Stochastic Subgradient Algorithms for Strongly Convex Optimization over Distributed Networks

N. Denizcan Vanli, Muhammed O. Sayin, and Suleyman S. Kozat, Senior Member, IEEE

Abstract—We study diffusion and consensus based optimization of a sum of unknown convex objective functions over distributed networks. The only access to these functions is via stochastic gradient oracles, each of which is only available at a different node, and a limited number of gradient oracle calls is allowed at each node. In this framework, we introduce a strongly convex optimization algorithm based on the stochastic gradient descent (SGD) updates. Particularly, we use a certain time-dependant weighted averaging of the SGD iterates, which yields an optimal convergence rate of $O\left(\frac{\sqrt{T}}{N}\right)$ after $T$ gradient updates for each node on a network of $N$ nodes. We then show that after $T$ gradient oracle calls, the SGD iterates at each node achieve a mean square deviation (MSD) of $O\left(\frac{\sqrt{T}}{N}\right)$. We provide the explicit description of the algorithm and illustrate the merits of the proposed algorithm with respect to the state-of-the-art methods in the literature.

Index Terms—Distributed processing, convex optimization, online learning, diffusion strategies, consensus strategies.

1 INTRODUCTION

T
he demand for large-scale networks consisting of multiple agents (i.e., nodes) with different objectives is steadily growing due to their increased efficiency and scalability compared to centralized distributed structures [1]–[11]. A wide range of problems in the context of distributed and parallel processing can be considered as a minimization of a sum of objective functions, where each function (or information on each function) is available only to a single agent or node [10], [11]. In such practical applications, it is essential to process the information in a decentralized manner since transferring the objective functions as well as the entire resources (e.g., data) may not be feasible or possible [1], [2]. For example, in a distributed data mining scenario, privacy considerations may prohibit sharing of the objective functions [10], [11]. Similarly, in a distributed wireless network, energy considerations may limit the communication rate between agents [1], [2]. In such settings, parallel or distributed processing algorithms, where each node performs its own processing and share information subsequently, are preferable over the centralized methods [1], [2], [10], [11].

Here, we consider minimization of a sum of unknown convex objective functions, where each agent (or node) observes only its particular objective function via the stochastic gradient oracles. Particularly, we seek to minimize this sum of functions with a limited number of gradient oracle calls at each agent. In this framework, we introduce a distributed online convex optimization algorithm based on the SGD iterates that efficiently minimizes this cost function. Specifically, each agent uses a time-dependant weighted combination of the SGD iterates and achieves the presented performance guarantees, which matches with the lower bounds presented in [14], only with a relatively small excess term caused by the unknown network model. The proposed method is comprehensive, in that any communication strategy, such as adapt-then-combine (ATC) [1], combine-then-adapt (CTA) [1], and consensus [4], are incorporated into our algorithm in a straightforward manner as pointed out in the paper. We compare the performance of our algorithm respect to the state-of-the-art methods [4], [9], [15] in the literature and present the outstanding performance improvements for various well-known network topologies.

Distributed networks are successfully used in parameter estimation problems [1], [2], [9], and recently used for convex optimization via projected subgradient techniques [4], [8], [10], [11]. In [9], the authors illustrate the performance of the least mean squares (LMS) algorithm over distributed networks using different diffusion strategies. We emphasize that this problem can also be casted as a distributed convex optimization problem, hence our results can be applied to these problems in a straightforward manner. In [8], the authors consider the cooperative optimization of the cost function under convex inequality constraints. However, the problem formulation as well as the convergence results in this paper are significantly different than the ones in [8]. In [4], the authors present a deterministic analysis of the SGD iterates and our results builds on them by illustrating a stronger convergence bound in expectation while also providing MSD analyses of the SGD iterates.

In [10], [11], the authors consider the distributed convex optimization problem and present the probability-1 and mean square convergence results of the SGD iterates. In this paper, on the other hand, we provide the expected convergence rate of our algorithm and the MSD of the SGD iterates at any time instant.

