NON-ASYMPTOTIC RATES FOR COMMUNICATION EFFICIENT DISTRIBUTED ZEROTH ORDER STRONGLY CONVEX OPTIMIZATION

Anit Kumar Sahu* *Dusan Jakovetic*† *Dragana Bajovic*$^\$$ *Soummya Kar**

* Department of Electrical and Computer Engineering, Carnegie Mellon University
{anits,soummyak}@andrew.cmu.edu
$^\$$ Faculty of Technical Sciences, University of Novi Sad
dbajovic@uns.ac.rs
†Faculty of Sciences, University of Novi Sad
djakovet@uns.ac.rs

ABSTRACT

This paper focuses on the problem of communication efficient distributed zeroth order minimization of a sum of strongly convex loss functions. Specifically, we develop distributed stochastic optimization methods for zeroth order strongly convex optimization that are based on an adaptive probabilistic *sparsifying* communications protocol. Under standard assumptions on the cost functions and the noises corrupting the function evaluations, we establish with the proposed method $O(1/(C_{\mathrm{comm}})^{2/3-\zeta})$ mean square error (MSE) convergence rates, for the zeroth order optimization, where C_{comm} is the number of per-node communications and $\zeta > 0$ is arbitrarily small. In the distributed setting considered, the established rate is the best known rate in terms of the MSE-communication cost trade off for zeroth order optimization. Finally, through empirical evaluations we illustrate the proposed algorithm's theoretical guarantees.

Index Terms— Distributed Optimization, Stochastic Optimization, Zeroth Order Optimization, Multi-agent Networks.

1. INTRODUCTION

We study zeroth order distributed strongly convex stochastic optimization over networks. There are N interconnected agents, that aim to collaboratively minimize the sum of their locally known strongly convex costs. Distributed stochastic optimization has had increasing interest of late, e.g., [1–4]. These references consider algorithms which have access to a stochastic first order or a second order oracle. However, in this paper, we focus on zeroth order distributed stochastic optimization methods, where at each time instant (iteration) k, each node queries a stochastic zeroth order oracle ($\mathcal{SZO}$) to get unbiased estimates of function values at a queried point. Such kind of scenarios arise in typical *black box* settings, where only the evaluations of a loss function are known or can be queried for and there is no access to first order gradient or second order Hessian information that can be retrieved. Our focus is on examining the tradeoffs between performance and *communication cost*, measured by the number of per-node transmissions to neighboring nodes in the network; and *computational cost*, measured by the number of per-node queries made to the $\mathcal{SZO}$.

Contributions. Our main contributions are as follows. We develop a novel method for communication efficient zeroth order distributed stochastic optimization. The method is based on a communication protocol which probabilistically sparsifies the message exchanges in the network along iterations. More precisely, each node, at each iteration k, participates in communication (transmits and receives messages in its neighborhood) with probability p_k (independently from the past and from the others), where the parameter p_k decays to zero at a carefully tuned rate. For the proposed method, we establish the $O(1/(C_{\mathrm{comm}})^{2/3-\zeta})$ mean square error (MSE) convergence rate in terms of communication cost[1] C_{comm}, where $\zeta > 0$ is arbitrarily small. At the same time, the method achieves the order-optimal $O(1/(C_{\mathrm{comp}})^{1/2})$ MSE rate in terms of computational cost[2] C_{comp}, that is not improvable even in the centralized setting. The achieved $O(1/(C_{\mathrm{comm}})^{2/3-\zeta})$ MSE-communication rate is significantly faster than existing zeroth order optimization schemes in the distributed setting (see, for example [5–8]), that achieve at best the $O(1/(C_{\mathrm{comm}})^{1/2})$ rate.

Related Work.In the context of distributed stochastic strongly convex optimization, first order schemes with static networks ([2, 9]), deterministic time-varying networks [1, 3, 4] and random time-varying networks albeit with access to exact first order information [10, 11] have been considered. The aforementioned works explicitly characterize the convergence rates in terms of the iteration counter k, that translates

The work of DJ and DB was supported in part by the EU Horizon 2020 project I-BiDaaS, project number 780787. The work of D. Jakovetic was also supported in part by the Serbian Ministry of Education, Science, and Technological Development, grant 174030. The work of AKS and SK was supported in part by National Science Foundation under grant CCF-1513936.

