2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Communication-Efficient Decentralized Local SGD over Undirected Networks

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

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683272

***Abstract*— We consider the distributed learning problem where a network of n agents seeks to minimize a global function F. Agents have access to F through noisy gradients, and they can locally communicate with their neighbors over an undirected network. We study the Decentralized Local SGD method, where agents perform a number of local gradient steps and occasionally exchange information with their neighbors. Previous algorithmic analysis efforts have focused on the specific network topology (star topology), where a leader node aggregates all agents' information. We generalize that setting to an arbitrary undirected network by analyzing the trade-off between the number of communication rounds and the computational effort of each agent. We bound the expected optimality gap in terms of the number of iterates T, the number of workers n, and the spectral gap of the underlying network. Our main results show that by using only $R = \Omega(n)$ communication rounds, one can achieve an error that scales as $O(1/nT)$, where the number of communication rounds is independent of T and only depends on the number of agents.**

I. Introduction

Stochastic Gradient Descent (SGD) is arguably the most commonly used algorithm for the optimization of parameters of machine learning models. SGD tries to minimize a function F by iteratively updating parameters as: $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \hat{\mathbf{g}}^t$, where $\hat{\mathbf{g}}^t$ is a stochastic gradient of F at $\mathbf{x}^t$ and η_t is the learning rate. However, given the massive scale of many modern ML models and datasets, and taking into account data ownership, privacy, fault tolerance, and scalability, decentralized training approaches have recently emerged as a suitable alternative over centralized ones, e.g., parameter server [2], federated learning [3]–[5], decentralized stochastic gradient descent [6]–[9], decentralized momentum SGD [10], decentralized ADAM [11], among others [12], [13].

A naive parallelization of the SGD consists of having multiple workers computing stochastic gradients in parallel, with a central node or fusion center, where local gradients are aggregated and sent back to the workers. However, such a structure induces a large communication overhead, where at each iteration of the algorithm, all workers need to send their gradients to the central node, and then the central node needs to send the workers the aggregated information. Local SGD (Parallel SGD or Federated SGD) [14], [15] presents a suitable solution to reduce such communication overheads. Specifically, each machine independently runs SGD locally and then aggregates by a central node from time to time only. We refer interested readers to [16]–[18] for a number of recent unifying approaches for distributed SGD.

The main advantage of the parallelization approach is that it allows distributed stochastic gradient computations. Such an advantage has been recently shown to translate into linear speedups with respect to the nodes available [8], [19]. However, linear speedups come at the cost of an increasing number of communication rounds. Recently, in [20], [21], the authors showed that the number of communication rounds required is $\Omega(n \text{ polylog}(T))$ for strongly convex functions. Later in [19], the authors showed that such communication complexity could be improved to $\Omega(n)$, still maintaining the linear speedup in the number of nodes n.

This paper focuses on smooth and strongly convex functions with a general noise model, where agents have access to stochastic gradients of a function F, but no central node or fusion center exists. Instead, agents can communicate or exchange information with each other via a network. The network imposes communication constraints because a worker or node can only exchange information with those directly connected. We prove that linear speedup with only $\Omega(n)$ communication rounds is still achievable in this setting.

Our work builds upon two main recent results on the analysis of Local SGD and Decentralized SGD, namely [17], and [19]. In [17], the authors introduced a unifying theory for decentralized SGD and Local updates. In particular, they study the problem of decentralized updates where agents are connected over an arbitrary time-varying network. On the other hand, in [19], the authors propose a new analysis of Local SGD, where a fixed leader/follower network topology is used. They introduce a new analysis that shows that linear speedups with respect to the number of workers n can be achieved, with a communication complexity proportional to $\Omega(n)$ only. A similar problem was also studied in [22], where Local Decentralized SGD was proposed, which allows for multiple local steps and multiple Decentralized SGD steps. We extend the result of [19] to the setting of networked agents. In particular, we study the cases in which the length of the local SGD update is fixed or increasing, and we show that a linear speedup in the number of agents can be obtained for a fixed number of communication rounds in terms of n.

