2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Decentralized Federated Learning for Over-Parameterized Models

Tiancheng Qin, S. Rasoul Etesami, and Cesár A. Uribe

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992924

***Abstract*— Modern machine learning, especially deep learning, features models that are often highly expressive and over-parameterized. They can interpolate the data by driving the empirical loss close to zero. We analyze the convergence rate of decentralized stochastic gradient descent (SGD), which is at the core of decentralized federated learning (DFL), for these over-parameterized models. Our analysis covers the setting of decentralized SGD with time-varying networks, local updates and heterogeneous data. We establish strong convergence guarantees with or without the assumption of convex objectives that either improves upon the existing literature or is the first for the regime.**

***Index Terms*— Decentralized Federated Learning, Decentralized Optimization, Local SGD, Overparameterization**

I. Introduction

Federated Learning [1] has gained much attention as an important learning paradigm where many agents collaboratively train a model while keeping the training data decentralized. Federated Learning has shown great potential in communication efficiency and its capability of preserving data privacy [2], and has exhibited outstanding performance in real-world applications such as keyboard prediction [3] and healthcare [4], [5].

The fundamental and most well-studied Federated Learning algorithm has been the Local Stochastic Gradient Descent (or Local SGD, also known as Federated Averaging) algorithm, where agents communicate with a central server, and during a communication round, a number of local SGD iterations are performed at each agent before the central server computes the average [6], [7]. There has been a number of works studying the theoretical convergence guarantees of Local SGD in various settings [8]–[12].

However, having a central server can sometimes incur a single point of failure or cause communication traffic jam that harms the algorithm's performance [13]. As an alternative, Decentralized Federated Learning (DFL) has gained much popularity recently, where agents conly synchronize with their neighbors in a communication network to achieve model consensus. Numerical experiments have also shown that decentralized algorithms are able to outperform their centralized counterparts [14]. While DFL can be traced back to decentralized optimization and decentralized SGD, which has a long history [14]–[16], a number of recent works have adapted decentralized SGD to the DFL setting [13], [17]–[19]. Specifically, [18] provides a unified theory of Decentralized SGD with time-varying networks and local updates, and [13] further incoporates compressed communication to the framework.

On the other hand, [20] makes a key observation for explaining the fast convergence of SGD in modern machine learning , which says modern machine learning architectures are often highly expressive, they are over-parameterized, and can interpolate the data by driving the empirical loss close to zero. For such over-parameterized models, a faster convergence rate of SGD was proven [20], [21]. Recently, [10] studies the convergence rate of Local SGD for these over-parameterized models and provides better theoretical convergence guarantees for Local SGD in the setting.

Motivated by the above studies, in this paper, we analyze the convergence rate of decentralized SGD for these over-parameterized models. Our analysis covers the setting of decentralized SGD with time-varying networks, local updates and heterogeneous data. We establish strong convergence guarantees with or without the assumption of convex objectives that either improves upon the existing literature or is the first for the regime.

A. Contributions

To summarize our main results, in this work we show:

- For strongly convex loss functions, an error bound of $\mathcal{O}(\exp(-T))$ can be achieved, where T is the total number of iterations. Before our work, the best-known convergence rate was $\mathcal{O}(\exp(-pT/\tau))$ [18], where $p \leq 1, \tau \geq 1$ are parameters related to network connectivity. (Illustration in Assumption 5)
- For general convex loss functions, we establish an error bound of $\mathcal{O}(1/T)$ under a mild data similarity assumption and an error bound of $\mathcal{O}(\tau/pT)$, otherwise. To the best of our knowledge, no convergence rate has been established in the past literature under this setting.
- For nonconvex loss functions, we prove an error bound of $\mathcal{O}(\tau/pT)$. To the best of our knowledge, no theoretical analysis of decentralized SGD in this setting existed in the literature.

The paper is organized as follows. Section II describes the problem statement, and assumptions. Section III states our main results and proof sketches. Section IV describes concluding remarks and future work. Omitted proofs are relegated to Appendix I.

This material is based upon work supported by the National Science Foundation under Grants No. EPCN-1944403, No. 2211815, and No. 2213568.

Tiancheng Qin and S. Rasoul Etesami are with the Department of Industrial and Enterprise Systems Engineering, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA. Email: (tq6, etesami1)@illinois.edu.

Cesár A. Uribe is with the Department of Electrical and Computer Engineering at Rice University, Houston, TX, 77005, USA. Email: cauribe@rice.edu.

II. Problem Formulation

We formalize the problem of n agents $[n] = \{1, 2, \ldots, n\}$ collaboratively learning an over-parameterized model as the following decentralized stochastic optimization problem:

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

where the function $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi_i\sim\mathcal{D}_i} f_i(\mathbf{x}, \xi_i)$ denotes the local loss function, ξ_i is a stochastic sample that agent i has access to, and $\mathcal{D}_i$ denotes the local data distribution over the sample space Ω_i of agent i.