[1] The communication cost is measured in terms of per-node number of transmissions.

[2] The computation cost is measured in terms of per-node number of queries made to the $\mathcal{SZO}$.

into computational cost $C_{\rm comp}$, i.e., number of gradient evaluations under suitable assumptions. More relevant to the current context, references [1, 3, 4] consider deterministically varying networks, assuming that the "union graph" over finite windows of iterations is connected. In contrast, we consider randomly time-varying networks connected only in mean with access to a $\mathcal{SZO}$ for our distributed zeroth order optimization scheme. In the context of distributed zeroth order optimization, [12] considers an algorithm for non-convex minimization over a static graph, where a random directions-random smoothing approach was employed. Reference [8] considers a zeroth order distributed stochastic approximation method and establishes the method's $O(1/k^{1/2})$ convergence rate in terms of the number of iterations, where the number of queries to the $\mathcal{SZO}$ at each iteration scales with the dimension of the optimizer. In contrast, the scheme proposed here utilizes only two calls of the $\mathcal{SZO}$ per node, per iteration, independently from the variable dimension d. However, all the aforementioned work in the distributed setup is aimed at attaining the optimal rate in terms of the iterations or explicitly in terms of the number of queries made to the stochastic oracle in question. In the context of distributed setups with random networks and access to stochastic oracles references [8] and [13] achieve order-optimal rates for zeroth and first order distributed strongly convex optimization respectively.[3] In the context of communication efficient distributed inference and optimization, adaptive communication protocols for first order schemes without explicit characterization of communication cost savings (see, for example [14–16]) and constant proportion of communication savings at the cost of deviating from the order-optimal rate (see, for example [17]) have been considered. In contrast to [14–16], we consider a communication efficient distributed zeroth optimization scheme, where we explicitly characterize the communication savings while ensuring order-optimal convergence rates as compared to [17]. In prior work [18, 19], we developed distributed algorithms with increasingly sparse communications for *statistical estimation problems*. This paper demonstrates that the concept of increasingly sparse communications can be exploited to develop communication-efficient distributed zeroth order stochastic optimization algorithms also. Technically, the setups in [18, 19] and the setup here are very different, requiring new analyses. Communication efficient distributed estimation schemes as proposed in [18, 19] involve *local correctness*, i.e., the optimizers of the sum of loss functions of the individual nodes is a subset of the optimizers of each local function, while in the current work, the setup is rendered *locally incorrect*. We skip the proofs due to space limitations. The proofs can be found in [20].

2. MODEL AND PROPOSED ALGORITHM

Our setup involves a network of N agents which collaborate through an iterative message passing scheme so as to solve the following unconstrained problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d}\sum_{i=1}^{N} f_i(\mathbf{x}), \tag{1}$$

where $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is a convex function available to node i, $i = 1, ..., N$. We make the following assumption on the functions $f_i(\cdot)$:

Assumption A1. For all $i = 1, ..., N$, function $f_i : \mathbb{R}^d \mapsto \mathbb{R}$ is twice continuously differentiable with Lipschitz continuous gradients. In particular, there exist constants $L, \mu > 0$ such that for all $\mathbf{x} \in \mathbb{R}^d$, $\forall i = 1, 2, \cdots, N$

$$\mu\,\mathbf{I} \preceq \nabla^2 f_i(\mathbf{x}) \preceq L\mathbf{I}.$$

From Assumption A1 we have that each f_i, $i = 1, \cdots, N$, is strongly convex with modulus μ. Using standard properties of convex functions, we have for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2,$$
$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\,\|\mathbf{x}-\mathbf{y}\|.$$

The optimization problem in (1) has a unique solution, which we denote by $\mathbf{x}^* \in \mathbb{R}^d$, where the uniqueness is guaranteed by assumption A1. Throughout the paper, we also use $f : \mathbb{R}^d \to \mathbb{R}$, $f(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x})$. We employ a distributed zeroth order optimization scheme to solve (1).

