$\mathcal{CREDO}$: A Communication-Efficient Distributed Estimation Algorithm

Anit Kumar Sahu
Department of ECE
Carnegie Mellon University
Pittsburgh
Email: anits@andrew.cmu.edu

Dusan Jakovetic
Department of Mathematics and Informatics
Faculty of Sciences, University of Novi Sad
21000 Novi Sad, Serbia
Email: djakovet@uns.ac.rs

Soummya Kar
Department of ECE
Carnegie Mellon University
Pittsburgh
Email: soummyak@andrew.cmu.edu

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. This paper presents Communication efficient REcursive Distributed estimatiOn algorithm, $\mathcal{CREDO}$ for networked multi-agent systems. $\mathcal{CREDO}$ caters to situations in which the agents collaboratively estimate a vector parameter by assimilating their latest sensed information and estimates from their time-varying neighborhood worker nodes over a (possibly sparse) communication graph, while adhering to a frugal communication scheme. The underlying inter-agent communication protocol is randomized and makes communications be increasingly (probabilistically) sparse. $\mathcal{CREDO}$ may be designed to achieve at each agent a $\Theta\left(\mathcal{C}_t^{-2+\zeta}\right)$ decay of the mean square error ($\zeta > 0$, arbitrarily small) with respect to per-node communication cost $\mathcal{C}_t$, which significantly improves over the existing $\Theta\left(\mathcal{C}_t^{-1}\right)$ rates. Simulations demonstrate $\mathcal{CREDO}$'s communication efficiency.**

A full version of this paper is accessible at: http://www.isit2018.org/

1. Introduction

This paper presents $\mathcal{CREDO}$, a **C**ommunication efficient **RE**cursive **D**istributed estimati**O**n algorithm for networked systems. More precisely, we consider the systems where N nodes, interconnected in a generic (possibly sparse) network, acquire noisy low-dimensional linear transformations of the unknown vector parameter $\boldsymbol{\theta}$ in a time-sequential yet streaming fashion over a slotted time frame. For the described observation model, we focus on *recursive distributed estimators*. Here, recursive means that each node continuously produces estimates of the underlying parameter at each time t. By distributed, we mean the estimators where each worker[1], at each time t, updates its local estimate by simultaneously processing its latest sensed information and neighborhood information. Recursive distributed estimators are relevant, e.g., with Internet of Things (IoT) applications. Therein, a system typically involves a heterogeneous network of entities without a central coordinator, where entities have localized knowledge and can exchange information among each other through an arbitrary pre-specified network. Further, in many IoT applications, data arrives in a steaming fashion, like in continuous monitoring of a large scale industrial plant or in continuous monitoring of a smart building.

This paper focuses on improving *communication efficiency* of recursive distributed estimators. This is highly relevant, e.g., with IoT applications, as the IoT devices have limited communication capabilities owing to on board power constraints and harsh environments. Among the existing recursive distributed estimators, including, e.g., consensus+innovations and diffusion methods [1]–[7], and related methods , e.g., [8]–[12], the best achievable $\Theta(1/t)$ rates of the mean square error have been already established. This also means that the best possible $\Theta(1/n_t)$ rate in terms of the number of processed per-node data samples n_t have been achieved, as with the works above each node typically processes one sample at a time. However, the existing methods achieve at best an $O(1/\mathcal{C}_t)$ communication rate, where C_t is the per-node (expected) communication cost.

In this paper, we present the algorithm $\mathcal{CREDO}$ which significantly improves the previously established communication rates to $\Theta(1/\mathcal{C}_t^{2-\zeta})$, where $\zeta > 0$ is arbitrarily small. $\mathcal{CREDO}$ utilizes a simple randomized communication protocol where neighborhood communication takes place with a certain probability that decays to zero as t grows, with the decay at a carefully crafted rate. Interestingly, despite sparsified communications, the method still achieves the optimal $\Theta(1/t)$ decay of MSE with time t, which then translates into significantly improved communication rates.