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

Notation: For a positive integer number $n \in \mathbb{Z}_+$, we let $[n] \triangleq \{1, \ldots, n\}$. For an integer $z \in \mathbb{Z}$, we denote the largest integer less than or equal to z by $\lfloor z \rfloor$. We let $\mathbf{1}_n$ be an n-

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

This work is supported by the NSF CAREER Award under Grant No. EPCN-1944403. Online version available at [1].

dimensional column vector with all entries equal to one. A nonnegative matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}_+^{n\times n}$ is called doubly stochastic if it is symmetric and the sum of the entries in each row equals 1. We denote the eigenvalues of a doubly stochastic matrix $\mathbf{W}$ by $1 = |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \geq 0$ and its spectral gap by $1-\rho$, where $\rho \triangleq |\lambda_2|$ denotes the size of the second largest eigenvalue of $\mathbf{W}$.

II. Problem Formulation

Let us consider a decentralized system with a set of $[n] = \{1, \ldots, n\}$ workers. We assume that the workers are connected via a weighted connected undirected graph $\mathcal{G} = ([n], \mathbf{W})$, where there is an edge between workers i and j if and only if they can directly communicate with each other. Here, $\mathbf{W} = [w_{ij}]$ is a symmetric doubly stochastic matrix whose ij-th entry $w_{ij} > 0$ denotes the weight on the edge $\{i, j\}$, and $w_{ij} = 0$ if there is no edge between i and j. We note that since $\mathbf{W}$ is the weighted adjacency matrix associated with the *connected* network $\mathcal{G}$, we have $\rho \in [0, 1)$, where $1-\rho$ is the spectral gap. The specific influence of the graph topology on ρ can be found in the literature [23]. For example, for path graphs one usually has $\rho = O(1/n^2)$, whereas for well connected graphs, like Erdős-Renyí random graph, we have $\rho = O(1/\log^2(n))$.

The workers' objective is to minimize a global function $F : \mathbb{R}^d \to \mathbb{R}$ by performing local gradient steps and occasionally communicating over the network and leveraging the samples obtained by the other agents. Similarly, as in [19], we assume that the function F is a smooth function and that all workers have access to F through noisy gradients. More precisely, we consider the following assumptions.

Assumption 1: Function $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable, μ-strongly convex and L-smooth with condition number $\kappa \triangleq L/\mu \geq 1$, that is, for every $x, y \in \mathbb{R}^d$ we have,

$$\frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2 \leq F(\mathbf{y}) - F(\mathbf{x}) - \langle \nabla F(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2.$$

Assumption 2: Each worker i has access to a gradient oracle which returns an unbiased estimate of the true gradient in the form $\mathbf{g}_i(\mathbf{x}) = \nabla F(\mathbf{x}) + \epsilon_i$, such that ϵ_i is a zero-mean conditionally independent random noise with its expected squared norm error bounded as

$$\mathbb{E}[\epsilon_i] = \mathbf{0}, \qquad \mathbb{E}[\|\epsilon_i\|^2 | \mathbf{x}] \leq c\|\nabla F(\mathbf{x})\|^2 + \sigma^2,$$

where $\sigma^2, c \geq 0$ are constants.

Let us denote the length of time horizon by $T \in \mathbb{Z}_+$. In Decentralized Local SGD, each worker $i \in [n]$ holds a local parameter $\mathbf{x}_i^t$ at iteration t and a set $\mathscr{I} \subset [T]$ of communication times. By writing all the variables and the gradient values at time t in a matrix form, we have

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

where $\mathbf{g}_i^t \triangleq \mathbf{g}(\mathbf{x}_i^t)$ is the stochastic gradient of node i at iteration t. In each time step t, the Decentralized Local SGD algorithm performs the following update:

$$\mathbf{X}^{t+1} = (\mathbf{X}^t - \eta_t \mathbf{G}(\mathbf{X}^t))\mathbf{W}_t,$$