Following previous works such as [8], [9], we assume throughout the paper that $f(\mathbf{x})$ is bounded below by $f^\star$ (i.e., a global minimum exists), $f_i(\mathbf{x}, \xi_i)$ is L-smooth for every $i \in [n]$, and $\nabla f_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic gradient of $f_i(\mathbf{x})$. Moreover, for some of our results, we will require functions $f_i(\mathbf{x}, \xi_i)$ to be μ-strongly convex with respect to the parameter $\mathbf{x}$ as defined next.

Assumption 1 (μ-strong convexity): There exists a constant $\mu \geq 0$, such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, i \in [n]$, and $\xi_i \in \Omega_i$, we have

$$f_i(\mathbf{x}, \xi_i) \geq f_i(\mathbf{y}, \xi_i) + \langle \nabla f_i(\mathbf{y}, \xi_i), \mathbf{x} - \mathbf{y}\rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2. \tag{2}$$

If $\mu = 0$, we simply say that each f_i is convex.

We characterize the over-parameterized setting, i.e., when the model can *interpolate* the data completely such that the loss at every data point is minimized simultaneously (usually means zero empirical loss) by the following two assumptions as in [20], [21]:

Assumption 2 (Interpolation): Let $\mathbf{x}^\star \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(\mathbf{x})$. Then, $\nabla f_i(\mathbf{x}^\star, \xi_i) = 0, \; \forall i \in [n], \; \xi_i \in \Omega_i$.

Assumption 3 (Strong Growth Condition (SGC)): There exists constant $\rho \geq 1$ such that $\forall \mathbf{x} \in \mathbb{R}^d, \; i \in [n]$,

$$\mathbb{E}_{\xi_i\sim\mathcal{D}_i}\|\nabla f_i(\mathbf{x}, \xi_i)\|^2 \leq \rho\|\nabla f(\mathbf{x})\|^2. \tag{3}$$

Notice that for the functions to satisfy SGC, local gradients at every data point must all be zero at the optimum $x^\star$. This means Assumption 3 implies Assumption 2.

Assumption 2 is commonly satisfied by modern machine learning model such as deep neural networks [22] and kernel machines [23]. [21] discussed functions satisfying Assumption 3 and showed that under additional assumptions on the data, the squared-hinge loss satisfies the assumption.

Finally, as in [10] we consider the following assumption that allows us to measure dissimilarity among local functions.

Assumption 4 (c-Bounded Optimality Gap (c-BOG)): For some constant $c \in [0, 1]$, we have

$$f_i(\mathbf{x}) - f_i^* \geq c(f(\mathbf{x}) - f^*), \quad \forall \mathbf{x} \in \mathbb{R}^d, \; i \in [n], \tag{4}$$

where $f_i^* = \min_{\mathbf{x}\in\mathbb{R}^d} f_i(\mathbf{x})$.

We note that Assumption 4 always holds if $c = 0$, and as the local loss functions become more similar, it will hold for larger values of c. In the case of homogeneous local loss functions, i.e., $f_i = f, \; \forall i$, Assumption 4 holds with $c = 1$.

Algorithm 1 Decentralized Local SGD

1: Input: $\mathbf{x}_i^0 = \mathbf{x}^0$ for $i \in [n]$, total number of iterations T, step-size η and the mixing matrix sequence $\{\mathbf{W}^t\}_{t=0}^{T-1}$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: **for** $i = 1, \ldots, n$ **do**
4: Sample ξ_i^t, compute $\mathbf{g}_i^t := \nabla f_i(\mathbf{x}_i^t, \xi_i^t)$
5: $\mathbf{x}_i^{t+\frac{1}{2}} = \mathbf{x}_i^t - \eta\mathbf{g}_i^t$
6: $\mathbf{x}_i^{t+1} = \sum_{j\in\mathcal{N}_i^t} w_{ji}\mathbf{x}_i^{t+\frac{1}{2}}$
7: **end for**
8: **end for**

III. Convergence of Decentralized SGD

This section reviews decentralized SGD and then analyzes its convergence rate under the over-parameterized setting.

In decentralized SGD, each agent can only exchange information (through gossip averaging) with its neighboring agents in the communication network. In every iteration t, the algorithm does the following: i) each agent performs stochastic gradient updates locally based on $\nabla f_i(\mathbf{x}, \xi_i)$, which is an unbiased estimation of $\nabla f_i(\mathbf{x})$, and ii) each agent performs consensus operations, where agents average their values with their neighbors.

The communication network at time t is encoded by a mixing matrix $\mathbf{W}^t$, where the neighbors of agent i at iteration t are denoted as $\mathcal{N}_i^t := \{j : \; w_{ij}^t > 0\}$.

The pseudo-code for the decentralized SGD algorithm is provided in Algorithm 1.