We now briefly review existing work, in addition to the above mentioned references [1]–[12]. In the context of distribution stochastic optimization, several highly scalable methods (see, for example [13]–[15]) for master-worker or similar types of architectures have been proposed which exhibit impressive performance in platforms such as Mapreduce and Spark. Such classes of distributed architectures are characterized by the presence of a central coordinator; this may be undesirable in systems and applications where the entire data is not centrally available and, is sensed in a streaming fashion and distributed across the worker nodes – the setup considered here. Communication efficient distributed recursive algorithms in the context of distributed optimization with no central node, where data is available a priori and is not collected in a streaming fashion has been addressed in [16]–[18] through increasingly sparse communication, adaptive communication scheme, and selective activation of nodes, respectively. However, the explicit characterization of the performance metric for instance MSE, in terms of the communication cost has not been addressed in the aforementioned references. Finally,

[1]We use nodes, agents and workers interchangeably throughout the paper.

this paper is a companion of [19] on the $\mathcal{CREDO}$ method. With respect to [19], the current paper focuses on the single time scale approximation variant of *CREDO*, while paper [19] considers on the more general two time scale algorithm version. Further, we include here novel numerical experiments, both on synthetic and real data. The single time scale method here is a special case of the two time scale method in [19]; hence the proofs of the convergence results presented here in Section 3 (b) are the special case of the corresponding results in [19].

2. Sensing Model and Preliminaries

The network consists of N deployed workers. Every worker n at time index t makes a noisy observation $\mathbf{y}_n(t)$, a noisy function of $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^M$. Formally the observation model for the n-th worker is given by,

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

where $\mathbf{H}_n \in \mathbb{R}^{M_n \times M}$ is the sensing matrix, where $M_n < M$, $\{\mathbf{y}_n(t)\} \in \mathbb{R}^{M_n}$ is the observation sequence for the n-th worker and $\{\gamma_n(t)\}$ is a zero mean temporally independent and identically distributed (i.i.d.) noise sequence at the n-th worker with nonsingular covariance $\boldsymbol{\Sigma}_n$, where $\boldsymbol{\Sigma}_n \in \mathbb{R}^{M_n \times M_n}$. The noise processes are independent across different workers. We state an assumption on the noise processes before proceeding further.

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 sensing matrix and the noise covariances at the worker nodes. Each worker node is interested in reconstructing the true underlying parameter $\boldsymbol{\theta}$. We assume a worker node is aware only of its local observation model and hence does not know about the observation matrix and noise processes of other worker nodes.

A. Preliminaries: Oracle and Distributed Estimation

We next give preliminaries on oracle and distributed estimation.

Oracle Estimation:

In the setup described above in (1), a hypothetical oracle node having access to all the sensed data at all nodes across all times would update its estimate in a recursive fashion as follows:

$$\mathbf{x}_c(t+1) = \mathbf{x}_c(t) + \underbrace{\frac{a}{t+1}\sum_{n=1}^{N}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{H}_n\mathbf{x}_c(t)\right)}_{\text{Global Innovation}},$$

where a is a positive constant. It is well known from standard stochastic approximation results that the sequence $\{\mathbf{x}_c(t)\}$ is strongly consistent and that the mean square error (MSE) $\mathbb{E}_\theta[\|\mathbf{x}_c(t) - \boldsymbol{\theta}\|^2] = \Theta(1/t)$. However, such an oracle based scheme owing to its requirement of instantaneous access to the sensed data across the network at all times is rendered infeasible in distributed multi-worker setting with time-varying sparse inter-worker interactions.