Algorithm 1 Decentralized Local SGD

Input: $\mathbf{x}_i^0 = \mathbf{x}^0$ for $i \in [n]$, total number of iterations T, the step-size sequence $\{\eta_t\}_{t=0}^{T-1}$ and $\mathscr{I} \subseteq [T]$.
for $t = 0, \ldots, T-1$ **do**
  for $j = 1, \ldots, n$ **do**
    evaluate a stochastic gradient $\mathbf{g}_j^t$
    if $t+1 \in \mathscr{I}$ **then**
      $\mathbf{x}_j^{t+1} = \sum_{i=1}^n w_{ij}(\mathbf{x}_i^t - \eta_t \mathbf{g}_i^t)$
    else
      $\mathbf{x}_j^{t+1} = \mathbf{x}_j^t - \eta_t \mathbf{g}_j^t$
    end if
  end for
end for

where $\mathbf{W}_t \in \mathbb{R}^{n\times n}$ is the connected matrix defined by

$$\mathbf{W}_t = \begin{cases} \mathbf{I_n} & \text{if } t \notin \mathscr{I}, \\ \mathbf{W} & \text{if } t \in \mathscr{I}. \end{cases} \tag{1}$$

Note that according to (1), the workers simply update their variables based on SGD at each $t \notin \mathscr{I}$, and communicate with the neighbors only at time instances $t \in \mathscr{I}$. The pseudo code for the Decentralized Local SGD is provided in Algorithm 1.

Our main objective will be to analyze the communication complexity and sample complexity of the outputs of Algorithm 1. In the next section, we state our main results.

III. Convergence and Communication Complexity

In this section, we analyze the convergence guarantee of the Decentralized Local SGD in terms of the number of communication rounds. We will show that for a specific choice of inter-communication intervals, one can achieve an approximate solution with an arbitrarily small optimality gap with a convergence rate of $O(\frac{1}{nT})$ using only $|\mathscr{I}| = \Omega(n)$ communications rounds, which is independent of the horizon length T. To that end, let us consider the following notations that will be used throughout the paper.

$$\bar{\mathbf{x}}^t \triangleq \frac{1}{n}\mathbf{X}^t\mathbf{1}_n, \qquad \bar{\epsilon}^t \triangleq \frac{1}{n}\sum_{i=1}^n \epsilon_i^t,$$
$$\bar{\mathbf{g}}^t \triangleq \frac{1}{n}\mathbf{G}(\mathbf{X}^t)\mathbf{1}_n, \quad \overline{\nabla F}(\mathbf{X}) \triangleq \frac{1}{n}\nabla F(\mathbf{X})\mathbf{1}_n,$$

and define $\xi^t \triangleq \mathbb{E}[F(\bar{\mathbf{x}}^t)] - F^*$ to be the optimality gap of the solution at time t. Let $0 < \tau_1 < \ldots < \tau_R \leq T$ be the communication times, and denote the length of $(i+1)$-th inter-communication interval by $H_i \triangleq \tau_{i+1} - \tau_i$, for $i = 0, \ldots, k-1$. Moreover, we let ρ_t be the size of the second largest eigenvalue of the connectivity matrix $\mathbf{W}_t$, i.e.,

$$\rho_t = \begin{cases} 1 & \text{if } t \notin \mathscr{I}, \\ \rho & \text{if } t \in \mathscr{I}. \end{cases}$$

Let us define the following parameters:

$$G_t \triangleq \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^t)\|_2^2], \; V_t \triangleq \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|_2^2].$$

To prove our main result, we first state two technical lemmas whose proofs can be found in Appendix I.

Lemma 1: Let Assumptions 1 and 2 hold. Then,

$$\xi^{t+1} \leq \xi^t(1-\mu\eta_t) - \frac{\eta_t}{2}G_t + \frac{\eta_t^2 L}{2}(1+\frac{c}{n})G_t + \frac{\eta_t L^2}{2}V_t + \frac{\eta_t^2 L\sigma^2}{2n}. \quad (2)$$

Lemma 2: Let Assumptions 1 and 2 hold. Set $\eta_k = 2/(\mu(k+\beta))$ with β to be specified later, then,

$$V_t \leq \frac{9(n-1)}{n}\sum_{k=0}^{t-1}\frac{cG_k+\sigma^2}{\mu^2(t+\beta)^2}\prod_{i=k}^{t-1}\rho_i^2. \quad (3)$$

We can now bound the optimality error in terms of the communication network's spectral gap at different times.