If we write all the variables and the gradient values in a matrix form,

$$\mathbf{X}^t \triangleq \left[\mathbf{x}_1^t, \cdots, \mathbf{x}_n^t\right] \in \mathbb{R}^{d\times n},$$
$$\mathbf{G}^t \triangleq \left[\mathbf{g}_1^t, \cdots, \mathbf{g}_n^t\right] \in \mathbb{R}^{d\times n},$$

then the update of decentralized SGD algorithm can be compactly written as:

$$\mathbf{X}^{t+1} = (\mathbf{X}^t - \eta\mathbf{G}^t)\mathbf{W}^t. \tag{5}$$

As in [18], we make the following mild assumptions on the mixing matrix $\{\mathbf{W}^t\}_{t=0}^{T-1}$ that reflects the setting of decentralized SGD with time-varying networks and local updates.

Assumption 5: The mixing matrices $\{\mathbf{W}^t\}_{t=0}^{T-1}$ are symmetric and doubly stochastic, i.e., $w_{ij}^t \geq 0, w_{ij}^t = w_{ji}^t, \mathbf{W}^t\mathbf{1}_n = 1, \; \forall t \in [T-1]$. Moreover, there exists two constants $p \in (0, 1]$ and integer $\tau \geq 1$ such that for all matrices $\mathbf{X} \in \mathbb{R}^{d\times n}$ and all integers $l \in \{0, \ldots, T/\tau\}$,

$$\|\mathbf{X}\mathbf{W}_{l,\tau} - \bar{\mathbf{X}}\|_F^2 \leq (1-p)\|\mathbf{X} - \bar{\mathbf{X}}\|_F^2, \tag{6}$$

where $\mathbf{W}_{l,\tau} = \mathbf{W}^{(l+1)\tau-1}\cdots\mathbf{W}^{l\tau}$ and $\bar{\mathbf{X}} := \mathbf{X}\frac{\mathbf{1}\mathbf{1}^T}{n}$.

Assumption 5 is a very mild assumption on the connectivity of the underlying communication network structure among the agents, and the setting incorporates Local SGD [9], Periodic Decentralized SGD [17] and Local Decentralized SGD [19] as special cases. We refer to [18] to a detailed discussion about the examples covered in the setting.

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 $\bar{\mathbf{x}}^{(t)}$ and $\bar{\mathbf{g}}^{(t)}$ be the average of agents' iterates and the average of their stochastic gradients at time t, respectively, i.e.,

$$\bar{\mathbf{x}}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^t, \quad \bar{\mathbf{g}}^t = \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i^t.$$

Moreover, define the following parameters

$$r_t = \mathbb{E}\|\bar{\mathbf{x}}^t - \mathbf{x}^*\|^2, \quad V_t = \frac{1}{n}\mathbb{E}\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2,$$
$$e_t = \mathbb{E}[f(\bar{\mathbf{x}}^t)] - f(\mathbf{x}^*), \quad h_t = \|\nabla f(\bar{\mathbf{x}}^t)\|^2,$$

which represent, respectively, the expected distance of the averaged iterates at time t to the optimum solution, the expected consensus error among agents at time t, the expected optimality gap and the gradient norm of the average iterates at time t.

For strongly convex loss functions we have the following rate.

Theorem 1 (Strongly convex functions): Let Assumptions 1, 2, 4 and 5 hold with $\mu > 0$. If we follow algorithm 1 with stepsize $\eta = 1/L$, we will have

$$\mathbb{E}\|\bar{\mathbf{x}}^{(T)} - \mathbf{x}^*\|^2 \leq \big(1 - \frac{\mu}{L}\big)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2, \tag{7}$$

where $\bar{\mathbf{x}}^{(t)} \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ is the average of all nodes iterates at time step t.

To prove Theorem 1, we need to first state the following proposition and lemma.

Proposition 1: For any $\mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{x}' \in \mathbb{R}^d$ and $\mathbf{W}$ which is a symmetric and doubly stochastic matrix, we have

$$\|\mathbf{X}\mathbf{W} - \mathbf{x}'\mathbf{1}_n\|_F^2 = \|(\mathbf{X} - \mathbf{x}'\mathbf{1}_n)\mathbf{W}\|_F^2 \leq \|\mathbf{X} - \mathbf{x}'\mathbf{1}_n\|_F^2. \tag{8}$$

Lemma 1: Let Assumptions 1 and 2 hold with $\mu > 0$. If we follow Algorithm 1 with stepsize $\eta = \frac{1}{L}$, we will have

$$\mathbb{E}_{\xi_i^t}\|\mathbf{x}_i^t - \eta\mathbf{g}_i^t - \mathbf{x}^*\|^2 \leq (1 - \frac{\mu}{L})\|\mathbf{x}_i^t - \mathbf{x}^*\|^2. \tag{9}$$

Now we state the proof of Theorem 1.

