2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Communication-Efficient Local SGD for Over-parametrized Models with Partial Participation

Tiancheng Qin, Jayesh Yevale, and S. Rasoul Etesami

Abstract— **We analyze the convergence rate of Local Stochastic Gradient Descent (SGD) for over-parameterized models, which is at the core of federated learning. In this model, we allow the server to randomly select a subset of agents and communicate with them at each communication round to optimize a global objective function. This captures the realistic scenarios where the communication link between the server and the agents may break down due to random link failures or adversarial attacks. We establish convergence guarantees for smooth objective functions without the convexity assumption that is the first for the regime. We also consider an extension of our results under a different random participation setting over general network structures (rather than a star network) in which an agent participates in the local optimization steps of its neighbors by some edge-dependent probability. We characterize the convergence rate of the proposed algorithm in terms of the number of communication rounds, which confirms the communication efficiency of our methods, and justify our results through a numerical experiment.**

I. Introduction

Distributed optimization has been an increasing trend in the past few decades. This is largely due to recent developments in the control of multi-agent systems or emerging applications in training large-scale machine learning models. The traditional algorithms consist of centralized schemes where the data is accumulated at the server for solving a global optimization problem. However, in practical applications, these traditional algorithms have witnessed serious limitations in handling massive datasets for various reasons, such as limited centralized computational capabilities or other privacy-related concerns. In that regard, distributed optimization provides a powerful framework to handle large-scale optimization problems while respecting privacy and data ownership among the participating agents.

A major issue in distributed optimization is related to communication efficiency between local agents as they collectively optimize a global objective function [1], [2]. One approach to address this issue is to use the Minibatch Stochastic Gradient Descent (SGD) algorithm, in which the model estimates the gradient step at the server by averaging the stochastic gradient steps evaluated at each agent and broadcast to each agent. The Minibatch SGD is well-known and used in various applications [3], [4]. Lately, Local SGD (Federated Averaging) [5], [6], a variant of Minibatch SGD, has been trending and is noticed prominently. It reduces the communication cost by performing multiple local SGD updates at clients before sending the information to the server. During the communication round, the server computes the average of the clients' updates and broadcasts this information back to the clients.

While the Local SGD algorithm assume all agents interact with the server at each communication round to achieve consensus, this assumption may not hold in many practical applications. For instance, the communication links between the agents and the server may fail due to adversarial attacks or random cyber-physical link failures, hindering the local agents not to have access to the most updated information on the server. That imposes a delay between the agents' updates which can further propagate over the entire system through repeated interactions with the server. As a result, the overall optimization performance would degrade due to communication failures. In this work, we aim to evaluate to performance of Local SGD under such a *partial participation* setting to address this communication bottleneck between the server and clients at each round.

A. Related Works

Local SGD is a building block for many optimization algorithms and has been extensively used in applications such as federated learning [7]. The good performance of Local SGD in such applications is noticeable in simulations [8] and prominent in other related problems such as mobile keyboard prediction [9]. Moreover, the theoretical convergence guarantees of Local SGD have been studied recently in various settings [10]–[13]. In [14], the convergence rate of $\mathcal{O}\left(1/nT\right)$ was obtained for strongly convex loss functions, where n is the total number of clients, and T is the total number of iterations. Also, in [15], the convergence rate of $\mathcal{O}(1/\sqrt{nT})$ was obtained for convex loss functions. Moreover, in [16], for non-convex loss functions, the convergence rate of $\mathcal{O}(1/\sqrt{nT})$ is observed. The provided research work is a considerable development in studying the theoretical convergence of the Local SGD algorithm. However, these works do not consider the theoretical performance of the Local SGD for over-parameterization models that satisfy the data interpolation assumption.

On the other hand, the previous works have shortcomings in explaining the quick convergence of Local SGD compared to Minibatch SGD, which was observed significantly in large-scale deep learning models [8]. Also, in the i.i.d. setting, i.e., when the local clients' loss functions are identical, the local SGD performs poorly compared to Minibatch SGD in [17] in terms of lower bounds of convergence rate. This

The authors are with Department of Industrial and Systems Engineering and Coordinated Science Lab, University of Illinois Urbana-Champaign, Urbana, IL, 61801, USA. Email: (tq6, jyeval2, etesami1)@illinois.edu.