2.1. Zeroth Order Optimization

We employ a distributed random directions stochastic approximation (RDSA) type method to solve (1). Each node i, $i = 1, ..., N$, in our setup maintains a local copy of its local estimate of the optimizer $\mathbf{x}_i(k) \in \mathbb{R}^d$ at all times. In the absence of first order information, each agent i queries the $\mathcal{SZO}$ at time k, to obtain noisy function values of $f_i(\mathbf{x}_i(k))$. An unbiased estimate of $f_i(\cdot)$ is obtained from the $\mathcal{SZO}$ which is then given by,

$$\widehat{f}_i(\mathbf{x}_i(k)) = f_i(\mathbf{x}_i(k)) + v_i(k), \tag{2}$$

where $v_i(k)$ is the measurement noise. In order to approximate the gradient, each agent makes queries to the $\mathcal{SZO}$ twice at each iteration. For instance, agent i queries for $f_i(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k})$ and $f_i(\mathbf{x}_i(k))$ at time k and obtains $\widehat{f}_i(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k})$ and $\widehat{f}_i(\mathbf{x}_i(k))$ respectively, where c_k is a carefully chosen time-decaying factor (to be specified soon) and $\mathbf{z}_{i,k}$ is a random vector such that $\mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\right] = \mathbf{I}_d$. Let $\mathcal{F}_k$ denote the history of the proposed algorithm up to time k which is given by the σ-algebra generated by the collection of random variables $\{\mathbf{L}(s)^4, v_i(s), \mathbf{z}_{i,s}\}$, $i = 1, ..., N$, $s = 0, ..., k-1$. Denote by $\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))$ the approximated gradient. By mean value theorem, we then have:

$$\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) = \frac{\widehat{f}_i\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right) - \widehat{f}_i\left(\mathbf{x}_i(k)\right)}{c_k}\mathbf{z}_{i,k}$$
$$\Rightarrow \mathbb{E}\left[\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))|\mathcal{F}_k\right] = \mathbb{E}\left[\mathbf{z}_{i,k}\mathbf{z}_{i,k}^\top\nabla f_i\left(\mathbf{x}_i(k)\right)|\mathcal{F}_k\right]$$
$$+ \underbrace{c_k\,\mathbb{E}\left[\left(\mathbf{z}_{i,k}^\top\nabla^2 f_i\left(\mathbf{e}_k\right)\mathbf{z}_{i,k}\right)\frac{\mathbf{z}_{i,k}}{2}|\mathcal{F}_k\right]}_{\mathbf{b}_i\left(\mathbf{x}_i(k)\right)}, \tag{3}$$

[3] Reference [8] utilizes a non-diminishing amount of communications across iterations, and hence achieves at best and $O(1/(C_{\rm comm})^{1/2})$ communication rates.

[4] To be specified soon

where $\mathbf{e}_k = \theta \mathbf{x}_i(k) + (1-\theta)\left(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}\right)$ and $\theta \in [0,1]$. Thus, we can write,

$$\begin{aligned}\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)) &= \nabla f_i\left(\mathbf{x}_i(k)\right) + \frac{\widehat{v}_i(k)\mathbf{z}_{i,k}}{c_k} \\ &+ \underbrace{\mathbb{E}\left[\widehat{\mathbf{g}}_i(\mathbf{x}_i(k))|\mathcal{F}_k\right] - \nabla f_i\left(\mathbf{x}_i(k)\right)}_{c_k\mathbf{b}(\mathbf{x}_i(k))},\end{aligned} \tag{4}$$

where $\widehat{v}_i(k) = (\widehat{f}_i(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k}) - f_i(\mathbf{x}_i(k) + c_k\mathbf{z}_{i,k})) - (\widehat{f}_i(\mathbf{x}_i(k)) - f_i(\mathbf{x}_i(k)))$.