Theorem 1: Let Assumptions 1 and 2 hold. Choose $\beta \geq \max\{9\kappa^2 c\ln(1+T/2\kappa^2)+2\kappa(1+c/n), 2\kappa^2\}$, and set $\eta_k = 2/(\mu(k+\beta))$. Then, the output of Algorithm 1 has the following property:

$$\mathbb{E}[F(\bar{\mathbf{x}}^T)] - F^* \leq \frac{\beta^2(F(\bar{\mathbf{x}}^0)-F^*)}{T^2} + \frac{2L\sigma^2}{n\mu^2 T} + \frac{9L^2\sigma^2}{\mu^3 T^2}\sum_{t=0}^{T-1}\frac{1}{t+\beta}\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2, \quad (4)$$

where $F^* \triangleq \min_{\mathbf{x}} F(\mathbf{x})$ is the optimum objective value.

Proof: If we combine (2) and (3), and substitute $\eta_t = 2/\mu(t+\beta)$, we get

$$\xi^{t+1} \leq \xi^t(1-\mu\eta_t) - \frac{1}{\mu(t+\beta)}G_t + \frac{2L}{\mu^2(t+\beta)^2}(1+\frac{c}{n})G_t + \frac{2L\sigma^2}{n\mu^2(t+\beta)^2} + \frac{9L^2}{\mu^3(t+\beta)^3}\sum_{k=0}^{t-1}(cG_k+\sigma^2)\prod_{i=k}^{t-1}\rho_i^2.$$

Moreover, if we multiply both sides of the above inequality by $(t+\beta)^2$ and use the following valid inequality

$$(1-\mu\eta_t)(t+\beta)^2 = (1-\frac{2}{t+\beta})(t+\beta)^2 < (t+\beta-1)^2,$$

we obtain,

$$\xi^{t+1}(t+\beta)^2 \leq \xi^t(t+\beta-1)^2 + \left(\frac{2L}{\mu^2}(1+\frac{c}{n}) - \frac{t+\beta}{\mu}\right)G_t + \frac{2L\sigma^2}{n\mu^2} + \frac{9L^2}{\mu^3(t+\beta)}\sum_{k=0}^{t-1}(c\mathbb{E}[G^k]+\sigma^2)\prod_{i=k}^{t-1}\rho_i^2.$$

Summing this relation for $t = 0, \ldots, T-1$, we get

$$\xi^T(T+\beta-1)^2 \leq \xi^0(\beta-1)^2 + \frac{2L\sigma^2}{n\mu^2}T + \frac{9L^2\sigma^2}{\mu^3}\sum_{t=0}^{T-1}\frac{1}{t+\beta}\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2 + \sum_{t=0}^{T-1}G_t\left(\sum_{k=t+1}^{T-1}\frac{9L^2c}{\mu^3(k+\beta)}\prod_{i=t}^{k-1}\rho_i^2 + \frac{2L}{\mu^2}(1+\frac{c}{n}) - \frac{t+\beta}{\mu}\right).$$

We can now bound the coefficient of G_t in the above expression as follows:

$$\begin{aligned}
&\sum_{k=t+1}^{T-1}\frac{9L^2c}{\mu^3(k+\beta)}\prod_{i=t}^{k-1}\rho_i^2 + \frac{2L}{\mu^2}(1+\frac{c}{n}) - \frac{t+\beta}{\mu} \\
&\leq \sum_{k=t+1}^{T-1}\frac{9L^2c}{\mu^3(k+\beta)} + \frac{2L}{\mu^2}(1+\frac{c}{n}) - \frac{t+\beta}{\mu} \\
&\leq \frac{9L^2c}{\mu^3}\ln(\frac{T-1+\beta}{\beta}) + \frac{2L}{\mu^2}(1+\frac{c}{n}) - \frac{\beta}{\mu} \\
&= \frac{1}{\mu}\left(9\kappa^2 c\ln(1+\frac{T-1}{\beta}) + 2\kappa(1+\frac{c}{n}) - \beta\right) \leq 0,
\end{aligned}$$

where in the third inequality we have used $\sum_{k=t_1+1}^{t_2} 1/k \leq \int_{t_1}^{t_2} dx/x = \ln(t_2/t_1)$. The last inequality also holds by the assumption of the theorem. As the coefficient of G_t is non-positive, we can simply drop it from the upper bound to get,

$$\xi^T(T+\beta-1)^2 \leq \xi^0(\beta-1)^2 + \frac{2L\sigma^2}{n\mu^2}T + \frac{9L^2\sigma^2}{\mu^3}\sum_{t=0}^{T-1}\frac{1}{t+\beta}\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2.$$

Finally, dividing both sides of the above inequality by the term $(T+\beta-1)^2$ completes the proof. ∎

Next, we specialize Theorem 1 to two specific choices of inter-communication time intervals, namely, fixed-length and increased-sized communication time intervals.