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-23-1-0107 and the NSF CAREER Award under Grant No. EPCN-1944403.

rate is hampered even more when used in heterogeneous settings, i.e., when the local clients' loss functions differ. The slower convergence is observed due to "client drift" in [14], and the Minibatch SGD oppresses every known Local SGD analysis [17].

A key observation for using interpolation and over-parameterized models in modern machine learning architecture is discussed in [18]. The over-parameterized models in modern machine learning are well used, in which the empirical loss is nearly zero by interpolation of data. For such models, the convergence rate is provided in [18], [19]. Also, in [20], it has been discussed that the batch size plays a vital role in the performance of SGD; if the selected batch size is more than a certain threshold, the performance guarantee is not affected. In [21], more local steps are incorporated along with an overparameterized model to produce the faster convergence of Local SGD for large-scale optimization problems compared to the faster convergence of Minibatch SGD. To the best of our knowledge, the performance of the Local SGD for overparametrized models with agents' partial participation has not been studied in the past literature.

B. Contributions and Organization

In this work, we assume agents' partial participation in the Local SGD algorithm, i.e., we allow the server to randomly select a subset of agents and communicate with them at each communication round to optimize a global objective function. This captures the realistic scenarios where the communication link between the server and the agents may break down due to random link failures or adversarial attacks. Under the assumption of over-parameterized models and partial participation, for nonconvex functions, the error bound of $\mathcal{O}(K/T) = \mathcal{O}(1/R)$ was obtained, where R is the number of communication rounds. According to our knowledge, this is the first error bound under this setting. Importantly, compared to the nonconvex case, the rates are comparable with local SGD, and no sacrifice is observed in terms of the magnitude of the rate. Before this work, [21] provided convergence of $\mathcal{O}(1/R)$ for nonconvex loss functions assuming full worker participation. Moreover, we extend our results to more general communication networks but under a slightly different partial participation setting where each agent can communicate with a random subset of its neighbors with i.i.d. edge selection probabilities.

In Section II, we describe the problem formulation. In Section III, we describe the proposed algorithm and the main theoretical convergence results for the star communication network with uniform partial participation. In Section IV, we extend our results to general network architectures with random edge-dependent partial participation among the agents. We provide a numerical experiment in Section V, and conclude the paper in Section VI.

II. Problem Formulation

We consider a stochastic distributed optimization problem in which a set $[n] = \{1, 2, \ldots, n\}$ of agents collaboratively want to solve an unconstrained optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{1}$$

where $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\mathcal{D}_i}[f_i(\mathbf{x}, \xi_i)]$ is a local deterministic objective function and ξ_i denotes a random sample that agent i selects from its local data set Ω_i with distribution $\mathcal{D}_i$. Moreover, we assume that there exists a global minimum f^* for the optimization problem (1), and the function $f_i(\mathbf{x}, \xi_i)$ is L-smooth for each $i \in [n]$. In addition, we assume that the gradient of $f_i(\mathbf{x}, \xi_i)$, denoted by $\nabla f_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic gradient of $f_i(x)$, that is

$$\mathbb{E}_{\mathcal{D}_i}[\nabla f_i(\mathbf{x}, \xi_i)] = \nabla f_i(x).$$

In this work, we focus on an over-parameterized setting, i.e., when the model can interpolate the data entirely so that the loss at every data point is minimized simultaneously (usually means zero empirical loss). Following [19], we characterize the over-parameterized setting by the following Strong Growth Condition (SGC).

Assumption 1: (Strong Growth Condition (SGC)). There exists a constant $\rho > 0$ such that for any agent $i \in [n]$ and any $\mathbf{x} \in \mathbb{R}^d$, we have

$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i} \|\nabla f_i(\mathbf{x}, \xi_i)\|^2 \leq \rho \|\nabla f(\mathbf{x})\|^2. \tag{2}$$

It is worth noting that in order to satisfy the SGC, at the global optimum x^* the local gradients computed at every data point must be zero. This indicates that the model can interpolate the data completely such that the loss at every data point is minimized simultaneously .

In fact, Assumption 1 has been well explored in the past literature. For instance, [19] provides some underlying functions satisfying SGC, including the squared-hinge loss function under additional assumptions on the data. Moreover, the SGC condition is often met in modern machine learning applications such as deep neural networks and kernel machines [18], [19]. In the next section, we will proceed to introduce Local SGD with partial participation for solving the optimization problem (1), and establish our main convergence rate results under the SGC Assumption 1.