Assumption A2. The $z_{i,k}$'s are drawn from a distribution P such that $\mathcal{D}(P) \doteq \sqrt{\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right]}$ is finite.

We provide two examples of two such distributions. If $\mathbf{z}_{i,k}$'s are drawn from $\mathcal{N}(0, \mathbf{I}_d)$, then $\sqrt{\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right]} = \sqrt{d(d+2)(d+4)}$. If $\mathbf{z}_{i,k}$'s are drawn uniformly from the l_2-ball of radius $\sqrt{d}$, then we have, $\|\mathbf{z}_{i,k}\| = \sqrt{d}$ and $\sqrt{\mathbb{E}\left[\|\mathbf{z}_{i,k}\|^6\right]} = d^{3/2}$.

2.1.1. *Communication Scheme*

Let the backbone graph over which we design the increasingly sparsified communication protocol be given by $G = (V, E)$, which is an undirected graph with N vertices, i.e. $|V| = N$ and E represents the edges. For each node i, at every time k, we introduce a binary random variable $\psi_{i,k}$, where

$$\psi_{i,k} = \begin{cases} \rho_k & \textit{with probability } \zeta_k \\ 0 & \textit{else}, \end{cases} \tag{5}$$

where $\psi_{i,k}$'s are independent both across time and the nodes, i.e., across k and i respectively which abstracts out the participation of the node i at time k in the neighborhood information exchange. We specifically take ρ_k and ζ_k of the form

$$\rho_k = \frac{\rho_0}{(k+1)^{\epsilon/2}}, \, \zeta_t = \frac{\zeta_0}{(k+1)^{(\tau/2-\epsilon/2)}}, \tag{6}$$

where $0 < \epsilon < \tau$ and $0 < \tau \leq 1$. Furthermore, define β_k to be

$$\beta_k = (\rho_k\zeta_k)^2 = \frac{\beta_0}{(k+1)^{\tau}}. \tag{7}$$

The random time-varying Laplacian $\mathbf{L}(k) \in \mathbb{R}^{N\times N}$ which abstracts the inter-node information exchange can be represented as follows:

$$\mathbf{L}_{i,j}(k) = \begin{cases} -\psi_{i,k}\psi_{j,k} & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ \sum_{l\neq i}\psi_{i,k}\psi_{l,k} & i = j. \end{cases} \tag{8}$$

The above protocol avoids directed graphs by enforcing the requirement of both nodes being active to be able to communicate with each other. We have, for $\{i,j\} \in E$:

$$\mathbb{E}\left[\mathbf{L}_{i,j}(k)\right] = -\beta_k, \, \mathbb{E}\left[\mathbf{L}_{i,j}^2(k)\right] = \frac{\rho_0\beta_0}{(k+1)^{\tau+\epsilon}}.$$

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

$$\overline{\mathbf{L}}_{i,j} = \begin{cases} -1 & \{i,j\} \in E, i \neq j \\ 0 & i \neq j, \{i,j\} \notin E \\ -\sum_{l\neq i} L_{i,l} & i = j. \end{cases} \tag{9}$$

We make the following assumption on $\overline{\mathbf{L}}$.

Assumption A3. The inter-agent communication graph is connected on average, i.e., $\overline{\mathbf{L}}$ is connected. In other words, $\lambda_2(\overline{\mathbf{L}}) > 0$, where $\lambda_2(\overline{\mathbf{L}})$ is the second largest eigenvalue of $\overline{\mathbf{L}}$.