Distributed Estimation:

We next consider distributed estimators where N worker nodes are interconnected in a generic network. Prior work typically assumes either a static connected network or a random network model described by a sequence of i.i.d. Laplacian matrices, such that the network is connected on average; see, e.g., [1]. The lack of global model information at each worker, makes it necessary to communicate at a properly crafted rate. In order to ensure that every worker is able to reconstruct the parameter, a communication rate needs to be crafted so as to ensure information flow among the worker nodes. In the case of setups with a restricted communication topology, the information loss at a node is compensated by incorporating an agreement or consensus potential into their updates which is then incorporated as follows:

$$\mathbf{x}_n(t+1) = \mathbf{x}_n(t) - \underbrace{\frac{b}{(t+1)}\sum_{l\in\Omega_n(t)}\left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{Neighborhood Consensus}}$$
$$+ \underbrace{\frac{a}{t+1}\mathbf{H}_n^\top\boldsymbol{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{H}_n\mathbf{x}_n(t)\right)}_{\text{Local Innovation}}, \tag{2}$$

where $\Omega_n(t)$ represents the (possibly random) neighborhood set of worker n at time t and a, b are appropriately chosen positive constants. It has been shown in previous work that under appropriate conditions (see, for example [20]), the estimate sequence $\{\mathbf{x}_n(t)\}$ is strongly consistent and the MSE decays as $\Theta(1/t)$.

Communication Efficiency

Define the communication cost $\mathcal{C}_t$ to be the expected per-node number of transmissions up to iteration t, i.e.,

$$\mathcal{C}_t = \mathbb{E}\left[\sum_{s=0}^{t-1}\mathbb{I}_{\{\textit{node } C \textit{ transmits at } s\}}\right], \tag{3}$$

where $\mathbb{I}_A$ represents the indicator of event A. The communication cost $\mathcal{C}_t$ for both the oracle estimator and the distributed estimators in [1], [2], [4], [6], [11], [12], [20] comes out to be $\mathcal{C}_t = \Theta(t)$, where we note that the time index t also matches the number of per node samples collected till time t. In other words, we have MSE decaying as $\Theta\left(\frac{1}{\mathcal{C}_t}\right)$. Both the paradigms achieve an order-optimal MSE decay rate $\Theta(1/t)$ in terms of the number of observations t. The $\Theta(1/t)$ decay rate of MSE in terms of the number of per-node data samples in the class of recursive estimators can not be improved. However, we show that $\mathcal{CREDO}$ improves the MSE decay in terms of communication cost $\mathcal{C}_t$ without compromising the $\Theta(1/t)$ decay rate.

3. $\mathcal{CREDO}$: A communication efficient distributed recursive estimator

A. The algorithm

In this section, we present the $\mathcal{CREDO}$ algorithm. The algorithm has a carefully controlled time decaying communication

rate-based protocol. Intuitively, we basically exploit the idea that, once the information flow starts in the network and a worker node is able to accumulate sufficient information about the parameter of interest, the need to communicate goes down with time. Technically speaking, for each node n, at every time t, we introduce a binary random variable $\psi_{n,t}$, where

$$\psi_{n,t} = \begin{cases} \rho_t & \textit{with probability } \zeta_t \\ 0 & \textit{else}, \end{cases} \tag{4}$$

where $\psi_{i,t}$'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

$$\rho_t = \frac{\rho_0}{(t+1)^{\epsilon/2}}, \zeta_t = \frac{\zeta_0}{(t+1)^{(1/2-\epsilon/2)}}, \tag{5}$$

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

$$\beta_t = (\rho_t \zeta_t)^2 = \frac{\beta_0}{(t+1)}. \tag{6}$$

In order to describe the update rule for $\mathcal{CREDO}$, we first define the network model and introduces the related quantities needed in the sequel. The inter-worker communication network is modeled as an *undirected* simple connected graph $G = (V, E)$, with $V = [1 \cdots N]$ and E denoting the set of nodes and communication links. The neighborhood of node n is given by $\Omega_n = \{l \in V \,|\, (n,l) \in E\}$. 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. The graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is positive definite, with eigenvalues ordered as $0 = \lambda_1(\mathbf{L}) \leq \lambda_2(\mathbf{L}) \leq \cdots \leq \lambda_N(\mathbf{L})$, where $\mathbf{D}$ is given by $\mathbf{D} = \text{diag}\,(d_1 \cdots d_N)$. Moreover, as the graph is connected, we have $\lambda_2(\mathbf{L}) > 0$. 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:

$$\mathbf{L}_{i,j}(t) = \begin{cases} -\psi_{i,t}\psi_{j,t} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l \neq i} \psi_{i,t}\psi_{l,t} & i = j. \end{cases} \tag{7}$$