Proof: Let $\mathbf{x}^* \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(\mathbf{x})$. Using Proposition 1 and Lemma 1 we have,

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}^{t+1} - \mathbf{x}^*\mathbf{1}_n\|_F^2 &\overset{(5)}{=} \mathbb{E}\|(\mathbf{X}^t - \eta\mathbf{G}^t)\mathbf{W}_t - \mathbf{x}^*\mathbf{1}_n\|_F^2 \\
&\overset{(8)}{\leq} \mathbb{E}\|\mathbf{X}^t - \eta\mathbf{G}^t - \mathbf{x}^*\mathbf{1}_n\|_F^2 \\
&= \mathbb{E}[\sum_{i=1}^n \mathbb{E}_{\xi_i^t}\|\mathbf{x}_i^t - \eta\mathbf{g}_i^t - \mathbf{x}^*\|^2] \\
&\overset{(9)}{\leq} (1 - \frac{\mu}{L})\mathbb{E}[\sum_{i=1}^n \|\mathbf{x}_i^t - \mathbf{x}^*\|^2] \\
&= (1 - \frac{\mu}{L})\mathbb{E}\|\mathbf{X}^t - \mathbf{x}^*\mathbf{1}_n\|_F^2.
\end{aligned}$$

Therefore, from Jensen's inequality we have

$$\begin{aligned}
\mathbb{E}\|\bar{\mathbf{x}}^{(T)} - \mathbf{x}^*\|^2 &\leq \frac{1}{n}\mathbb{E}\|\mathbf{X}^T - \mathbf{x}^*\mathbf{1}_n\|_F^2 \\
&\leq \frac{1}{n}(1 - \frac{\mu}{L})\mathbb{E}\|\mathbf{X}^{T-1} - \mathbf{x}^*\mathbf{1}_n\|_F^2 \\
&\leq \cdots \leq \frac{1}{n}\big(1 - \frac{\mu}{L}\big)^T \mathbb{E}\|\mathbf{X}^0 - \mathbf{x}^*\mathbf{1}_n\|_F^2 \\
&= \big(1 - \frac{\mu}{L}\big)^T \|\mathbf{x}^0 - \mathbf{x}^*\|^2
\end{aligned}$$

■

The next theorem states the convergence rates for general convex loss functions when mild assumption on data similarity is satisfied, i.e. Assumption 4 is satisfied with $c > 0$.

Theorem 2 (General convex functions): Let Assumptions 1, 2, 4 and 5 hold with $\mu = 0, c > 0$. If we follow Algorithm 1 with stepsize $\eta = \frac{1}{2L}$, and let $\hat{\mathbf{x}}^T \triangleq \frac{1}{T}\sum_{i=0}^{T-1} \bar{\mathbf{x}}^{(t)}$ we have

$$\mathbb{E}[f(\hat{\mathbf{x}}^T) - f^*] \leq \frac{2L\|\mathbf{x}^0 - \mathbf{x}^*\|^2}{cT}. \tag{10}$$

To prove Theorem 2, we need to first state the following proposition and lemma.

Proposition 2: Let Assumptions 1, 2, and 4 hold with $\mu \geq 0$. Let $\mathbf{x}^* \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(\mathbf{x})$. For all $\mathbf{x}_1, \ldots \mathbf{x}_n \in \mathbb{R}^d$ and $\bar{\mathbf{x}} \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$, we have

$$\frac{1}{n}\sum_{i=1}^n (f_i(\mathbf{x}_i) - f_i(\mathbf{x}^*)) \geq c(f(\bar{\mathbf{x}}) - f(\mathbf{x}^*)). \tag{11}$$

Lemma 2: Let Assumptions 1, 2, and 4 hold with $\mu = 0$. If we follow Algorithm 1 with stepsize $\eta_t = \frac{1}{2L}$, we will have

$$\mathbb{E}_{\xi_i^t}\|\mathbf{x}_i^t - \eta\mathbf{g}_i^t - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_i^t - \mathbf{x}^*\|^2 - \frac{1}{2L}(f_i(\mathbf{x}_i^t) - f_i(\mathbf{x}^*)) \tag{12}$$

Now we state the proof of Theorem 2.