Similar convergence analyses are recently illustrated in the computational learning theory [12]–[16]. In [12], the authors provide deterministic bounds on the learning performance (i.e., regret) of the SGD algorithm. In
these analyses are extended and a regret-optimal learning algorithm is proposed. In the same lines, in [13], the authors describe a method to make the SGD algorithm optimal for the strongly convex optimization. However, these approaches rely on the smoothness of the optimization problem. In [15], a different method to achieve the optimal convergence rate is proposed and its performance is analyzed. On the other hand, in this paper, the convex optimization is performed over a network of localized learners, unlike [12, 13, 15, 16]. Our results illustrate the convergence rates over any unknown communication graph, and in this sense build upon the analyses of the centralized learners. Furthermore, unlike [13, 15], our algorithm does not require the optimization problem to be sufficiently smooth.

Our main contributions are as follows. i) We introduce a distributed online convex optimization algorithm based on the SGD iterates that achieves an optimal convergence rate of $O\left(\frac{N\sqrt{T}}{T}\right)$ after $T$ gradient updates, for each and every node on the network. We emphasize that this convergence rate is optimal since it achieves the lower bounds presented in [14] up to constant terms. ii) We show that the SGD iterates at each node achieve a MSD of $O\left(\frac{N}{T}\right)$ after $T$ gradient updates. iii) Our analyses cover the well-known diffusion strategies such as the ATC and CTA algorithms, and the consensus strategy in a straightforward manner as illustrated in our paper. iv) We illustrate the highly significant performance gains of the introduced algorithm with respect to the state-of-the-art methods in the literature, and illustrate the highly significant performance gains under various network topologies.

The organization of the paper is as follows. In Section 2 we introduce the distributed convex optimization framework and provide the notations. We then introduce the main result of the paper, i.e., a SGD based convex optimization algorithm, in Section 3 and analyze the convergence rate of the introduced algorithm. In Section 4 we demonstrate the performance of our algorithm with respect to the state-of-the-art methods through simulations and then conclude the paper with several remarks in Section 5.

## 2 Problem Description

We consider the problem of distributed strongly convex optimization over a network of $N$ nodes. At each time $t$, each node $i$ observes a pair of regressor vectors and data, i.e., $u_{t,i}$ and $d_{t,i}$, where the pairs $(u_{t,i}, d_{t,i})$ are independent and identically distributed for all $t \geq 1$. The aim of each node is to minimize a strongly convex cost function $f_i(\cdot)$ over a convex set $\mathcal{W}$. However, we do not know $f_i(\cdot)$ and each node accesses to its $f_i(\cdot)$ only via a stochastic gradient oracle, which given some $w \in \mathcal{W}$, produces a random vector $g_i$, whose expectation $\mathbb{E}g_i = g_i$ is a subgradient of $f_i$ at $w$. Using these stochastic gradient oracles, each node estimates a parameter of interest $w_{t,i}$ on a common convex set $\mathcal{W}$ and calculates an estimate of the output as $d_{t,i} = w_{t,i}^T u_{t,i}$, i.e., by a first order linear method. After observing the true data, node $i$ suffers an expected loss of $\mathbb{E}f_i(w_{t,i})$, where $f_i(\cdot)$ is a strongly convex loss function, e.g., $\mathbb{E}f_i(w_{t,i}) = \mathbb{E}\left[l(w_{t,i}; u_i, d_i) + \frac{\lambda}{2} ||w_{t,i}||^2\right]$, where $l(w_{t,i}; u_i, d_i)$ is a Lipschitz-continuous convex loss function (with respect to the first variable) such as the square loss, i.e., $l(w_{t,i}; u_i, d_i) = (d_i - w_{t,i}^T u_i)^2$.

In this framework, at each time $t$, after observing the true data $d_{t,i}$, each node exchanges (i.e., diffuses) information with its neighbors in order to produce the next estimate of the parameter vector $w_{t+1,i}$. The network is assumed to form an irreducible graph but each node is not necessarily directly connected to one another. In this setup, our aim is to minimize the following cost function

$$f(w_{t,i}) \triangleq \sum_{j=1}^{N} f_j(w_{t,i}), \quad (1)$$

for each node $i$, over a convex set $\mathcal{W}$, with at most $t$ stochastic gradient oracle calls. Particularly, using at most $t$ calls to this oracle, we seek to find a parameter vector $w_{t,i}$ such that $f(w_{t,i})$ is as small as possible in expectation, i.e.,

$$\min_{w_{t,i} \in \mathcal{W}} \mathbb{E}f(w_{t,i}). \quad (2)$$

Here, the aim is to minimize a strongly convex cost function over a distributed network via $N$ different localized learners, where each learner is allowed to use at most $t$ calls to the gradient oracle until time $t$.