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 (4)-(7) not only decays the weight assigned to the links over time, but also, decays the probability of the existence of a link. Such a design is consistent with frameworks where the working nodes have finite power and hence not only the number of communications, but also, the quality of the communication decays over time. We have, for $\{i,j\} \in E$:

$$\begin{aligned} \mathbb{E}\left[\mathbf{L}_{i,j}(t)\right] &= -\left(\rho_t \zeta_t\right)^2 = -\beta_t = -\frac{c_3}{(t+1)} \\ \mathbb{E}\left[\mathbf{L}^2_{i,j}(t)\right] &= \left(\rho_t^2 \zeta_t\right)^2 = \frac{c_4}{(t+1)^{1+\epsilon}}. \end{aligned} \tag{8}$$

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

$$\textit{Var}\left(\mathbf{L}_{i,j}(t)\right) = \frac{\beta_0 \rho_0^2}{(t+1)^{1+\epsilon}} - \frac{a^2}{(t+1)^2}. \tag{9}$$

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

$$\mathbb{E}\left[\left\|\widetilde{\mathbf{L}}(t)\right\|^2\right] \leq N^2 \mathbb{E}\left[\widetilde{\mathbf{L}}^2_{i,j}(t)\right] = \frac{N^2 \beta_0 \rho_0^2}{(t+1)^{1+\epsilon}} - \frac{N^2 a^2}{(t+1)^2}, \tag{10}$$

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

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l \neq i} L_{i,l} & i = j. \end{cases} \tag{11}$$

We formalize the assumptions on the inter-worker communication graph and global observability.

Assumption M2. *We require the following global observability condition. The matrix* $\mathbf{G} = \sum_{n=1}^{N} \mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1} \mathbf{H}_n$ *is full rank.*

Assumption M2 is crucial for our distributed setup. It is to be noted that such an assumption is needed for even a setup with a centralized node which has access to all the data samples at each of the worker nodes at each time.

Assumption M3. *The inter-worker 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.*

Technically speaking, the communication graph need not be connected at all times. Hence, at any given time, only a few of the possible links could be active. The connectedness in average basically ensures that over time, the information from each worker node in the graph reaches other worker nodes over time in a symmetric fashion and thus ensuring information flow. Assumption M3 ensures that $\overline{\mathbf{L}}(t)$ is connected at all times as $\overline{\mathbf{L}}(t) = \beta_t \overline{\mathbf{L}}$. With the communication protocol established, we present the $\mathcal{CREDO}$ update, where every node n generates an estimate sequence $\{\mathbf{x}_n(t)\}$, with $\mathbf{x}_n(t) \in \mathbb{R}^M$, in the following way:

$$\begin{aligned} \mathbf{x}_n(t+1) = \mathbf{x}_n(t) &- \underbrace{\sum_{l \in \Omega_n} \psi_{n,t}\psi_{l,t}\left(\mathbf{x}_n(t) - \mathbf{x}_l(t)\right)}_{\text{Neighborhood Consensus}} \\ &+ \underbrace{\alpha_t \mathbf{H}_n^\top \mathbf{\Sigma}_n^{-1}\left(\mathbf{y}_n(t) - \mathbf{H}_n \mathbf{x}_n(t)\right)}_{\text{Local Innovation}}. \end{aligned} \tag{12}$$

Here, Ω_n denotes the neighborhood of node n with respect to the network encapsulated by $\overline{\mathbf{L}}$ and α_t is the innovation gain sequence which is given by $\alpha_t = a/(t+1)$. It is to be noted that a node n can send and receive information in its neighborhood at time t, when $\psi_{n,t} \neq 0$. At the same time, 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 $\psi_{l,t} \neq 0$. We next formalize an assumption on the innovation gain sequence $\{\alpha_t\}$ before proceeding further.