III. Convergence of Decentralized SGD

A powerful method for solving the distributed optimization problem (1) is the Local SGD algorithm, in which each agent performs local gradient steps and sends the latest model to the central server after every K steps. The server then computes the average of all the agents' parameters and broadcasts the averaged model to all agents. However, in many practical situations, not all agents are available at every communication round for various reasons, such as communication link failures, system delays, or adversarial attacks. In such scenarios, one can consider a variant of the Local SGD with agents' partial participation that we shall consider in this section.

More precisely, in Local SGD with partial participation, the server performs R communication rounds before terminating. During each communication round $r \in \{1, 2, \ldots, R\}$,

Algorithm 1 Local SGD with Partial Participation

Input: $\mathbf{x}_0$, stepsizes η_l, η_g.
for $r = 0, \ldots, R-1$ **do**
  The server samples agents S_r with $|S_r| = S$.
  for each agent $i \in S_r$ in parallel **do**
    $\mathbf{x}^i_{r,0} = \mathbf{x}_r$
    for $k = 0, \ldots, K-1$ **do**
      Sample $\xi^i_{r,k}$, compute $\nabla f_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k})$.
      $\mathbf{x}^i_{r,k+1} = \mathbf{x}^i_{r,k} - \eta_l \nabla f_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k})$
    end for
    Let $\Delta^i_r = \mathbf{x}^i_{r,K} - \mathbf{x}^i_{r,0}$ and send Δ^i_r to server.
  end for
  Server compute $\Delta_r = \frac{1}{S}\sum_{i\in S_r} \Delta^i_r$.
  $\mathbf{x}_{r+1} = \mathbf{x}_r + \eta_g \Delta_r$
end for

the server randomly selects a subset of agents $S_r \subseteq [n]$ satisfying $|S_r| = S$. S_r is randomly and independently selected without replacement such that for each member in S_r, we pick an agent from the entire set $[n]$ uniformly at random with equal probability. The selected agents independently execute local gradient steps for K iterations in the form of

$$\mathbf{x}^i_{r,k+1} = \mathbf{x}^i_{r,k} - \eta_l \nabla f_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k}), k = 0, \ldots, K-1.$$

At the end of each communication round, the server collects the updated information from the participating agents and performs a global update with the aggregation of local updates in the form of

$$\mathbf{x}_{r+1} = \mathbf{x}_r + \eta_g \sum_{i\in S_r} (\mathbf{x}^i_{r,K} - \mathbf{x}_r).$$

The pseudo-code for the Local SGD algorithm with partial participation is provided in Algorithm 1.

A. Convergence Rate Analysis

We now state our main results on the convergence rate of decentralized SGD under over-parameterized settings. To that end, let us first introduce some useful notations. Let $\bar{\Delta}_r$ be the average of agents' updates during communication round r as if all agents are performing local updates, i.e.,

$$\bar{\Delta}_r = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1} \eta_l \nabla f_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k}).$$

Then, we have

$$\mathbb{E}_r[\bar{\Delta}_r] = \mathbb{E}_r[\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{K-1} \eta_l \nabla f_i(\mathbf{x}^i_{r,k})].$$

Moreover, we define the following parameters

$$e_r = \mathbb{E}[f(\mathbf{x}_r)] - f(\mathbf{x}^*),$$
$$h_r = \mathbb{E}\|\nabla f(\mathbf{x}_r)\|^2,$$
$$V_r = \frac{1}{n}\mathbb{E}\sum_{i=1}^{n}\sum_{k=1}^{K} \|\mathbf{x}^i_{r,k} - \mathbf{x}_r\|^2,$$

which represent the expected optimality gap, the expected gradient norm of $\mathbf{x}_r$, and the expected cumulative consensus error among agents during communication round r, respectively. It is worth noting that the randomness for agents' partial participation contains two parts: the random sampling and the stochastic gradient. However, in the following, we shall use $\mathbb{E}_r[\cdot]$ to represent the expectation with respect to both types of randomness, i.e., expectation conditioned on the filtration adapted to the history of random variables $\{\mathbf{x}_r\}$ up till the start of communication round r.

We now state our main result on the convergence rate of Local SGD under over-parameterized settings and partial participation for non-convex functions.