Throughout the paper, we make the following standard assumptions:

1) Each $f_i$ is $\lambda$-strongly convex over $\mathcal{W}$, where $\mathcal{W} \subseteq \mathbb{R}^p$ is a closed convex set.

2) The norm of each subgradient $g_i(w) \triangleq \nabla_w f_i(w)$ has a finite variance, i.e., $\mathbb{E}||g_i(w)||^2 \leq G^2$, for any $w \in \mathcal{W}$.

3) The communication graph $H$ forms a doubly stochastic matrix such that $H$ is irreducible and aperiodic, hence for some $\theta \in \mathbb{R}$ and $0 \leq \gamma < 1$, we have

$$\sum_{i=1}^{N} \left|H_{ij} - \frac{1}{N}\right| \leq \theta \gamma^j, \forall j = 1, \ldots, N. \quad (3)$$

4) Initial weights at each node is identically initialized to avoid any bias, i.e., $w_{1,i} = w_{1,j}, \forall i, j \in \{1, \ldots, N\}$.

We emphasize that these assumptions are widely used to analyze the convergence of online algorithms such as in [4, 10–13, 15, 16]. Particularly, the first assumption may not hold for some loss functions such as the square loss, i.e., $l(w_{t,i}; u_i, d_i) = (d_i - w_{t,i}^T u_i)^2$. 

1 Throughout the paper, all vectors are column vectors and represented by boldface lowercase letters. Matrices are represented by boldface uppercase letters. For a vector $x$, $||x|| = \sqrt{x^T x}$ is the $l^2$-norm. Here, 0 (and 1) denotes a vector with all zero (and one) elements and the dimensions can be understood from the context. For a matrix $H$, $H_{ij}$ represents its entry at the $i$th row and $j$th column. $\Pi_{\mathcal{W}}$ denotes the Euclidean projection onto $\mathcal{W}$, i.e., $\Pi_{\mathcal{W}}(w') = \arg\min_{w \in \mathcal{W}} \|w - w'\|$. 

$$\Pi_{\mathcal{W}}(w') = \arg\min_{w \in \mathcal{W}} \|w - w'\|.$$
However, one can incorporate a small regularization term, i.e., \( I(w_{t,i}; u_i, d_i) + \frac{\lambda}{2} \|w_{t,i}\|^2 \) in order to make it strongly convex. The second assumption is only needed for the technicalities of our proof and our algorithm does not need \( G \). Nevertheless, this assumption is a widely used one to analyze the performance of SGD based algorithms (4, 4, 12, 13, 13) that results from the Lipschitz continuity of the cost function. The third assumption indicates that the convergence of the communication graph is geometric. This assumption holds when the communication graph is strongly connected, doubly stochastic, and each node gives a nonzero weight to the iterates of its neighbors. Hence, this is a practical assumption and its proof can be obtained using the tools of the finite-state Markov chain theory (for the complete proof, see [10]). The last assumption is basically an unbiasedness condition, which is reasonable since the objective weight \( w^* \) is completely unknown to us. Even though the initial weights are not identical, our analyses still hold, however with small excess terms.

### 3 Main Result

In this section, we present the main result of the paper, i.e., introduce an algorithm based on the SGD updates that achieves an expected regret upper bound of \( O\left(\frac{N^2 N}{\lambda (T+1)}\right) \) after \( T \) iterations. The proposed method uses time dependent weighted averages of the SGD updates at each node together with the ATC strategy to efficiently diffuse the iterates to achieve this performance. However, as discussed in Remark 1 of this paper, our algorithm can be extended to consensus and different diffusion strategies such as the CTA method. The complete description of the algorithm can be found in Algorithm 1.