A. Fixed-Length Communication Intervals

A simple way to select the communication times $\mathscr{I}$, is to partition the entire training time T into R subintervals of length at most H, i.e. $\tau_i = iH$ for $i = 1, \ldots, R-1$ and $\tau_R = \min\{RH, T\}$. In that case, we can bound the error term in (4) as follows:

$$\begin{aligned}
\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2 &\leq \sum_{l=0}^{\lfloor\frac{t-1}{H}\rfloor}\sum_{k=lH}^{(l+1)H-1}\prod_{i=k}^{t-1}\rho_i^2 \\
&\leq H\sum_{l=0}^{\lfloor\frac{t-1}{H}\rfloor}\prod_{i=(l+1)H-1}^{t-1}\rho_i^2 \\
&= H\sum_{l=0}^{\lfloor\frac{t-1}{H}\rfloor}\rho^{2(\lfloor\frac{t-1}{H}\rfloor-l)} \\
&= H\sum_{l=0}^{\lfloor\frac{t-1}{H}\rfloor}\rho^{2l} \leq \frac{H}{1-\rho^2},
\end{aligned}$$

where in the first inequality and by some abuse of notation we set $\prod_{i=k}^{t-1}\rho_i^2 = 1$ if $k > t-1$. As a result, we can upper-bound the error term in (4) by

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{1}{t+\beta}\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2 &\leq \frac{H}{1-\rho^2}\sum_{t=0}^{T-1}\frac{1}{t+\beta} \\
&\leq \frac{H}{1-\rho^2}\ln(\frac{T+\beta}{\beta-1}).
\end{aligned}$$

Corollary 1: Suppose that the assumptions of Theorem 1 hold, and workers communicate at least once every H iterations. Then,

$$\mathbb{E}[F(\bar{\mathbf{x}}^T)] - F^* \leq \frac{\beta^2(F(\bar{\mathbf{x}}^0) - F^*)}{T^2} + \frac{2L\sigma^2}{n\mu^2 T} + \frac{9L^2\sigma^2 H}{\mu^3 T^2(1-\rho^2)} \ln(1 + \frac{T}{\beta - 1}).$$

Using Corollary 1, if we choose $H = \mathcal{O}(\frac{T}{n\ln(T)})$, then Algorithm 1 achieves a linear speedup in the number of workers, which is equivalent to a communication complexity of $R = \Omega(n\ln(T))$.

B. Increased-Size Communication Intervals

Here, we consider a more interesting choice of communication times with varying interval lengths. In other words, we allow the length of consecutive inter-communication intervals $H_i \triangleq \tau_{i+1} - \tau_i$ to grow linearly over time. The following Theorem presents a performance guarantee for this choice of communication times.

Theorem 2: Suppose assumptions of Theorem 1 hold. Choose the maximum number of communications $1 \leq R \leq \sqrt{2T}$ and set $a \triangleq \lceil 2T/R^2 \rceil \geq 1$, $H_i = a(i+1)$ and $\tau_i = \min(ai(i+1)/2, T)$ for $i = 1, \ldots, R$. Then, using Algorithm 1, we have

$$\mathbb{E}[F(\bar{\mathbf{x}}^T)] - F^* \leq \frac{\beta^2(F(\bar{\mathbf{x}}^0) - F^*)}{T^2} + \frac{2L\sigma^2}{n\mu^2 T} + \frac{144L^2\sigma^2}{(1-\rho^2)\mu^3 TR}. \tag{5}$$

Proof: Define $\tau_0 = 0$, for any t satisfying $\tau_j \leq t < \tau_{j+1}$. We can write,

$$\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2 \leq \sum_{l=0}^{j}\sum_{k=\tau_l}^{\tau_{l+1}-1}\prod_{i=k}^{t-1}\rho_i^2 \leq H_j \sum_{l=0}^{j}\prod_{i=\tau_{l+1}-1}^{t-1}\rho_i^2 = H_j\sum_{l=0}^{j}\rho^{2(j-l)} \leq \frac{H_j}{1-\rho^2}.$$