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 agent in the graph reaches other agents in a *balanced* fashion, thus ensuring information flow. With the communication protocol in place, we now state the optimizer update rule. For arbitrary deterministic initializations $\mathbf{x}_i(0) \in \mathbb{R}^d$, $i = 1, ..., N$, the optimizer update rule at node i and $k = 0, 1, ...,$ of the consensus+innovations form [21] and is given as follows:

$$\begin{aligned}\mathbf{x}_i(k+1) = \mathbf{x}_i(k) &- \sum_{j\in\Omega_i}\psi_{i,k}\psi_{j,k}\left(\mathbf{x}_i(k) - \mathbf{x}_j(k)\right) \\ &- \alpha_k\widehat{\mathbf{g}}_i(\mathbf{x}_i(k)),\end{aligned} \tag{10}$$

where $\widehat{\mathbf{g}}_i(\cdot)$ is as defined in (4) and Ω_i represents the neighborhood of agent i at time k. The weight sequences $\{\alpha_k\}$, $\{c_k\}$ and $\{\beta_k\}$ are given by $\alpha_k = \alpha_0/(k+1)$, $c_k = c_0/(k+1)^{\delta}$ and $\beta_k = \beta_0/(k+1)^{\tau}$ respectively, where $\alpha_0, c_0, \beta_0 > 0$. We state an assumption on the weight sequences and measurement noises before proceeding further.

Assumption A4. The constants α_0, $\delta > 0$ and $\tau \in (0,1)$ are chosen such that, $\sum_{k=1}^{\infty}\frac{\alpha_k^2}{c_k^2} < \infty$.

Assumption A5. For each $i = 1, ..., N$, the sequence of measurement noises $\{v_i(k)\}$ satisfies for all $k = 0, 1, ...$:

$$\mathbb{E}[\, v_i(k)\,|\,\mathcal{F}_k] = 0, \; \mathbb{E}[\, v_i(k)^2\,|\,\mathcal{F}_k\,] \leq c_f\|\mathbf{x}_i(k)\|^2 + \sigma^2, \text{ a.s.},$$

where c_f and σ^2 are nonnegative constants.

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

$$\mathcal{C}_t = \mathbb{E}\left[\sum_{s=0}^{t-1}\mathbb{I}_{\{\textit{node C transmits at s}\}}\right], \tag{11}$$

where $\mathbb{I}_A$ represents the indicator of event A. Note that the per-node communication cost in (11) is the same as the network average of communication costs across all nodes, as the activation probabilities are homogeneous across nodes.

3. CONVERGENCE RATES

In this section, we state the results concerning the convergence rate of the proposed zeroth order optimization algorithm.

Theorem 3.1. *1) Consider the optimizer estimate sequence* $\{\mathbf{x}(k)\}$ *generated by the algorithm* (10). *Let assumptions A1-A5 hold. Then, for each node* i*'s optimizer estimate* $\mathbf{x}_i(k)$ *and the solution* $\mathbf{x}^\star$ *of problem* (1), $\forall k \geq k_3$ *there holds:*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] \leq 2M_k + \frac{32NL^2\Delta_{1,\infty}\alpha_0^2}{\mu^2\lambda_2^2\left(\overline{\mathbf{L}}\right)c_0^2\beta_0^2(k+1)^{2-2\tau-2\delta}}$$
$$\frac{8L^2\mathcal{D}^2(P)c_0^2}{\mu^2(k+1)^{2\delta}} + \frac{4\Delta_{1,\infty}\alpha_0^2}{\lambda_2^2\left(\overline{\mathbf{L}}\right)\beta_0^2c_0^2(k+1)^{2-2\tau-2\delta}}$$
$$+ \frac{4N\alpha_0\left(dc_f q_\infty(N,d,\alpha_0,c_0) + dN\sigma_1^2\right)}{\mu c_0^2(k+1)^{1-2\delta}}, \tag{12}$$