#### Theorem 1. Under Assumptions 1-4, Algorithm 1 with learning rate \( \mu_t = \frac{2}{(t+1)} \) and weighted parameters \( \hat{w}_{t,i} \), when applied to any independent and identically distributed regressor vectors and data, i.e., \((u_{t,i}, d_{t,i})\) for all \( t \geq 1 \) and \( i = 1, \ldots, N \), achieves the following convergence guarantee

\[
\mathbb{E}[f(\hat{w}_{T,i}) - f(w^*)] \leq \frac{8NG^2}{\lambda(T+1)} \left( \frac{3}{2} + \frac{4\gamma \sqrt{N}}{1 - \gamma} \right),
\]

(4)

for all \( T \geq 1 \), and the SGD iterates achieve the following MSD guarantee

\[
\mathbb{E}\|w_{T+1} - w^*\|^2 \leq \frac{16CG^2}{\lambda^2(T+1)} \left( \frac{3}{2} + \frac{4\gamma \sqrt{N}}{1 - \gamma} \right),
\]

(5)

where \( w_{t,i} = \frac{1}{N} \sum_{i=1}^{N} w_{t,i} \).

This theorem illustrates the convergence rate of our algorithm over distributed networks. The upper bound on the regret \( O\left(\frac{N^2 N}{\lambda (T+1)}\right) \) results due to [3] since the algorithm suffers from a “diffusion regret” to sufficiently exchange (i.e., diffuse) the information among the nodes. This convergence rate matches with the lower bounds presented in [14] up to constant terms, hence is optimal in a minimax sense. The computational complexity of the introduced algorithm is on the order of the computational complexity of the SGD iterates up to constant terms. Furthermore, the communication load of the proposed method is the same as the communication load of the SGD algorithm. However, by using a time dependant averaging of the SGD iterates, our algorithm achieves a significantly improved performance as shown in Theorem 1 and illustrated through our simulations in Section 4.

#### Proof. In order to efficiently manage the recursions, we first consider the projection operation and let

\[
x_{t,i} \triangleq \Pi_{\mathcal{W}}(\psi_{t+1,i}) - \psi_{t+1,i},
\]

(6)

where \( \mathbb{E}[|x_{t,i}|^2] \leq \mathbb{E}[|\mu_t g_{t,i}|^2] \leq G^2 \mu_t^2 \) with \( g_{t,i} \triangleq \nabla w_{t,i} f_i(w_{t,i}) \). Then, we can compactly represent the averaged estimation parameter \( \bar{w}_t = \frac{1}{N} \sum_{i=1}^{N} w_{t,i} \) in a recursive manner as follows

\[
\bar{w}_{t+1} = \frac{1}{N} \sum_{j=1}^{N} \left[ \sum_{i=1}^{N} H_{ij} (w_{t,i} - \mu_t g_{t,i} + x_{t,i}) \right]
\]

\[
= \frac{1}{N} \sum_{i=1}^{N} (x_{t,i} - \mu_t g_{t,i}),
\]

(7)

where the last line follows since \( H \) is right stochastic, i.e., \( H1 = 1 \).

Hence, the MSD of these average iterates with respect to \( w^* \) can be obtained as follows

\[
\mathbb{E}[||w_{t+1} - w^*||^2] = \mathbb{E} \left[ \frac{1}{N^2} \sum_{i=1}^{N} (x_{t,i} - \mu_t g_{t,i})^2 \right]
\]

\[
+ ||w_t - w^*||^2 + \frac{2}{N} \sum_{i=1}^{N} (x_{t,i} - \mu_t g_{t,i})^T (w_t - w^*) \cdot
\]

(8)

We then upper bound the right hand side (RHS) of (8) term by term as follows.

We first upper bound the first term of the RHS of (8)
as follows
\[
E \frac{1}{N^2} \left( \sum_{i=1}^{N} (x_{t,i} - \mu_i g_{t,i}) \right)^2
\leq E \frac{1}{N^2} \left( \sum_{i=1}^{N} ||x_{t,i}|| + \mu_i ||g_{t,i}|| \right)^2
\leq 4G^2 \mu_i^2. \tag{9}
\]