Proof: Let $\mathbf{x}^* \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(\mathbf{x})$. Using Proposition 1,2, Lemma 2 and similar to the proof of Theorem 1 we have,

$$\begin{aligned}
\mathbb{E}\|\mathbf{X}^{t+1} - \mathbf{x}^*\mathbf{1}_n\|_F^2 &\leq \mathbb{E}[\sum_{i=1}^n \mathbb{E}_{\xi_i^t}\|\mathbf{x}_i^t - \eta\mathbf{g}_i^t - \mathbf{x}^*\|^2] \\
&\overset{(12)}{\leq} \mathbb{E}[\sum_{i=1}^n \|\mathbf{x}_i^t - \mathbf{x}^*\|^2 - \frac{1}{2L}(f_i(\mathbf{x}_i^t) - f_i(\mathbf{x}^*))] \\
&\overset{(11)}{\leq} \mathbb{E}\big[\|\mathbf{X}^t - \mathbf{x}^*\mathbf{1}_n\|_F^2 - \frac{cn}{2L}(f(\bar{\mathbf{x}}^t) - f(\mathbf{x}^*))\big].
\end{aligned}$$

Summing over $t = 0, \ldots, T-1$ and notice $\mathbb{E}\|\mathbf{X}^T - \mathbf{x}^*\mathbf{1}_n\|_F^2 \geq 0$ we have

$$\frac{cn}{2L}\sum_{t=0}^{T-1} \mathbb{E}[f(\bar{\mathbf{x}}^t) - f(\mathbf{x}^*)] \leq \|\mathbf{X}^0 - \mathbf{x}^*\mathbf{1}_n\|_F^2 = n\|\mathbf{x}^0 - \mathbf{x}^*\|^2.$$

Theorem 2 now follows from Jensen's inequality. ■

When local data can be arbitrarily dissimilar, i.e. Assumption 4 is satisfied with $c = 0$, we provide the following convergence rate for general convex loss functions.

Theorem 3 (General convex functions): Let Assumptions 1, 2, 4 and 5 hold with $\mu = 0, c = 0$. If we follow Algorithm 1 with stepsize $\eta = \frac{1}{2L}$, and let $\hat{\mathbf{x}}^T \triangleq \frac{1}{T}\sum_{i=0}^{T-1} \bar{\mathbf{x}}^{(t)}$ we have

$$\mathbb{E}[f(\hat{\mathbf{x}}^T) - f^*] \leq \frac{40L\tau\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2}{pT}. \tag{13}$$

To prove Theorem 3, we need to first state the following lemma.

Lemma 3: Let Assumptions 1, 2 hold with $\mu = 0$. If we follow Algorithm 1 with stepsize $\eta \leq \frac{1}{4L}$, then,

$$r_{t+1} \leq r_t - \eta e_t + \frac{3}{2}L\eta V_t. \tag{14}$$

Lemma 4: Let Assumptions 1,2,5 hold with $\mu = 0$. If we follow Algorithm 1 with stepsize $\eta = \frac{p}{28L\tau}$, then,

$$V_t \leq \frac{30L\tau}{p}\eta^2 \sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor \frac{t-j}{\tau} \rfloor} e_j. \tag{15}$$

Now we state the proof of Theorem 3.

Proof: Summing (14) over $t = 0, \ldots, T-1$ we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\eta e_t \leq r_0 - r_T + \frac{3}{2}L\eta\sum_{t=0}^{T-1}V_t \\
\overset{(15)}{\leq}& r_0 + \frac{3}{2}L\eta\sum_{t=0}^{T-1}\frac{30L\tau}{p}\eta^2\sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor \frac{t-j}{\tau} \rfloor}e_j \\
=& r_0 + \frac{45L^2\tau\eta^3}{p}\sum_{j=0}^{T-2}e_j\sum_{t=j+1}^{T-1}(1-\frac{p}{4})^{\lfloor \frac{t-j}{\tau} \rfloor} \\
\leq& r_0 + \frac{180L^2\tau^2\eta^3}{p^2}\sum_{j=0}^{T-2}e_j.
\end{aligned}$$

Substituting $\eta = \frac{p}{28L\tau}$ and notice $\frac{1}{28} \times (1 - \frac{180}{28^2}) \geq \frac{1}{40}$, we have,

$$\frac{p}{40L\tau}\sum_{t=0}^{T-1}e_t \leq r_0.$$

Theorem 3 now follows from Jensen's inequality. ∎

Finally, for the case of non-convex loss functions, we have the following result.

Theorem 4 (Non-convex functions): Let Assumption 3 hold. If we follow Algorithm 1 with stepsize $\eta = \frac{p}{28L\tau\rho}$, we will have

$$\min_{0\leq t\leq T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^t)\|^2 \leq \frac{100L\tau\rho(f(\mathbf{x}_0)-f^*)}{pT}. \tag{16}$$

To prove Theorem 4, we need to first state the following lemma.

Lemma 5: Let Assumption 3 hold. If we follow Algorithm 1 with stepsize $\eta \leq \frac{1}{6L\rho}$, we have

$$e_{t+1} \leq e_t - \frac{1}{3}\eta h_t + \frac{2}{3}\eta L^2 V_t \tag{17}$$

Lemma 6: Let Assumptions 3 and 5 hold. If we follow Algorithm 1 with stepsize $\eta = \frac{p}{28L\tau\rho}$, then,

$$V_t \leq \frac{15\tau\rho}{p}\eta^2\sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor \frac{t-j}{\tau} \rfloor}h_j. \tag{18}$$

Now we state the proof of Theorem 4.