where, $k_3 = \max\{k_0, k_1, k_2\}$,
$k_0 = \inf\left\{k | \frac{\mu}{2} > (L-\mu)\sqrt{N}dc_k + \frac{2c_f\alpha_k}{c_k^2}\right\}$,
$k_1 = \inf\left\{k | \frac{\mu}{2} > \frac{\sqrt{N}}{2}d(P)Lc_k + \frac{2dc_f\alpha_k}{c_k^2}\right\}$,
$k_2 = \inf\{k | \frac{\beta_k}{2}\lambda_2\left(\overline{\mathbf{L}}\right) > 4|E|\beta_k\rho_k\}$,
$\Delta_{1,\infty} = 6dc_f q_\infty(N,d,\alpha_0,c_0) + 6dN\sigma_1^2$ *and* $q_\infty(N,d,\alpha_0,c_0) = \mathbb{E}\left[\|\mathbf{x}(k_0) - \mathbf{x}^o\|^2\right] + \frac{\sqrt{N}d(P)L\alpha_0c_0}{2\delta} + \frac{Nd^2(P)L^2\alpha_0^2c_0^2}{4(1+2\delta)} + 4\frac{\|\nabla F(\mathbf{x}^o)\|^2}{\mu^2} + \frac{d\alpha_0^2\left(2c_fN\|\mathbf{x}^o\|^2 + N\sigma^2\right)}{c_0^2(1-2\delta)}$. M_k *is a term which decays faster than the rest of the terms.*
2) In particular, the RHS of (12) *decays as* $(k+1)^{-\delta_1}$*, where* $\delta_1 = \min\{1-2\delta, 2-2\tau-2\delta, 2\delta\}$*. By, optimizing over* τ *and* δ*, we obtain that for* $\tau = 1/2$ *and* $\delta = 1/4$*,*

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^*\|^2\right] = O\left(\frac{1}{k^{\frac{1}{2}}}\right), \ \forall i.$$

3) The MSE-communication rate is given by,

$$\mathbb{E}\left[\|\mathbf{x}_i(k) - \mathbf{x}^\star\|^2\right] = \Theta\left(\frac{1}{\mathcal{C}_k^{2/3-\zeta}}\right).$$

Theorem 3.1 asserts that the MSE-communication rate can be improved to $\Theta\left(\mathcal{C}_k^{-2/3+\zeta}\right)$ while keeping the MSE decay rate at $O\left(k^{-\frac{1}{2}}\right)$ by the proposed zeroth order distributed algorithm. The performance of the zeroth order optimization scheme depends explicitly on the connectivity of the expected Laplacian through a $\frac{1}{\lambda_2^2(\overline{\mathbf{L}})}$ scaling. In particular, communication graphs which are well connected, i.e., have higher values of $\lambda_2\left(\overline{\mathbf{L}}\right)$ will have lower MSE as compared to a counterpart with lower values of $\lambda_2\left(\overline{\mathbf{L}}\right)$. However, the network connectivity quality, i.e., $\lambda_2\left(\overline{\mathbf{L}}\right)$, does not affect the convergence rate in k.

4. SIMULATIONS

In this section, we provide evaluations of the proposed algorithm on the Abalone dataset [22]. To be specific, we consider ℓ_2-regularized empirical risk minimization for the Abalone dataset, where the regularization function is given by $\Psi_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ and the loss function is the squared loss. We consider a 10 node network. The Abalone dataset has 4177 data points out of which 577 data points are kept aside as the test set and the other 3600 is divided equally among the 10 nodes resulting in each node having 360 data points. The vectors $\mathbf{z}_{i,k}$'s are sampled from a normal distribution with unit covariance. The measurement noises $v_{i,k}$ are sampled from a standard normal distribution. For the proposed algorithm, we compare it with a zeroth order scheme employing the static Laplacian (Benchmark). The data points at each node are sampled without replacement in a contiguous manner. Figure 1 compares the test error for the schemes, where it can be clearly observed that the test error is indistinguishable in terms of the number of iterations or equivalently in terms of the number of queries to the stochastic zeroth oracle. Figure 2 demonstrates the superiority of the proposed algorithm in terms of the test error versus communication cost as compared to the benchmark, as predicted by Theorem 3.1. For example, at the same relative test error level of 0.3, the proposed algorithm uses up to 3x less number of transmissions as compared to the benchmark scheme.