Theorem 1: Let Assumption 1 hold. If we follow Algorithm 1 with stepsize satisfying $\eta_g\eta_l \leq \frac{1}{6LK\rho}$ and $\eta_l \leq \frac{1}{4KL\rho}$, we will have

$$\min_{0\leq r\leq R-1} \mathbb{E}\|\nabla f(\mathbf{x}_r)\|^2 \leq \frac{6(f(\mathbf{x}_0) - f^*)}{K\eta_g\eta_l R}. \tag{3}$$

Specifically, if we choose $\eta_g\eta_l = \frac{1}{6LK\rho}$, we will have

$$\min_{0\leq r\leq R-1} \mathbb{E}\|\nabla f(\mathbf{x}_r)\|^2 \leq \frac{36L\rho(f(\mathbf{x}_0) - f^*)}{R}.$$

Essentially, Theorem 1 establishes an $\mathcal{O}(\frac{1}{R})$ convergence rate of Local SGD for over-parameterized models under the partial participation setting. This improves the previous results of $\mathcal{O}(\frac{1}{\sqrt{R}})$ convergence rate of Local SGD under the partial participation setting but without the over-parameterizerd assumption [14], [22]. To prove Theorem 1, we first establish the following useful lemma.

Lemma 1: Let Assumption 1 hold. If we follow Algorithm 1 with stepsize satisfying $\eta_g\eta_l \leq \frac{1}{6LK\rho}$, we have

$$e_{r+1} \leq e_r - \frac{1}{3}K\eta_g\eta_l h_r + L^2\eta_g\eta_l V_r \tag{4}$$

Proof: For simplicity of notation, we adopt the convention that summations are always over $k \in [K]$ or $i \in [n]$ unless stated otherwise. From the L-smoothness property we have:

$$\begin{aligned}\mathbb{E}_r[f(\mathbf{x}_{r+1})] &= \mathbb{E}_r[f(\mathbf{x}_r + \eta_g\Delta_r)] \\ &\leq f(\mathbf{x}_r) - \langle \nabla f(\mathbf{x}_r), \mathbb{E}_r[\eta_g\Delta_r]\rangle + \frac{L}{2}\mathbb{E}_r[\|\eta_g\Delta_r\|^2].\end{aligned}$$

Using the fact that

$$\mathbb{E}_r[\Delta_r] = \mathbb{E}_r[\frac{1}{S}\sum_{i\in S_r}\Delta^i_r] = \mathbb{E}_r[\Delta^{s_1}_r] = \mathbb{E}_r[\bar{\Delta}_r],$$

we can write,

$$\begin{aligned}&\mathbb{E}_r[f(\mathbf{x}_{r+1})] \\ &\leq f(\mathbf{x}_r) - \langle \nabla f(\mathbf{x}_r), \mathbb{E}_r[\eta_g\bar{\Delta}_r]\rangle + \frac{L}{2}\mathbb{E}_r[\|\eta_g\Delta_r\|^2] \\ &= f(\mathbf{x}_r) - K\eta_g\eta_l\|\nabla f(\mathbf{x}_r)\|^2 \\ &+ \underbrace{\eta_g\langle \nabla f(\mathbf{x}_r), K\eta_l\nabla f(\mathbf{x}_r) - \mathbb{E}_r[\bar{\Delta}_r]\rangle}_{A_1} \\ &+ \underbrace{\frac{L}{2}\mathbb{E}_r[\|\eta_g\Delta_r\|^2]}_{A_2}.\end{aligned} \tag{5}$$

We can bound the first term A_1 as

$$
\begin{aligned}
A_1 &= \eta_g\langle \nabla f(\mathbf{x}_r), K\eta_l \nabla f(\mathbf{x}_r) - \eta_l \frac{1}{n}\sum_{i,k} \nabla f_i(\mathbf{x}^i_{r,k})\rangle \\
&= \langle \sqrt{K\eta_g\eta_l}\nabla f(\mathbf{x}_r), \sqrt{\frac{\eta_g\eta_l}{K}}(\frac{1}{n}\sum_{i,k}(\nabla f_i(x^i_{r,k}) - \nabla f_i(\mathbf{x}_r)))\rangle \\
&\leq \frac{1}{2}K\eta_g\eta_l\|\nabla f(\mathbf{x}_r)\|^2 \\
&\quad + \frac{\eta_g\eta_l}{2K}\|\frac{1}{n}\sum_{i,k}(\nabla f_i(x^i_{r,k}) - \nabla f_i(\mathbf{x}_r))\|^2 \\
&\leq \frac{1}{2}K\eta_g\eta_l\|\nabla f(\mathbf{x}_r)\|^2 \\
&\quad + \frac{\eta_g\eta_l}{2n}\sum_{i,k}\|\nabla f_i(x^i_{r,k}) - \nabla f_i(\mathbf{x}_r)\|^2 \\
&\leq \frac{1}{2}K\eta_g\eta_l\|\nabla f(\mathbf{x}_r)\|^2 + \frac{L^2\eta_g\eta_l}{2n}\sum_{i,k}\|x^i_{r,k} - \mathbf{x}_r\|^2.
\end{aligned}
$$

