotically efficient adaptive strategies. *SIAM Journal on Control and Optimization*, 51(3):2200–2229, 2013.
22. Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *arXiv preprint arXiv:1701.03961*, 2017.
23. Jinchao Li and Ali H Sayed. Modeling bee swarming behavior through diffusion adaptation with asymmetric information sharing. *EURASIP Journal on Advances in Signal Processing*, 2012(1):18, 2012.
24. Qiang Liu and Alexander T Ihler. Distributed estimation, information loss and exponential families. In *Advances in Neural Information Processing Systems*, pages 1098–1106, 2014.
25. C. G. Lopes and A. H. Sayed. Diffusion least-mean squares over adaptive networks: Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7):3122–3136, July 2008.
26. Chenxin Ma, Virginia Smith, Martin Jaggi, Michael Jordan, Peter Richtarik, and Martin Takac. Adding vs. averaging in distributed primal-dual optimization. In *International Conference on Machine Learning*, pages 1973–1982, 2015.
27. Chenxin Ma and Martin Takáč. Partitioning data on features or samples in communication-efficient distributed optimization? *arXiv preprint arXiv:1510.06688*, 2015.
28. Gonzalo Mateos and Georgios B Giannakis. Distributed recursive least-squares: Stability and performance analysis. *IEEE Transactions on Signal Processing*, 60(7):3740–3754, 2012.
29. Gonzalo Mateos, Ioannis D Schizas, and Georgios B Giannakis. Performance analysis of the consensus-based distributed lms algorithm. *EURASIP Journal on Advances in Signal Processing*, 2009:68, 2009.
30. A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48 61, Jan. 2009.
31. Angelia Nedić, Alex Olshevsky, and César A Uribe. Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs. *arXiv preprint arXiv:1410.1977*, 2014.
32. J. Pfanzagl. Asymptotic optimum estimation and test procedures. In *Proceedings of the Prague Symposium on Asymptotic Statistics*, volume 1, Sept. 3 - 6 1973.
33. S. S. Ram, A. Nedic, and V. V. Veeravalli. Incremental stochastic subgradient algorithms for convex optimization. *SIAM Journal on Optimization*, 20(2):691–717, June 2009.
34. S.S. Ram, A. Nedić, and V.V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.
35. S.S. Ram, V.V. Veeravalli, and A. Nedic. Distributed and recursive parameter estimation in parametrized linear state-space models. *to appear in IEEE Transactions on Automatic Control*, 55(2):488– 492, February 2010.
36. A. K. Sahu, D. Jakovetic, and S. Kar. CREDO: A communication-efficient distributed estimation algorithm. 2018. submitted to IEEE International Symposium on Information Theory, ISIT 2018.

37. Anit Kumar Sahu, Dusan Jakovetic, and Soummya Kar. Communication optimality tradeoffs for distributed estimation. *arXiv preprint arXiv:1801.04050*, 2018.
38. Anit Kumar Sahu and Soummya Kar. Distributed sequential detection for Gaussian shift-in-mean hypothesis testing. *IEEE Transactions on Signal Processing*, 64(1):89–103, 2016.
39. Anit Kumar Sahu, Soummya Kar, José MF Moura, and H Vincent Poor. Distributed constrained recursive nonlinear least-squares estimation: Algorithms and asymptotics. *IEEE Transactions on Signal and Information Processing over Networks*, 2(4):426–441, 2016.
40. D.J. Sakrison. Efficient recursive estimation; application to estimating the parameters of a covariance function. *International Journal of Engineering Science*, 3(4):461–483, 1965.
41. C.J. Stone. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics*, 3(2):267–284, Mar. 1975.
42. Konstantinos Tsianos, Sean Lawlor, and Michael G Rabbat. Communication/computation tradeoffs in consensus-based distributed optimization. In *Advances in neural information processing systems*, pages 1943–1951, 2012.
43. Konstantinos I Tsianos, Sean F Lawlor, Jun Ye Yu, and Michael G Rabbat. Networked optimization with adaptive communication. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 579–582. IEEE, 2013.
44. Zifeng Wang, Zheng Yu, Qing Ling, Dimitris Berberidis, and Georgios B Giannakis. Decentralized rls with data-adaptive censoring for regressions over large-scale networks. *arXiv preprint arXiv:1612.08263*, 2016.
45. Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pages 592–617, 2013.

Figure legend

The list of Figures is as follows.

Fig. 1. Network Deployment of 10 agents.

Fig. 2. *Comparison of the proposed and benchmark estimators in terms of relative MSE: Number of Iterations*. The blue line represents the benchmark, while the red line represents the proposed estimator.

Fig. 3. *Comparison of the proposed and benchmark estimators in terms of relative MSE: Communication Cost Per Node*. The blue line represents the benchmark, while the red line represents the proposed estimator.