We next turn our attention to \(E [-g_{T,i}^T (w_t - w^*)] \) term in (8) and upper bound this term as follows
\[
E [-g_{T,i}^T (w_t - w^*)] = E [-g_{T,i}^T (w_t - w_{t,i} + w_{t,i} - w^*)]
\leq E [-g_{T,i}^T (w_t - w_{t,i}) + f_i(w^*) - f_i(w_{t,i})
- \frac{\lambda}{2} ||w^* - w_{t,i}||^2]
\leq E [-g_{T,i}^T (w_t - w_{t,i}) + g_{T,i}(w_t - w_{t,i}) + f_i(w^*)
- f_i(w_{t,i}) - \frac{\lambda}{2} ||w^* - w_{t,i}||^2 - \frac{\lambda}{2} ||w_t - w_{t,i}||^2]
\leq E \left[ f_i(w^*) - f_i(w_{t,i}) + \left( \frac{||g_{T,i}||}{2} + ||g_{T,i}|| \right) \right]
\times ||w_{t,i} - w_{t,i}|| - \frac{\lambda}{2} ||w^* - w_{t,i}||^2 - \frac{\lambda}{2} ||w_t - w_{t,i}||^2, \tag{10}
\]

where (10) follows from the \(\lambda\)-strong convexity, i.e.,
\[
E f_i(w^*) \geq E \left[ f_i(w_{t,i}) + g_{T,i}^T (w^* - w_{t,i}) + \frac{\lambda}{2} ||w^* - w_{t,i}||^2 \right], \tag{11}
\]
(11) also follows from the \(\lambda\)-strong convexity, i.e.,
\[
E f_i(w_{t,i}) \geq E \left[ f_i(w_t) + g_{T,i}^T (w_t - w_{t,i}) + \frac{\lambda}{2} ||w_t - w_{t,i}||^2 \right], \tag{12}
\]
(12) follows from the Cauchy-Schwarz inequality, and \(g_{T,i} \triangleq \nabla w_i f_i(w_t)\). Summing (12) from \(i = 1\) to \(N\), we obtain
\[
- \sum_{i=1}^{N} g_{T,i}^T (w_t - w^*) \leq \left[ 2G \sum_{i=1}^{N} ||w_{t,i}|| + f(w^*)
- f(w_t) - \frac{\lambda N}{2} \sum_{i=1}^{N} \left( ||w^* - w_{t,i}||^2 + ||w_{t,i} - w_{t,i}||^2 \right) \right]
\leq E \left[ f(w^*) - f(w_t) + 2G \sum_{i=1}^{N} ||w_{t,i} - w_{t,i}||
- \frac{\lambda N}{2} ||w_t - w^*||^2 \right], \tag{13}
\]
where the last line follows from the Jensen’s inequality due to the convexity of the norm operator.

We finally turn our attention to \(E x_{T,i}^T (w_t - w^*) \) term in (8) and upper bound this term as follows
\[
E x_{T,i}^T (w_t - w^*) = E \left[ x_{T,i}^T (w_t - \psi_{t+1,i})
+ x_{T,i}^T (\psi_{t+1,i} - w^*) \right]
\leq E \mu_i \left[ ||w_{t,i}|| + \psi_{t+1,i} \right], \tag{14}
\]
where the last line follows from the definition of the Euclidean projection.

Putting (9), (13), and (14) back in (8), we obtain
\[
E ||w_{t+1} - w^*||^2 \leq E \left[ (1 - \lambda \mu_i) ||w_t - w^*||^2 + 4G^2 \mu_i^2
+ \frac{2\mu_i}{N} \left\{ f(w^*) - f(w_t) + G \sum_{i=1}^{N} (2 ||w_{t,i} - w^*||
+ ||w_t - \psi_{t+1,i}||) \right\} \right]. \tag{15}
\]

We next upper bound the terms \(||w_t - w_{t,i}||\) and \(||w_t - \psi_{t+1,i}||\) in (15) as follows. In order to obtain a compact representation, we first let \(W_t \triangleq [w_{t,1}, \ldots, w_{t,N}], G_t \triangleq [g_{t,1}, \ldots, g_{t,N}], \) and \(X_t \triangleq [x_{t,1}, \ldots, x_{t,N}].\) Then, we obtain the recursion on \(W_t\) as follows
\[
W_t = W_{t-1} H^{-1} + \sum_{z=1}^{t-1} (X_{t-z} - \mu_{t-z} G_{t-z}) H^2. \tag{16}
\]