Moreover, we can bound the second term A_2 as

$$
\begin{aligned}
A_2 &\leq \frac{LK}{2S}\eta_g^2\eta_l^2\mathbb{E}_r[\sum_{k,i\in S_r}\|\nabla f_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k})\|^2] \\
&= \frac{LK}{2S}\eta_g^2\eta_l^2\frac{S}{n}\mathbb{E}_r[\sum_{k,i}\|\nabla f_i(\mathbf{x}^i_{r,k}, \xi^i_{r,k})\|^2] \\
&\overset{(2)}{\leq} \frac{LK\rho}{2n}\eta_g^2\eta_l^2\mathbb{E}_r[\sum_{k,i}\|\nabla f(\mathbf{x}^i_{r,k})\|^2] \\
&\leq \frac{LK\rho}{n}\eta_g^2\eta_l^2\mathbb{E}_r[\sum_{k,i}(\|\nabla f(\mathbf{x}_r)\|^2 + \|\nabla f(\mathbf{x}^i_{r,k}) - \nabla f(\mathbf{x}_r)\|^2)] \\
&\leq LK^2\rho\eta_g^2\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2 \\
&\quad + L^3K\rho\eta_g^2\eta_l^2\frac{1}{n}\mathbb{E}_r[\sum_{i,k}\|x^i_{r,k} - \mathbf{x}_r\|^2],
\end{aligned}
$$

where in the second inequality we have used the SGC Assumption 1. By substituting the bounds obtained for A_1 and A_2 into (5) and taking expectation from both sides, we can write

$$
\begin{aligned}
\mathbb{E}[f(\mathbf{x}_{r+1})] &\leq \mathbb{E}[f(\mathbf{x}_r)] \\
&\quad - K\eta_g\eta_l(\frac{1}{2} - LK\rho\eta_g\eta_l)\mathbb{E}[\|\nabla f(\mathbf{x}_r)\|^2] \\
&\quad + \frac{L^2\eta_g\eta_l}{n}(\frac{1}{2} + LK\rho\eta_g\eta_l)\mathbb{E}[\sum_{i,k}\|x^i_{r,k} - \mathbf{x}_r\|^2] \\
&\leq \mathbb{E}[f(\mathbf{x}_r)] - \frac{1}{3}K\eta_g\eta_l\mathbb{E}[\|\nabla f(\mathbf{x}_r)\|^2] \\
&\quad + \frac{L^2\eta_g\eta_l}{n}\mathbb{E}[\sum_{i,k}\|x^i_{r,k} - \mathbf{x}_r\|^2],
\end{aligned}
$$

where the last inequality is due to the stepsize condition $\eta_g\eta_l \leq \frac{1}{6LK\rho}$. This completes the proof. ∎

Next, we need the following lemma to bound the cumulative consensus error V_r.

Lemma 2: Let Assumptions 1 hold. If we follow Algorithm 1 with stepsize $\eta_l \leq \frac{1}{4KL\rho}$, then,