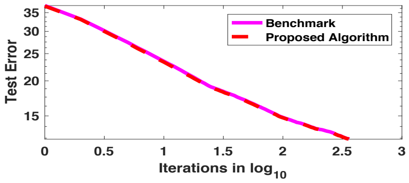

Fig. 1: Test Error vs Iterations

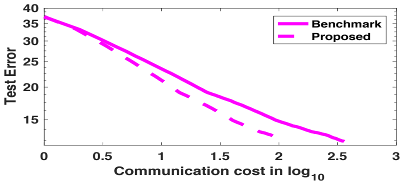

Fig. 2: Test Error vs Communication Cost

5. CONCLUSION

We have developed a communication efficient distributed stochastic zeroth order optimization method for smooth strongly convex optimization, where by employing a random directions stochastic approximation type consensus+innovations algorithm. Through the analysis of the considered method, we have established the order optimal $O(k^{-1/2})$ MSE convergence rate while improving the MSE-communication rate to $\Theta\left(\mathcal{C}_k^{-2/3+\zeta}\right)$. In particular, we have also quantified the mean square error of the generated optimizer estimate sequence in terms of the algorithm parameters. Future work includes extending the current approach to a broader class of convex and non-convex functions.

6. REFERENCES

[1] D. Yuan, Y. Hong, D. W. C. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, April 2018.

[2] K. Tsianos and M. Rabbat, "Distributed strongly convex optimization," *50th Annual Allerton Conference on-Communication, Control, and Computing*, Oct. 2012.

[3] N. D. Vanli, M. O. Sayin, and S. S. Kozat, "Stochastic subgradient algorithms for strongly convex optimization over distributed networks," *IEEE Transactions on network science and engineering*, vol. 4, no. 4, pp. 248–260, Oct.-Dec. 2017.

[4] A. Nedic and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.

[5] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE transactions on automatic control*, vol. 37, no. 3, pp. 332–341, 1992.

[6] J. Kiefer and J. Wolfowitz, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

[7] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, "Optimal rates for zero-order convex optimization: The power of two function evaluations," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806, 2015.

[8] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approach," 2018, submitted to CDC 2018, Available at https://www.dropbox.com/s/kfc2hgbfcx5yhr8/MainCDC2018KWSA.pdf.

[9] Z. J. Towfic, J. Chen, and A. H. Sayed, "Excess-risk of distributed stochastic learners," *IEEE Transactions on Information Theory*, vol. 62, no. 10, Oct. 2016.

[10] I. Lobel and A. E. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Trans. Automat. Contr.*, vol. 56, no. 6, pp. 1291–1306, Jan. 2011.

[11] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

[12] D. Hajinezhad, M. Hong, and A. Garcia, "Zeroth order nonconvex multi-agent optimization over networks," *arXiv preprint arXiv:1710.09997*, 2017.

[13] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," 2018, submitted to CDC 2018, Available at https://www.dropbox.com/s/zylonzrhypy29zj/MainCDC2018.pdf.

[14] 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.

[15] 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.

[16] 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.

[17] Z. Wang, Z. Yu, Q. Ling, D. Berberidis, and G. B. Giannakis, "Decentralized rls with data-adaptive censoring for regressions over large-scale networks," *arXiv preprint arXiv:1612.08263*, 2016.

[18] A. K. Sahu, D. Jakovetic, and S. Kar, "Communication optimality trade-offs for distributed estimation," *arXiv preprint arXiv:1801.04050*, 2018.

[19] A. K. Sahu, D. Jakovetic, D. Bajovic, and S. Kar, "Communication efficient distributed weighted non-linear least squares estimation," 2018, submitted to EURASIP Journal on Advances in Signal Processing, Available at https://users.ece.cmu.edu/~anits/comm_ci_eurasipSUBMIT.pdf.

[20] ——, "Non-asymptotic rates for communication efficient distributed zeroth order strongly convex optimization," 2018, available at https://www.dropbox.com/s/53rfp208rmysym3/globalsip2018.pdf.

[21] S. Kar and J. M. Moura, "Consensus+ innovations distributed inference over networks: cooperation and sensing in networked systems," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 99–109, 2013.

[22] "Libsvm regression datasets," https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html.