Therefore, we have

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{1}{t+\beta}\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2 &\leq \sum_{t=\tau_0}^{\tau_1-1}\frac{1}{t+\beta}\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2 + \sum_{j=1}^{R-1}\sum_{t=\tau_j}^{\tau_{j+1}-1}\frac{1}{t+\beta}(\sum_{k=0}^{t-1}\prod_{i=k}^{t-1}\rho_i^2) \\
&\leq \sum_{t=\tau_0}^{\tau_1-1}\frac{t}{t+\beta} + \frac{1}{1-\rho^2}\sum_{j=1}^{R-1}\sum_{t=\tau_j}^{\tau_{j+1}-1}\frac{H_j}{t+\beta} \\
&\leq H_0 + \frac{1}{1-\rho^2}\sum_{j=1}^{R-1}\frac{H_j^2}{\tau_j+\beta} \\
&= a + \frac{1}{1-\rho^2}\sum_{j=1}^{R-1}\frac{2a^2(j+1)^2}{aj(j+1)+2\beta} \\
&\leq a + \frac{2}{1-\rho^2}\sum_{j=1}^{R-1}\frac{a^2(j+1)^2}{aj(j+1)} \leq \frac{4aR}{1-\rho^2}.
\end{aligned}$$

If we substitute the values of R and a into the above relation, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1}\frac{t-\tau(t)}{t+\beta} &\leq \frac{4aR}{1-\rho^2} \leq \frac{4(\frac{2T}{R^2}+1)R}{1-\rho^2} \\
&= \frac{1}{1-\rho^2}(\frac{8T}{R} + 4R) \leq \frac{16T}{(1-\rho^2)R},
\end{aligned}$$

where the last inequality holds because $R \leq \sqrt{2T}$. The above relation, together with Theorem 1, concludes the proof. ■

According to Theorem 2, if we choose the number of communication rounds to be $R = \Omega(n)$, then Algorithm 1 achieves an error that scales as $O(\frac{1}{nT})$, *i.e.* a linear speedup in the number of workers when $T = \Omega(n^2)$.

IV. Conclusion

In this paper, we considered the problem of computation versus communication trade-off for Decentralized Local SGD over arbitrary undirected connected graphs. We have shown that by choosing inter-communication intervals appropriately, one can achieve a linear speedup in the number of workers while keeping the total number of communications bounded by the number of workers.

In this work, we restricted our attention to undirected networks and homogeneous objective functions. Therefore, generalizing our work to directed networks in which workers have access to heterogeneous objective functions (or heterogeneous data sets) would be an interesting research direction.

Appendix I: Omitted Proofs

A. Proof of Lemma 1

Define $\mathscr{F}^t := \{\mathbf{x}_i^k, \mathbf{g}_i^k | 1 \leq i \leq n, 0 \leq k \leq t-1\} \cup \{\mathbf{x}_i^t | 1 \leq i \leq n\}$ to be the history of all the iterates up to time t. Then, we can bound the optimality error in terms of V_t and G_t as follows:

$$\begin{aligned}
\bar{\mathbf{x}}^{t+1} &= \frac{1}{n}\mathbf{X}^{t+1}\mathbf{1}_n = \frac{1}{n}((\mathbf{X}^t - \eta_t\mathbf{G}(\mathbf{X}^t))\mathbf{W}_t)\mathbf{1}_n \\
&= \frac{1}{n}\mathbf{X}^t\mathbf{W}_t\mathbf{1}_n - \eta_t\frac{1}{n}\mathbf{G}(\mathbf{X}^t))\mathbf{W}_t\mathbf{1}_n \\
&= \bar{\mathbf{x}}^t - \eta_t\bar{\mathbf{g}}^t.
\end{aligned}$$

By Assumption 1, we have

$$\mathbb{E}[F(\bar{\mathbf{x}}^{t+1}) - F(\bar{\mathbf{x}}^t)] \leq -\eta_t\mathbb{E}[\langle\nabla F(\bar{\mathbf{x}}^t), \bar{\mathbf{g}}^t\rangle] + \frac{\eta_t^2 L}{2}\mathbb{E}[\|\bar{\mathbf{g}}^t\|_2^2]. \tag{6}$$