$$V_r \leq 20K^3\rho\eta_l^2 h_r. \tag{6}$$

Proof: For any $i \in [n]$ and $k \in [K]$, we have

$$
\begin{aligned}
&\mathbb{E}_r[\|\mathbf{x}^i_{r,k} - \mathbf{x}_r\|^2] \\
&\quad = \mathbb{E}_r[\|\mathbf{x}^i_{r,k-1} - \mathbf{x}_r - \eta_l\nabla f_i(\mathbf{x}^i_{r,k-1}, \xi^i_{r,k-1})\|^2] \\
&\quad \leq (\frac{2K}{2K-1})\mathbb{E}_r[\|\mathbf{x}^i_{r,k-1} - \mathbf{x}_r\|^2] \\
&\quad + 2K\eta_l^2\mathbb{E}_r[\|\nabla f_i(\mathbf{x}^i_{r,k-1}, \xi^i_{r,k-1})\|^2] \\
&\quad \overset{(2)}{\leq} (\frac{2K}{2K-1})\mathbb{E}_r[\|\mathbf{x}^i_{r,k-1} - \mathbf{x}_r\|^2] \\
&\quad + 4K\rho\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2 \\
&\quad + 4KL^2\rho\eta_l^2\mathbb{E}_r[\|\mathbf{x}^i_{r,k-1} - \mathbf{x}_r\|^2] \\
&\quad \leq \frac{K}{K-1}\mathbb{E}_r[\|\mathbf{x}^i_{r,k-1} - \mathbf{x}_r\|^2] \\
&\quad + 4K\rho\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2,
\end{aligned}
$$

where the second inequality holds by the SGC Assumption 1 and the last inequality uses the fact that $\eta_l \leq \frac{1}{4KL\rho}$. Unrolling the above inequality recursively, we get

$$
\begin{aligned}
\mathbb{E}_r[\|\mathbf{x}^i_{r,k} - \mathbf{x}_r\|^2] &\leq 4K\rho\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2 \cdot \sum_{p=0}^{k-1}(1 + \frac{1}{K-1})^p \\
&\leq 4K\rho\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2 K((1 + \frac{1}{K-1})^K - 1) \\
&\leq 20K^2\rho\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2.
\end{aligned}
$$

Therefore, we can write

$$\frac{1}{n}\sum_{i,k}\mathbb{E}_r[\|\mathbf{x}^i_{r,k} - \mathbf{x}_r\|^2] \leq 20K^3\rho\eta_l^2\|\nabla f(\mathbf{x}_r)\|^2.$$

Taking unconditional expectations from both sides and using the definition of h_r completes the proof. ∎

Using the above lemmas, we are now ready to prove Theorem 1.

Proof of Theorem 1: By summing inequality (4) over $r = 0, \ldots, R-1$, we obtain

$$
\begin{aligned}
\frac{1}{3}K\eta_g\eta_l\sum_{r=0}^{R-1}h_r &\leq e_0 - e_R + L^2\eta_g\eta_l\sum_{r=0}^{R-1}V_r \\
&\overset{(6)}{\leq} e_0 + L^2\eta_g\eta_l\sum_{r=0}^{R-1}20K^3\rho\eta_l^2 h_r \\
&\leq e_0 + \frac{1}{6}K\eta_g\eta_l\sum_{r=0}^{R-1}h_r,
\end{aligned}
$$

where the second inequality uses Lemma 2 and the last inequality is due to the fact that $\eta_l \leq \frac{1}{11KL\rho}$. Therefore, we have

$$\frac{1}{6}K\eta_g\eta_l\sum_{r=0}^{R-1}h_r \leq e_0,$$

which also implies

$$\min_{0\leq r\leq R-1} h_r \leq \frac{6e_0}{K\eta_g\eta_l R}.$$

This completes the proof of Theorem 1.

IV. Extention to Network Setting

The analysis of the previous section can be viewed as a decentralized SGD over a star network with one center node (server) and n leaf nodes (agents). Therefore, a natural question is on how to extend our convergence rate results to general network structures with over-parameterized models. Unfortunately, extending the uniformly sampled partial participation setting to general network structures seems quite challenging due to the high correlations that may occur between the neighboring agents. For that reason, and motivated by applications such as random link failures in communication networks, in the section, we focus on Local SGD with a different partial participation setting and provide a theoretical convergence guarantee for general network structures.

More precisely, we consider the case where an agent can only exchange information (through gossip averaging) with its neighboring agents in a fixed communication network. However, due to random link failures or adversarial attacks, each link in this communication network has a probability p of failure at any given timestep. We assume that the probability of each link failing is independent of the others. Suppose the fixed underlying communication network is encoded by a symmetric doubly stochastic mixing matrix $W = (w_{ij})$, where w_{ij} corresponds to the weight that agent i is influenced by agent j at each communication. Then, at every iteration $t = 1, 2, \ldots$, Decentralized SGD can be summarized as the following:

- i) the current communication network is observed and encoded into a mixing matrix denoted by W^t.
- ii) each agent performs stochastic gradient updates locally based on $\nabla f_i(\mathbf{x}, \xi_i)$.
- each agent performs consensus operations, where agents average their values with their neighbors.