Letting \(1_i\) denote the basis function for the \(i\)th dimension, i.e., only the \(i\)th entry of \(1_i\) is 1 whereas the rest is 0, we have
\[
E ||w_t - w_{t,i}|| = E \left[ ||W_t \left( \frac{1}{N} 1 - 1_i \right) || \right]
\leq E \left[ ||w_{t-1,i}|| + 2 \sum_{z=1}^{t-1} \mu_{t-z} \left[ G_{t-z} \left( \frac{1}{N} 1 - H^2 1_i \right) \right] \right]
\leq 2G\theta \sqrt{N} \sum_{z=1}^{t-1} \mu_{t-z} \gamma^z, \tag{17}
\]
where the last line follows from (3). Following similar lines to (17), an upper bound for the term \(||w_t - \psi_{t+1,i}||\) can be obtained as
\[
E ||w_t - \psi_{t+1,i}|| \leq \mu_t + 2G\theta \sqrt{N} \sum_{z=1}^{t-1} \mu_{t-z} \gamma^z. \tag{18}
\]

Putting (17) and (18) back in (15), we obtain
\[
E \left[ f(w_t) - f(w^*) \right] \leq 3NG^2 \mu_t + 6N \sqrt{NG^2} \theta \sum_{z=1}^{t-1} \mu_{t-z} \gamma^z
+ \frac{N}{2\mu_t} E \left( (1 - \lambda \mu_t) ||w_t - w^*||^2 - ||w_{t+1} - w^*||^2 \right). \tag{19}
\]

Since we also have \(E [f(w_{t,i}) - f(w_t)] \geq E [NG ||w_t - w_{t,i}||]\), we obtain
\[
E [f(w_{t,i}) - f(w^*)] \leq 3NG^2 \mu_t + 8N \sqrt{NG^2} \theta \sum_{z=1}^{t-1} \mu_{t-z} \gamma^z
+ \frac{N}{2\mu_t} E \left( (1 - \lambda \mu_t) ||w_t - w^*||^2 - ||w_{t+1} - w^*||^2 \right). \tag{20}
\]

Multiplying both sides of (20) by \(t\) and summing from
Dividing both sides by \( t \) yields [16]
\[
\mathbb{E} \left[ \sum_{t=1}^{T} t |f(w_{t,j}) - f(w^*)| \right] \leq 8N\sqrt{NG^2\theta} \sum_{t=1}^{T} \sum_{z=1}^{t} \mu_{t-z} \gamma^z
\]

\[
+ \mathbb{E} \left[ \frac{N(1-\lambda\mu_t)}{2\mu_t} \|w_1 - w^*\|^2 - \frac{TN}{2\mu_t} \|w_{T+1} - w^*\|^2 \right]
\]

\[
+ \sum_{t=2}^{T} \frac{N}{2} \left( \frac{(1-\lambda\mu_t)}{\mu_t} - \frac{t-1}{\mu_{t-1}} \right) \|w_t - w^*\|^2
\]

\[
+ 3NG^2 \sum_{t=1}^{T} \mu_t.
\]

Observing that
\[
\sum_{t=1}^{T} \sum_{z=1}^{t} \mu_{t-z} \gamma^z \leq \sum_{z=1}^{T} \sum_{t=1}^{T} \mu_{t-z} \gamma^z \leq \frac{\gamma}{1-\gamma} \sum_{t=1}^{T} \mu_t,
\]

and inserting \( \mu_t = \frac{2}{\lambda(t+1)} \), we obtain
\[
\mathbb{E} \left[ \sum_{t=1}^{T} t |f(w_{t,j}) - f(w^*)| \right] \leq \frac{8N\sqrt{NG^2\theta}}{1-\gamma} \sum_{t=1}^{T} \frac{2t}{\lambda(t+1)}
\]

\[
- \frac{\lambda N T(T+1)}{4} \mathbb{E} \|w_{T+1} - w^*\|^2 + \frac{6NG^2 T}{\lambda}
\]

\[
\leq \frac{4NG^2 T}{\lambda} \left( \frac{3}{2} + \frac{4\gamma \theta \sqrt{N}}{1-\gamma} \right),
\]

Dividing both sides by \( \sum_{t=1}^{T} t = \frac{T(T+1)}{2} \), we obtain
\[
\mathbb{E} \left[ f \left( \frac{2}{T(T+1)} \sum_{t=1}^{T} t w_{t,j} \right) - f(w^*) \right]
\]

\[
+ \frac{\lambda N}{2} \mathbb{E} \|w_{T+1} - w^*\|^2 \leq \frac{8NG^2}{\lambda(T+1)} \left( \frac{3}{2} + \frac{4\gamma \theta \sqrt{N}}{1-\gamma} \right).
\]