We bound the first term on the right side of (6) by conditioning on $\mathscr{F}^t$ as follows:

$$\begin{aligned}
\mathbb{E}[\langle\nabla F(\bar{\mathbf{x}}^t), \bar{\mathbf{g}}^t\rangle|\mathcal{F}^t] &= \frac{1}{n}\sum_{i=1}^{n}\langle\nabla F(\bar{\mathbf{x}}^t), \mathbb{E}[\mathbf{g}_i^t|\mathbf{x}_i^t]\rangle \\
&= \frac{1}{2}\|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{1}{2n}\sum_{i=1}^{n}\|\nabla F(\mathbf{x}_i^t)\|^2 \\
&\quad - \frac{1}{2n}\sum_{i=1}^{n}\|\nabla F(\bar{\mathbf{x}}^t) - \nabla F(\mathbf{x}_i^t)\|^2
\end{aligned}$$

$$\geq \mu(F(\bar{\mathbf{x}}^t) - F^*) + \frac{1}{2n}\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^t)\|^2 - \frac{L^2}{2n}\sum_{i=1}^n \|\bar{\mathbf{x}}^t - \mathbf{x}_i^t\|^2. \tag{7}$$

Taking expectation from (7), we obtain

$$\mathbb{E}[\langle \nabla F(\bar{\mathbf{x}}^t), \bar{\mathbf{g}}^t\rangle] \geq \mu\xi^t + \frac{1}{2}G_t - \frac{L^2}{2}V_t.$$

Next, we bound the second term on the right side of (6) by conditioning on $\mathscr{F}^t$ as follows:

$$\begin{aligned}\mathbb{E}[\|\bar{\mathbf{g}}^t\|^2|\mathscr{F}^t] &= \mathbb{E}[\|\overline{\nabla F}(\mathbf{X}^t) + \bar{\epsilon}^t\|^2|\mathscr{F}_t] \\ &= \|\overline{\nabla F}(\mathbf{X}^t)\|^2 + \mathbb{E}[\|\bar{\epsilon}^t\|^2|\mathscr{F}^t] \\ &\leq \frac{1}{n}\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^t)\|^2 + \frac{1}{n^2}\sum_{i=1}^n (\sigma^2 + c\|F(\mathbf{x}_i^t)\|^2).\end{aligned}$$

Taking expectation from the above expression, we have

$$\mathbb{E}[\|\bar{\mathbf{g}}^t\|^2] \leq (1 + \frac{c}{n})G_t + \frac{\sigma^2}{n}. \tag{8}$$

Substituting (7), (8) into (6) completes the proof.

B. Auxiliary Lemmas and Proof of Lemma 2

Lemma 3: Let ρ be the second largest eigenvalue of the doubly stochastic matrix $\mathbf{W}$. Then, for any matrix $\mathbf{Y} \in \mathbb{R}^{d\times n}$, $\mathbf{Y}\mathbf{1}_n = 0$, we have $\|\mathbf{Y}\mathbf{W}\|_F^2 \leq \rho^2\|\mathbf{Y}\|_F^2$.

Lemma 4: Let Assumptions 1 and 2 hold. Then,

$$V_{t+1} \leq \rho_t^2\Big(V_t\left(1 - 2\eta_t\mu + \eta_t^2L^2\right) + \frac{n-1}{n}\eta_t^2\sigma^2 + \frac{n-1}{n}\eta_t^2 cG_t\Big).$$

Proof: Let us define $Q = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. Then, we have

$$\begin{aligned}nV_{t+1} &= \mathbb{E}[\|\mathbf{X}^{t+1}(\mathbf{I} - Q)\|_F^2] \\ &= \mathbb{E}[\|(\mathbf{X}^t - \eta_t\mathbf{G}(\mathbf{X}^t))\mathbf{W}_t(\mathbf{I} - Q)\|_F^2] \\ &= \mathbb{E}[\|(\mathbf{X}^t - \eta_t\mathbf{G}(\mathbf{X}^t))(\mathbf{I} - Q)\mathbf{W}_t\|_F^2] \\ &\leq \rho_t^2\mathbb{E}[\|(\mathbf{X}^t - \eta_t\mathbf{G}(\mathbf{X}^t))(\mathbf{I} - Q)\|_F^2],\end{aligned} \tag{9}$$

where the last inequality holds by Lemma 3. Define

$$\mathbf{x}_i^{t+1/2} := \mathbf{x}_i^t - \eta_t\mathbf{g}_i^t, \qquad \bar{\mathbf{x}}^{t+1/2} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{t+1/2}.$$