Assumption M4. *Let $\lambda_{min}(\cdot)$ denote the smallest eigenvalue. We require that a satisfies*[2]*, $a \min\{\lambda_{min}(\mathbf{\Gamma}), \lambda_{min}\left(\overline{\mathbf{L}} \otimes \mathbf{I}_M + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top\right), \beta_0^{-1}\} \geq 1$, where $\otimes$ denotes the Kronecker product.*

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

B. Convergence results

We now present the main results of the presented algorithm $\mathcal{CREDO}$, while the proofs can be found in [19]. The first result concerns with the consistency of the estimate sequence $\{\mathbf{x}_n(t)\}$.

Theorem 3.1. *Let assumptions M1-M4 hold. Consider the sequence $\{\mathbf{x}_n(t)\}$ generated by* (12) *at each worker n. Then, for each n, we have*

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\lim_{t\to\infty} \mathbf{x}_n(t) = \boldsymbol{\theta}\right) = 1. \tag{13}$$

Theorem 3.1 asserts that the estimate sequence generated by $\mathcal{CREDO}$ at any worker n is strongly consistent, i.e., $\mathbf{x}_n(t) \to \boldsymbol{\theta}$ almost surely (a.s.) as $t \to \infty$, despite the increasingly sparse communication protocol utilized. We now state a main result concerning the MSE communication rate for $\mathcal{CREDO}$.

Theorem 3.2. *Let the hypothesis of Theorem 3.1 hold. Then, we have,*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2\right] = \Theta\left(\mathcal{C}_t^{-\frac{2}{\epsilon+1}}\right), \tag{14}$$

where $\epsilon < 1$ and is as defined in (5)*.*

$\mathcal{CREDO}$ has communication cost $\mathcal{C}_t = \Theta\left(t^{0.5(1+\epsilon)}\right)$. It can be shown from standard arguments in stochastic approximation that updates with $\beta_t = a(t+1)^{-1-\delta}$ with $\delta > 0$, even though they result in a communication cost of $\mathcal{C}_t = \Theta(t^{0.5(1+\epsilon-\delta)})$, lead to an algorithm that does not generate estimate sequences which converge to $\boldsymbol{\theta}$.

4. Simulation Experiments

This section corroborates our theoretical findings through simulation examples.

A. Synthetic Data

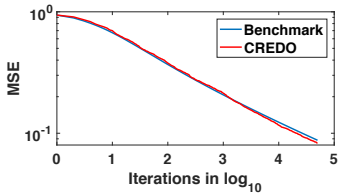
We compare the presented communication-efficient distributed estimator, $\mathcal{CREDO}$, with the distributed recursive estimator (referred to as *benchmark* hereby) in (2) which utilizes all inter-neighbor communications at all times. We consider an undirected graphs with $N = 20$ nodes, with relative degrees[3] of nodes slated at 0.5315. The graph was generated as connected graph instances of the random geometric graph model with radius $r = \sqrt{\ln(N)/N}$. We set $M = 10$ and $M_n = 1$, for all $n = 1, ..., N$, while each node makes a scalar observation at each time t. The noises are Gaussian and are i.i.d. both in time and across nodes and have the covariance matrix equal to $0.25I$. The sampling matrices $\mathbf{H}_n$'s are chosen to be 2-sparse, i.e., every nodes observes a linear combination of two arbitrary entries of the vector parameter. The non-zero entries of the $\mathbf{H}_n$'s are sampled from a standard normal distribution. The sampling matrices $\mathbf{H}_n$'s at the same time satisfy Assumption M2. The benchmark estimator's consensus weight is set to $0.1(t+1)^{-1}$. For $\mathcal{CREDO}$, we set $\rho_t = 0.1(t+1)^{-0.01}$, $\zeta_t = (t+1)^{-0.49}$, i.e., $\epsilon = 0.01$, $\tau_1 = 1$. The Laplacian associated with the benchmark estimator and the expected Laplacian associated with $\mathcal{CREDO}$ are equal , i.e., $\overline{\mathbf{L}} = \mathbf{L}$. The innovation weight is set to $\alpha_t = (3.68(t+20))^{-1}$. Note that all the theoretical results in the paper hold unchanged for the "time-shifted" α_t used here. The purpose of the shift in the innovation potential is to avoid large innovation weights in the initial iterations. As a performance metric, we use the relative MSE estimate averaged across nodes, $\frac{1}{N}\sum_{n=1}^{N} \|\mathbf{x}_n(t) - \boldsymbol{\theta}\|^2 / \|\mathbf{x}_n(0) - \boldsymbol{\theta}\|^2$ further averaged across 50 independent runs. Here, $\mathbf{x}_n(0)$ is node n's initial estimate. With both estimators, at each run, at all nodes, we set $\mathbf{x}_n(0) = 0$. Figure 1a plots the estimated relative MSE versus time t in log-log scale from which we can see that the MSE decay of $\mathcal{CREDO}$ coincides with that of the benchmark estimator. Figure 1b, illustrates the savings in terms of the communication cost made by $\mathcal{CREDO}$ as compared to the benchmark estimator.