Hence, using the weighted average \( \tilde{w}_{T,j} \triangleq \frac{1}{T(T+1)} \sum_{t=1}^{T} t w_{t,j} \) instead of the original SGD iterates, we can achieve a convergence rate of \( O \left( \frac{N^{1/4}}{\sqrt{T}} \right) \), where the SGD iterates achieve a MSD of \( O \left( \frac{N^{1/2}}{T} \right) \). This concludes the proof of the theorem. \( \square \)

**Remark 1.** Algorithm [4] can be generalized into CTA and consensus strategies in a straightforward manner, while the performance guarantee in Theorem 1 still holds up to constant terms, i.e., we still have a convergence rate of \( O \left( \frac{N^{1/4}}{\sqrt{T}} \right) \). As an example, for the consensus strategy, the lines 5–7 of Algorithm 1 is replaced by the following update
\[
w_{t+1,i} = \Pi_{\mathcal{W}} \left( \sum_{j=1}^{N} H_{ji} w_{t,j} - \mu_t g_{t,i} \right).
\]

Hence, we have the following recursion on the parameter vectors
\[
W_t = W_1 H^{t-1} - \sum_{z=1}^{t-1} (X_{t-z} - \mu_{t-z} G_{t-z}) H^{z-1},
\]

instead of the one in [16]. Therefore, the new upper bounds in [17] and [18] are looser by a factor of \( 1/\gamma \).

### 4 Simulations

In this section, we examine the performance of the proposed algorithms for various distributed network topologies, namely the star, the circle, and a random network topologies (which are shown in Fig. 1). In all cases, we have a network of \( N = 20 \) nodes where at each node \( i = 1, \ldots, N \) at time \( t \), we observe the data \( d_{t,i} = w_0^T u_{t,i} + v_{t,i} \), where the regression vector \( u_{t,i} \) and the observation noise \( v_{t,i} \) are generated from i.i.d. zero mean Gaussian processes for all \( t \geq 1 \). The variance of the observation noise is \( \sigma_{v,i}^2 = 0.1 \) for all \( i = 1, \ldots, N \), whereas the auto-covariance matrix of the regression vector \( u_{t,i} \in \mathbb{R}^2 \) is randomly chosen for each node \( i = 1, \ldots, N \) such that the signal-to-noise ratio over the network varies between ~10dB to 10dB. The parameter of interest \( w_0 \in \mathbb{R}^2 \) is randomly chosen from a zero mean Gaussian process and normalized to have a unit norm, i.e., \( ||w_0|| = 1 \). We use the well-known Metropolis combination rule [11] to set the combination weights.

Throughout the experiments, we consider the squared error loss, i.e., \( l(w_{t,i}; u_{t,i}, d_{t,i}) = (d_{t,i} - w_{t,i}^T u_{t,i})^2 \) as our loss function. The notation is as follows: D-CSS-SGD represents the distributed constant step-size SGD algorithm of [9], D-VSS-SGD represents the distributed variable step-size SGD algorithm of [4], D-UW-SGD represents the distributed version of the uniform weighted SGD algorithm of [15], and D-TVW-SGD represents the distributed time variant weighted SGD algorithm proposed in this paper. The step-sizes of the D-CSS-SGD, D-SS-SGD-1, and D-SS-SGD-3 algorithms are set to 0.05, 0.1, and 0.2, respectively, at each node and the learning rates of the D-VSS-SGD and D-UW-SGD algorithms are set to 1/(\( \lambda(t+1) \)), whereas the learning rate of the D-TVW-SGD algorithm is set to 2/(\( \lambda(t+1) \)), where \( \lambda = 0.01 \). These learning rates chosen to guarantee a fair performance comparison between these algorithms according to the corresponding algorithm descriptions stated in this paper and in [4], [15].