The weight matrix W^t at time t is constructed in the following way:

$$\forall i \neq j, W^t_{i,j} = \begin{cases} W_{i,j} & \text{if link } (i,j) \text{ does not fail,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\forall i, W^t_{i,i} = 1 - \sum_{j \neq i} W^t_{i,j}$$

The pseudo-code for the decentralized SGD algorithm is provided in Algorithm 2. In order to analyze the convergence rate of Decentralized SGD, we need the following mild assumption on the connectivity of the underlying fixed network.

Assumption 2: The mixing matrix W is symmetric and doubly stochastic, i.e., $w_{ij} \geq 0, w_{ij} = w_{ji}, \mathbf{W}\mathbf{1}_n = 1$. Moreover, there exists a constant $q \in [0, 1)$ such that for all vector $\mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{1}_n^T \mathbf{x} = 0$, we have[1]

$$\|W\mathbf{x}\|^2 \leq q\|\mathbf{x}\|^2. \tag{7}$$

[1]Here, q is related to the spectrum of W.

Algorithm 2 Decentralized SGD

```
Input: x_i^0 = x^0 for i ∈ [n], total number of iterations T, step-size η and the mixing matrix W.
for t = 0, ..., T − 1 do
    Observe current network and encode W^t
    for i = 1, ..., n do
        Sample ξ_i^t, compute g_i^t := ∇f_i(x_i^t, ξ_i^t)
        x_i^{t+1/2} = x_i^t − η g_i^t
        x_i^{t+1} = Σ_{j∈N_i^t} w_{ji}^t x_i^{t+1/2}
    end for
end for
```

We can then derive the following proposition based on Assumption 2.

Proposition 1: Let $\tilde{W} = \mathbb{E}[W_t]$. Then, $\tilde{W}$ is a symmetric and doubly stochastic matrix. Moreover, let $\tilde{q} = (1-p)q + p \in [0, 1)$. Then, for all vectors $\mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{1}_n^T \mathbf{x} = 0$, we have

$$\|W\mathbf{x}\|^2 \leq \tilde{q}\|\mathbf{x}\|^2.$$

Proof: The expected matrix $\tilde{W}$ can be computed as

$$\tilde{W}_{i,j} = (1-p)W_{i,j} \text{ if } i \neq j,$$
$$\tilde{W}_{i,i} = W_{i,i} + p(1 - W_{i,i}).$$

From the above expressions, it is easy to verify that $\tilde{W}$ is a symmetric and doubly stochastic matrix. Moreover, $\forall \mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{1}_n^T \mathbf{x} = 0$, we have

$$\begin{aligned}(\tilde{W}\mathbf{x})_i &= \sum_j \tilde{W}_{i,j} x_j \\ &= \sum_{j \neq i} (1-p) W_{i,j} x_j + (p + (1-p)W_{i,i}) x_i \\ &= (1-p)(W\mathbf{x})_i + p x_i.\end{aligned}$$

Using Jensen's inequality, we can write

$$(\tilde{W}\mathbf{x})_i^2 \leq (1-p)(W\mathbf{x})_i^2 + p x_i^2.$$

Finally, using the above relation together with Assumption 2, we can write

$$\begin{aligned}\|\tilde{W}\mathbf{x}\|^2 &\leq (1-p)\|W\mathbf{x}\|^2 + p\|\mathbf{x}\|^2 \\ &\overset{(7)}{\leq} ((1-p)q + p)\|\mathbf{x}\|^2,\end{aligned}$$

which completes the proof. ∎

Theorem 2: Let Assumptions 1 and 2 hold. If we follow Algorithm 2 with stepsize $\eta = \frac{(1-p)q+p}{28L\rho}$, we will have

$$\min_{0 \leq t \leq T-1} \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^t)\|^2 \leq \frac{100L\rho(f(\mathbf{x}_0) - f^*)}{((1-p)q + p)T}. \tag{8}$$

Proof: The proof follows by using Proposition 1 to Theorem 4 in [23], which gives us the desired convergence result for the Decentralized SGD for over-parameterized models under the random link failure. ∎