B. Real Data

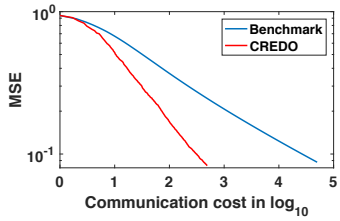
In order to evaluate the performance of $\mathcal{CREDO}$, we ran experiments on Abalone dataset ([21]). For the Abalone dataset (4177 data points, 8 features), we divided the samples into 10 equal parts of 360 points each, with 577 data points as the test set. For the 10 node network, we constructed a random geometric graph. We added Gaussian noise to the dependent variable, i.e., the age of Abalone. The training dataset, with respect to the sensing model (1), has dynamic regressors (a regressor here corresponds to a feature vector of one data point), i.e, time-varying $\mathbf{H}_n$'s for each agent n. Thus, we perform a pre-processing step where we average the training data points' regressors at each node to obtain an averaged $\overline{\mathbf{H}}_n$, which is then subsequently used at every iteration t in the update (12). A consistency check is done by ensuring that $\sum_{n=1} \overline{\mathbf{H}}_n^\top \mathbf{\Sigma}_n^{-1} \overline{\mathbf{H}}_n$ is invertible and thus global observability holds. As the number of data points at each node are the same, we sample along iterations t data points at each node without replacement.The test errors obtained in the abalone dataset is 0.06 of the initial test error. In figures 1c and 1d, we plot the evolution of the test error for the abalone dataset as a function of the number of iterations and the communication cost. It can be seen that while $\mathcal{CREDO}$ matches the final test error of that of the benchmark algorithm, it requires almost thrice as less number of communications.

[2]Note that, $\mathbf{\Gamma}$ and $\overline{\mathbf{L}} \otimes \mathbf{I}_M + \mathbf{G}_H \mathbf{\Sigma}^{-1} \mathbf{G}_H^\top$ are positive definite matrices.

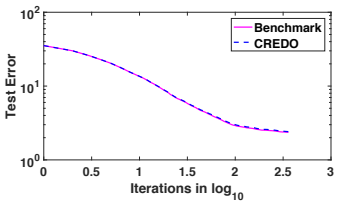
[3]Relative degree is the ratio of the number of links in the graph to the number of possible links in the graph.