In Fig. 4 (the top row), we compare the normalized time accumulated error performances of these algorithms under different network topologies in terms of the global normalized cumulative error (NCE) measure, i.e., \( \text{NCE}(t) = \frac{1}{M^T} \sum_{i=1}^{N} \sum_{t=1}^{T} (d_{t,i} - w_{t,i}^T u_{t,i})^2 \). Additionally, in the bottom row of Fig. 4 we compare the performance of the algorithms in terms of the global MSD measure, i.e., \( \text{MSD}(t) = \frac{1}{M^T} \sum_{i=1}^{N} ||w_0 - w_{t,i}||^2 \). As seen in the Fig. 4, the proposed D-TVW-SGD algorithm significantly outperforms its competitors and achieves a smaller error performance. This superior performance of our algorithm is obtained thanks to the time dependent weighting of the regression parameters, which is used to obtain a faster convergence rate with respect to the rest of the algorithms. Hence, by using a certain time varying weighting of the SGD iterates, we obtain a significantly improved convergence performance compared to the state-of-the-art approaches in the literature.
5 Conclusion

In this paper, we study distributed strongly convex optimization over distributed networks, where the aim is to minimize a sum of unknown convex objective functions. We introduce an algorithm that uses a limited number of gradient oracle calls to these objective functions and achieves the optimal convergence rate of $O \left( \frac{\sqrt{T}}{T} \right)$ after $T$ gradient updates at each node. This performance is obtained by using a certain time dependent weighting of the SGD iterates at each node. The computational complexity and the communication load of the proposed approach is the same with the state-of-the-art methods in the literature up to constant terms. We also prove that the SGD iterates at each node achieve a mean square deviation (MSD) of $O \left( \frac{\sqrt{D}}{T} \right)$ after $T$ gradient oracle calls. We illustrate the superior convergence rate of our algorithm with respect to the state-of-the-art methods in the literature through simulations.

References

[1] A. H. Sayed, “Adaptive networks,” Proc. IEEE, vol. 102, no. 4, pp. 460-497, 2014.
[2] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, “Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior,” IEEE Signal Process. Mag., vol. 30, no. 5, pp. 155–171, 2013.
[3] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, G. B. Giannakis, “Distributed compression-estimation using wireless sensor networks,” IEEE Signal Process. Mag., vol. 23, no. 4, pp. 27–41, 2006.
[4] P. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi, “Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 11, pp. 2483–2493, 2013.
[5] F. Angiulli, S. Basta, S. Lodi, and C. Sartori, “Distributed strategies for mining outliers in large data sets,” IEEE Trans. Knowl. Data Eng., vol. 25, no. 7, pp. 1520–1532, 2013.
[6] X. Zhang, L. Chen, M. Wang, “Efficient parallel processing of distance join queries over distributed graph,” IEEE Trans. Knowl. Data Eng., vol. PP, no. 99, 2014.
[7] J. Luts, “Real-time semiparametric regression for distributed data sets,” IEEE Trans. Knowl. Data Eng., vol. PP, no. 99, 2014.
[8] Z. J. Towfic and A. H. Sayed, “Adaptive penalty-based distributed stochastic convex optimization,” IEEE Trans. Signal Process., vol. 62, no. 15, pp. 3924–3938, 2014.
[9] C. G. Lopes and A. H. Sayed, “Diffusion least-mean squares over adaptive networks: formulation and performance analysis,” IEEE Trans. Signal Process., vol. 56, pp. 3122-3136, 2008.
[10] S. Ram, A. Nedic, and V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” J. Optim. Theory Appl., vol. 147, pp. 516–545, 2010.
[11] A. Nedic and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” IEEE Trans. Autom. Control, vol. 54, no. 1, pp. 48-61, 2009.
[12] E. Hazan, A. Agarwal, and S. Kale, “Logarithmic regret algorithms for online convex optimization,” Machine Learning, vol. 69, no. 2-3, pp. 169-192, 2007.
[13] E. Hazan and S. Kale, “Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization,” J. Mach. Learn. Res., vol. 15, pp. 2489-2512, 2014.
[14] A. Agarwal, P. L. Bartlett, P. Ravikumar, M. J. Wainwright, “Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization,” IEEE Trans. Inf. Theory, vol. 58, no. 9, pp. 3235–3249, 2012.
[15] A. Rakhlin, O. Shamir, and K. Sridharan, “Making gradient descent optimal for strongly convex stochastic optimization,” ICML-J, pp. 449-456, 2012.
[16] S. Lacoste-Julien, M. W. Schmidt, and F. Bach, “A simpler approach to obtaining an O(1/T) convergence rate for the projected stochastic subgradient method,” http://arxiv.org/abs/1212.2002, 2012.