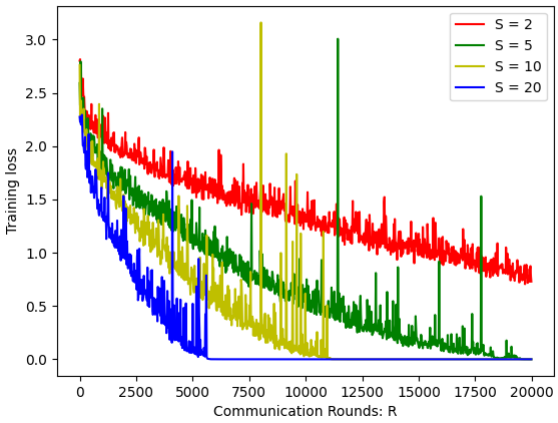

Fig. 1. Training loss versus the number of communication rounds R.

V. Numerical Analysis

In this section, we simulate the training loss of Algorithm 1 versus the number of communication rounds for different values of participation S. We distribute the Cifar10 dataset [24] to $n = 20$ nodes and apply Local SGD to train a ResNet18 neural network [25]. The neural network has 11 million trainable parameters and, after sufficient training rounds, can achieve close to 0 training loss, thus satisfying the interpolation property.

For this set of experiments, we run the Local SGD Algorithm 1 with $K = 10$ number of local steps and for $R = 20000$ communication rounds with different values of participation $S = 2, 5, 10, 20$ per communication round. The training error of the global model along the process has been illustrated in Figure 1, which decreases at nearly the same rate characterized in Theorem 1, as expected.

VI. Conclusion

In this paper, we analyzed the convergence rate of Local SGD with partial agents' participation for over-parameterized models. We established an error bound of $\mathcal{O}(1/R)$ for R number of communication rounds, which is the first error bound under this setting. Importantly, compared to the nonconvex case, the rates are comparable with Local SGD with full participation without any sacrifice in terms of the magnitude of the convergence rate. Moreover, we extended our results to more general communication networks but under a different partial participation setting.

One future research direction is to extend our results to other models of partial participation where the agents are not necessarily uniformly sampled or the link failures are not i.i.d. Moreover, one can consider analyzing the effect of noise (also known as client drift) that occurred due to partial participation during communication with the server.

References

[1] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang, "Zipml: Training linear models with end-to-end low precision, and a little bit of deep learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 4035–4043.

[2] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," *arXiv preprint arXiv:1712.01887*, 2017.

[3] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction using mini-batches." *Journal of Machine Learning Research*, vol. 13, no. 1, 2012.

[4] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan, "Better mini-batch algorithms via accelerated gradient methods," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[5] S. U. Stich, "Local SGD converges fast and communicates little," *arXiv preprint arXiv:1805.09767*, 2018.

[6] L. Mangasarian, "Parallel gradient distribution in unconstrained optimization," *SIAM Journal on Control and Optimization*, vol. 33, no. 6, pp. 1916–1925, 1995.

[7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[9] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage, "Federated learning for mobile keyboard prediction," *arXiv preprint arXiv:1811.03604*, 2018.

[10] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[11] E. Gorbunov, F. Hanzely, and P. Richtárik, "Local SGD: Unified theory and new efficient methods," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3556–3564.

[12] T. Qin, S. R. Etesami, and C. A. Uribe, "Communication-efficient decentralized local sgd over undirected networks," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 3361–3366.

[13] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," *arXiv preprint arXiv:2101.11203*, 2021.

[14] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[15] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local SGD on identical and heterogeneous data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4519–4529.

[16] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 5693–5700.

[17] B. Woodworth, K. K. Patel, S. Stich, Z. Dai, B. Bullins, B. Mcmahan, O. Shamir, and N. Srebro, "Is local SGD better than minibatch SGD?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 334–10 343.

[18] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3325–3334.

[19] S. Vaswani, F. Bach, and M. Schmidt, "Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1195–1204.

[20] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[21] T. Qin, S. R. Etesami, and C. A. Uribe, "Faster convergence of local SGD for over-parameterized models," *arXiv preprint arXiv:2201.12719*, 2022.

[22] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," *arXiv preprint arXiv:2101.11203*, 2021.

[23] T. Qin, S. R. Etesami, and C. A. Uribe, "Decentralized federated learning for over-parameterized models," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 5200–5205.

[24] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.