Proof: Summing (17) over $t = 0, \ldots, T-1$ we have

$$\begin{aligned}
&\frac{1}{3}\sum_{t=0}^{T-1}\eta h_t \leq e_0 - e_T + \frac{2}{3}L^2\eta\sum_{t=0}^{T-1}V_t \\
\overset{(18)}{\leq}& e_0 + \frac{2}{3}L^2\eta\sum_{t=0}^{T-1}\frac{15\tau\rho}{p}\eta^2\sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor \frac{t-j}{\tau} \rfloor}h_j \\
=& e_0 + \frac{10L^2\tau\rho\eta^3}{p}\sum_{j=0}^{T-2}h_j\sum_{t=j+1}^{T-1}(1-\frac{p}{4})^{\lfloor \frac{t-j}{\tau} \rfloor} \\
\leq& e_0 + \frac{40L^2\tau^2\rho\eta^3}{p^2}\sum_{j=0}^{T-2}h_j.
\end{aligned}$$

Substituting $\eta = \frac{p}{28L\tau\rho}$ and notice $\rho \geq 1$, $\frac{1}{28*3} \times (1 - \frac{120}{28^2}) \geq \frac{1}{100}$, we have,

$$\frac{p}{100L\tau\rho}\sum_{t=0}^{T-1}h_t \leq e_0 \quad \Rightarrow \quad \min_{0\leq t\leq T-1}h_t \leq \frac{100L\tau\rho e_0}{pT}.$$

Thus we proved Theorem 4. ∎

IV. Conclusion

Inspired by the supurior performance of DFL algorithms both in practice and in numerical experiments, in this paper we theoretically analyzed the convergence rate of decentralized SGD for over-parameterized models. Our analysis covers the setting of decentralized SGD with time-varying networks, local updates and heterogeneous data. We established strong convergence guarantees with or without the assumption of convex objectives that either improves upon the existing literature or is the first for the regime.

Appendix I: Omitted Proofs

For the proof of Lemma 1, 2 and 5 we refer the reader to [10].

We first state some propositions that would be useful for our proof.

Proposition 3: Let $f : \mathbb{R}^d \to \mathbb{R}$ be an L-smooth function and $\mathbf{x}^* \in \operatorname{argmin}_{x\in\mathbb{R}^d} f(\mathbf{x})$. Then,

$$\frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \tag{19}$$

Proof: Proof can be found in [24]. ∎

Proposition 4: Let Assumptions 1 and 2 hold with $\mu = 0$. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{g}_i^t\|^2 \leq 4Le_t + 2L^2V_t. \tag{20}$$

Proof: We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{g}_i^t\|^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{x}_i^t,\xi_i^t)\|^2 \\
\leq& \frac{2}{n}\sum_{i=1}^{n}(\mathbb{E}\|\nabla f_i(\bar{\mathbf{x}}^t,\xi_i^t)\|^2 + \mathbb{E}\|\nabla f_i(\bar{\mathbf{x}}^t,\xi_i^t) - \nabla f_i(\mathbf{x}_i^t,\xi_i^t)\|^2) \\
\overset{(19)}{\leq}& \frac{2}{n}\big(2L\sum_{i=1}^{n}\mathbb{E}[f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}^*)] + \sum_{i=1}^{n}L^2\mathbb{E}\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2\big) \\
=& 4Le_t + 2L^2V_t.
\end{aligned}$$

∎

Proposition 5: Let Assumptions 3 hold. Then,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{g}_i^t\|^2 \leq 2\rho h_t + 2L^2\rho V_t. \tag{21}$$

Proof: We have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{g}_i^t\|^2 \overset{(3)}{\leq} \frac{\rho}{n}\mathbb{E}\sum_{i=1}^{n}\|\nabla f(\mathbf{x}_i^t)\|^2$$
$$\leq \frac{2\rho}{n}\mathbb{E}\sum_{i=1}^{n}\|\nabla f(\bar{\mathbf{x}}^t)\|^2 + \frac{2\rho}{n}\mathbb{E}\sum_{i=1}^{n}\|\nabla f(\mathbf{x}_i^t) - \nabla f(\bar{\mathbf{x}}^t)\|^2$$
$$\leq 2\rho h_t + 2L^2\rho V_t.$$