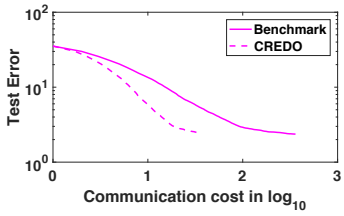
(a) Comparison of relative MSE: Number of Iterations



(b) Comparison of relative MSE: Communication cost per node



(c) Abalone Test Error: Number of Iterations



(d) Abalone Test Error: Communication cost per node

Fig. 1: Comparison of the $\mathcal{CREDO}$ and benchmark estimator

5. Conclusion

In this paper, we have presented a communication efficient distributed recursive estimation scheme $\mathcal{CREDO}$, for which we established strong consistency of the estimation sequence. The communication efficiency of $\mathcal{CREDO}$ was characterized in terms of the dependence of the MSE on the communication cost. To be specific, we have established that the MSE of $\mathcal{CREDO}$ can be as good as $\Theta\left(\mathcal{C}_t^{-2+\epsilon}\right)$, where $\zeta > 0$ and ζ is arbitrarily small. Future research directions include coming up with communication schemes, which are adaptive in terms of the connectivity of a node, and local decision making in terms of whether to communicate or not based on neighborhood information.

References

[1] S. Kar and J. M. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 674–690, 2011.

[2] F. S. Cattivelli and A. H. Sayed, "Diffusion lms strategies for distributed estimation," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1035–1048, 2010.

[3] D. Bajović, J. M. F. Moura, J. Xavier, and B. Sinopoli, "Distributed inference over directed networks: Performance limits and optimal design," *arXiv preprint arXiv:1504.07526*, 2015.

[4] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3122–3136, July 2008.

[5] A. K. Sahu and S. Kar, "Distributed sequential detection for Gaussian shift-in-mean hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 89–103, 2016.

[6] 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*, vol. 59, no. 8, pp. 3889–3902, August 2011.

[7] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4129–4144, 2014.

[8] A. Nedić, A. Olshevsky, and C. A. Uribe, "Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs," *arXiv preprint arXiv:1410.1977*, 2014.

[9] S. Ram, A. Nedić, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.

[10] P. Braca, S. Marano, and V. Matta, "Enforcing consensus while monitoring the environment in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 3375–3380, 2008.

[11] S. Ram, V. Veeravalli, and A. Nedic, "Distributed and recursive parameter estimation in parametrized linear state-space models," *to appear in IEEE Transactions on Automatic Control*, vol. 55, no. 2, pp. 488– 492, February 2010.

[12] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, June 2009.

[13] Y. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression," in *Conference on Learning Theory*, 2013, pp. 592–617.

[14] C. Heinze, B. McWilliams, and N. Meinshausen, "Dual-loco: Distributing statistical estimation using random projections," in *Artificial Intelligence and Statistics*, 2016, pp. 875–883.

[15] C. Ma, V. Smith, M. Jaggi, M. Jordan, P. Richtarik, and M. Takac, "Adding vs. averaging in distributed primal-dual optimization," in *International Conference on Machine Learning*, 2015, pp. 1973–1982.

[16] K. Tsianos, S. Lawlor, and M. G. Rabbat, "Communication/computation tradeoffs in consensus-based distributed optimization," in *Advances in neural information processing systems*, 2012, pp. 1943–1951.

[17] K. I. Tsianos, S. F. Lawlor, J. Y. Yu, and M. G. Rabbat, "Networked optimization with adaptive communication," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 579–582.

[18] D. Jakovetic, D. Bajovic, N. Krejic, and N. K. Jerinkic, "Distributed gradient methods with variable number of working nodes." *IEEE Trans. Signal Processing*, vol. 64, no. 15, pp. 4080–4095, 2016.

[19] A. K. Sahu, D. Jakovetic, and S. Kar, "Communication optimality trade-offs in distributed estimation," 2018, arxiv preprint. [Online]. Available: http://users.ece.cmu.edu/~anits/comm_ci_jmlr_prefinal.pdf

[20] S. Kar, J. M. F. Moura, and H. V. Poor, "Distributed linear parameter estimation: Asymptotically efficient adaptive strategies," *SIAM Journal on Control and Optimization*, vol. 51, no. 3, pp. 2200–2229, 2013.

[21] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

Acknowledgment

The work of AKS and SK was supported in part by NSF under grant CCF-1513936. The work of DJ was supported by Ministry of Education, Science and Technological Development, Republic of Serbia, grant no. 174030.