■

Proof of Lemma 3

Proof: Using Proposition 4 we have

$$r_{t+1} = \mathbb{E}\|\bar{\mathbf{x}}^{t+1} - \mathbf{x}^*\|^2 = \mathbb{E}\|\bar{\mathbf{x}}^t - \eta\bar{\mathbf{g}}^t - \mathbf{x}^*\|^2$$
$$= r_t + \frac{\eta^2}{n^2}\mathbb{E}\|\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t, \xi_i^t)\|^2 - \frac{2\eta}{n}\mathbb{E}[\sum_{i=1}^{n}\langle\nabla f_i(\mathbf{x}_i^t), \bar{\mathbf{x}}^t - \mathbf{x}^*\rangle]$$
$$\overset{(20)}{\leq} r_t + 4L\eta^2 e_t + 2L^2\eta^2 V_t - \frac{2\eta}{n}\underbrace{\mathbb{E}[\sum_{i=1}^{n}\langle\nabla f_i(\mathbf{x}_i^t), \bar{\mathbf{x}}^t - \mathbf{x}^*\rangle]}_{T_1}.$$

For T_1 we can bound it by

$$T_1 = \mathbb{E}[\sum_{i=1}^{n}\langle\nabla f_i(\mathbf{x}_i^t), \bar{\mathbf{x}}^t - \mathbf{x}^*\rangle]$$
$$= \mathbb{E}[\sum_{i=1}^{n}\langle\nabla f_i(\mathbf{x}_i^t), \bar{\mathbf{x}}^t - \mathbf{x}_i^t + \mathbf{x}_i^t - \mathbf{x}^*\rangle]$$
$$\geq \mathbb{E}\sum_{i=1}^{n}\big(f_i(\bar{\mathbf{x}}^t) - f_i(\mathbf{x}_i^t) - \frac{L}{2}\|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2 + f_i(\mathbf{x}_i^t) - f_i(\mathbf{x}^*)\big)$$
$$= ne_t - \frac{Ln}{2}V_t.$$

Therefore we have

$$r_{t+1} \leq r_t + 4L\eta^2 e_t + 2L^2\eta^2 V_t - \frac{2\eta}{n}(ne_t - \frac{Ln}{2}V_t)$$
$$\overset{(\eta\leq\frac{1}{4L})}{\leq} r_t - \eta e_t + \frac{3}{2}L\eta V_t.$$

■

Proof of Lemma 4

Proof: Let $m = \lfloor\frac{t}{\tau}\rfloor - 1$, and using the fact that $\|\mathbf{X} - \bar{\mathbf{X}}\|_F^2 \leq \|\mathbf{X}\|_F^2,\ \forall\mathbf{X}$ and $\|a+b\|^2 \leq (1+q)\|a\|^2 + (1+\frac{1}{q})\|b\|^2,\ \forall q > 0$, we have

$$nV_t = \mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^{m\tau} - (\bar{\mathbf{X}}^{m\tau} - \bar{\mathbf{X}}^t)\|_F^2 \leq \mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^{m\tau}\|_F^2$$
$$= \mathbb{E}\|\mathbf{X}^{m\tau}\prod_{i=m\tau}^{t-1}\mathbf{W}^i - \eta\sum_{i=m\tau}^{t-1}\mathbf{G}^i\prod_{j=i}^{t-1}\mathbf{W}^j - \bar{\mathbf{X}}^{m\tau}\|_F^2$$
$$\overset{(6)}{\leq}(1+\frac{p}{2})(1-p)nV_{m\tau} + (1+\frac{2}{p})\mathbb{E}\|\eta\sum_{i=m\tau}^{t-1}\mathbf{G}^i\prod_{j=i}^{t-1}\mathbf{W}^j\|_F^2$$
$$\leq (1-\frac{p}{2})nV_{m\tau} + (1+\frac{2}{p})(t-m\tau)\eta^2\sum_{i=m\tau}^{t-1}\mathbb{E}\|\mathbf{G}^t\|_F^2$$
$$\overset{(20)}{\leq}(1-\frac{p}{2})nV_{m\tau} + \frac{6\tau}{p}\eta^2\sum_{i=m\tau}^{t-1}n(4Le_i + 2L^2V_i).$$

Substituting $\eta = \frac{p}{28L\tau}$ and $p \leq 1$ we have

$$V_t \leq (1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1}V_i + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i. \quad (22)$$

Similarly, let $m = \lfloor\frac{t}{\tau}\rfloor$ we can have

$$V_t \leq (1+\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1}V_i + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i. \quad (23)$$

Now, using (23) we can prove by recursion that for $m\tau \leq i \leq (m+1)\tau - 1$ we have

$$(1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1}V_i + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i$$
$$\leq\big((1-\frac{p}{2})V_{m\tau} + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i\big)(1+\frac{p}{16\tau})^{i-m\tau}.$$

Similarly, using (22) we can prove by recursion that for $(m+1)\tau \leq i \leq (m+2)\tau - 1$ we have

$$(1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1}V_i + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i$$
$$\leq\big((1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{(m+1)\tau-1}V_i$$
$$+\frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i\big)(1+\frac{p}{64\tau})^{i-(m+1)\tau}.$$

Therefore, let $m = \lfloor\frac{t}{\tau}\rfloor - 1$ we have

$$V_t \leq \big((1-\frac{p}{2})V_{m\tau} + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i\big)(1+\frac{p}{16\tau})^{\tau}(1+\frac{p}{64\tau})^{\tau}$$
$$\leq \big((1-\frac{p}{2})V_{m\tau} + \frac{24\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i\big)(1+\frac{p}{4})$$
$$\leq (1-\frac{p}{4})V_{m\tau} + \frac{30\tau L\eta^2}{p}\sum_{i=m\tau}^{t-1}e_i$$
$$\leq \cdots \leq V_0 + \frac{30L\tau}{p}\eta^2\sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor\frac{t-j}{\tau}\rfloor}e_j$$
$$= \frac{30L\tau}{p}\eta^2\sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor\frac{t-j}{\tau}\rfloor}e_j.$$

■

Proof of Lemma 6

Proof: The proof of Lemma 6 is similar to the proof of Lemma 4, where the difference is that instead of using Proposition 4 to bound $\mathbb{E}\|\mathbf{G}^t\|_F^2$ we use Proposition 5 to bound it. Actually, similar to the proof of Lemma 4 we have,

$$nV_t \leq (1-\frac{p}{2})nV_{m\tau} + (1+\frac{2}{p})(t-m\tau)\eta^2\sum_{i=m\tau}^{t-1}\mathbb{E}\|\mathbf{G}^t\|_F^2$$
$$\overset{(21)}{\leq} (1-\frac{p}{2})nV_{m\tau} + \frac{6\tau}{p}\eta^2\sum_{i=m\tau}^{t-1}n(2\rho h_t + 2L^2\rho V_t).$$

Substituting $\eta = \frac{p}{28L\tau\rho}$ and $p\rho \geq 1$ we have

$$V_t \leq (1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1} V_i + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i. \quad (24)$$

Similarly, let $m = \lfloor \frac{t}{\tau} \rfloor$ we can have

$$V_t \leq (1+\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1} V_i + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i. \quad (25)$$

Now, using (25) we can prove by recursion that for $m\tau \leq i \leq (m+1)\tau - 1$ we have

$$\begin{aligned}&(1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1} V_i + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i \\ \leq&\big((1-\frac{p}{2})V_{m\tau} + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i\big)(1+\frac{p}{16\tau})^{i-m\tau}.\end{aligned}$$

Similarly, using (24) we can prove by recursion that for $(m+1)\tau \leq i \leq (m+2)\tau - 1$ we have

$$\begin{aligned}&(1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{t-1} V_i + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i \\ \leq&\big((1-\frac{p}{2})V_{m\tau} + \frac{p}{64\tau}\sum_{i=m\tau}^{(m+1)\tau-1} V_i \\ &+\frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i\big)(1+\frac{p}{64\tau})^{i-(m+1)\tau}.\end{aligned}$$

Therefore, let $m = \lfloor \frac{t}{\tau} \rfloor - 1$ we have

$$\begin{aligned}V_t \leq& \big((1-\frac{p}{2})V_{m\tau} + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i\big)(1+\frac{p}{16\tau})^{\tau}(1+\frac{p}{64\tau})^{\tau} \\ \leq& \big((1-\frac{p}{2})V_{m\tau} + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i\big)(1+\frac{p}{4}) \\ \leq& (1-\frac{p}{4})V_{m\tau} + \frac{12\tau\rho\eta^2}{p}\sum_{i=m\tau}^{t-1} h_i \\ \leq& \cdots \leq \frac{12\tau\rho\eta^2}{p}\eta^2\sum_{j=0}^{t-1}(1-\frac{p}{4})^{\lfloor\frac{t-j}{\tau}\rfloor}h_j.\end{aligned}$$

■

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–7, 2020.

[5] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," *Journal of Healthcare Informatics Research*, vol. 5, no. 1, pp. 1–19, 2021.

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

[7] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," *arXiv preprint arXiv:1907.02189*, 2019.

[8] 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.

[9] 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.

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

[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, "The role of local steps in local sgd," *arXiv preprint arXiv:2203.06798*, 2022.

[13] W. Liu, L. Chen, and W. Zhang, "Decentralized federated learning: Balancing communication and computing costs," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 8, pp. 131–143, 2022.

[14] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.

[16] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[17] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

[18] 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.

[19] X. Li, W. Yang, S. Wang, and Z. Zhang, "Communication-efficient local decentralized sgd methods," *arXiv preprint arXiv:1910.09126*, 2019.

[20] 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.

[21] 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.

[22] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou, "Empirical analysis of the hessian of over-parametrized neural networks," *arXiv preprint arXiv:1706.04454*, 2017.

[23] M. Belkin, S. Ma, and S. Mandal, "To understand deep learning we need to understand kernel learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 541–549.

[24] X. Zhou, "On the fenchel duality between strong convexity and lipschitz continuous gradient," *arXiv preprint arXiv:1803.06573*, 2018.