Then, we can write

$$\begin{aligned}&\mathbb{E}[\|(\mathbf{X}^t - \eta_t\mathbf{G}(\mathbf{X})\|_F^2] = \mathbb{E}[\sum_{i=1}^n \|\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}\|^2] \\ &= \sum_{i=1}^n \|\mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}]\|^2 \\ &+ \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2} - \mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}]\|^2\right].\end{aligned} \tag{10}$$

Let us consider the first term on the right side of (10). By taking conditional expectation, we obtain

$$\begin{aligned}&\sum_{i=1}^n \|\mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}|\ \mathscr{F}^t]\|^2 \\ &= \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t - \eta_t(\nabla F(\mathbf{x}_i^t) - \overline{\nabla F}(\mathbf{X}^t))\|^2 \\ &= \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 + \sum_{i=1}^n \eta_t^2\|\nabla F(\mathbf{x}_i^t) - \overline{\nabla F}(\mathbf{X}^t)\|^2 \\ &- 2\eta_t\sum_{i=1}^n \langle \nabla F(\mathbf{x}_i^t), \mathbf{x}_i^t - \bar{\mathbf{x}}^t\rangle.\end{aligned} \tag{11}$$

By using L-smoothness of F, we have

$$\begin{aligned}\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^t) - \overline{\nabla F}(\mathbf{X}^t)\|^2 &= \frac{1}{n}\sum_{\{i,j\}} \|F(\mathbf{x}_i^t) - F(\mathbf{x}_j^t)\|^2 \\ &\leq \frac{L^2}{n}\sum_{\{i,j\}} \|\mathbf{x}_i^t - \mathbf{x}_j^t\|^2 = L^2\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2.\end{aligned} \tag{12}$$

Moreover, by μ-strong convexity of F, we have

$$\begin{aligned}\sum_{i=1}^n \langle F(\mathbf{x}_i^t), \mathbf{x}_i^t - \bar{\mathbf{x}}^t\rangle &= \sum_{i=1}^n \langle F(\mathbf{x}_i^t), \frac{1}{n}\sum_{j=1}^n (\mathbf{x}_i^t - \mathbf{x}_j^t)\rangle \\ &= \frac{1}{n}\sum_{\{i,j\}} \langle F(\mathbf{x}_i^t) - F(\mathbf{x}_j^t), \mathbf{x}_i^t - \mathbf{x}_j^t\rangle \\ &\geq \frac{\mu}{n}\sum_{\{i,j\}} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = \mu\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2,\end{aligned} \tag{13}$$

where the last inequality follows from $\langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \mu\|\mathbf{x} - \mathbf{y}\|^2$. Finally, by combining (11)-(13) we obtain,

$$\begin{aligned}&\sum_{i=1}^n \|\mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}|\mathscr{F}^t]\|^2 \\ &\qquad \leq \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2\left(1 - 2\eta_t\mu + \eta_t^2L^2\right).\end{aligned}$$

Next, let us consider the second term on the right side of (10). We have,

$$\begin{aligned}&\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2} - \mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}]\right\|^2|\mathscr{F}^t\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{x}_i^{t+1/2} - \mathbb{E}[\mathbf{x}_i^{t+1/2}] - (\bar{\mathbf{x}}^{t+1/2} - \mathbb{E}[\bar{\mathbf{x}}^{t+1/2}])\right\|^2|\mathscr{F}^t\right] \\ &= \eta_t^2\sum_{i=1}^n \mathbb{E}\left[\left\|\epsilon_i^t - \bar{\epsilon}^t\right\|^2|\mathscr{F}^t\right] \\ &= \eta_t^2\left(\sum_{i=1}^n \mathbb{E}\left[\left\|\epsilon_i^t\right\|^2|\mathscr{F}^t\right] - n\mathbb{E}\left[\left\|\bar{\epsilon}^t\right\|^2|\mathscr{F}^t\right]\right) \\ &= \eta_t^2\sum_{i=1}^n \mathbb{E}\left[\left\|\epsilon_i^t\right\|^2|\mathscr{F}^t\right](1 - \frac{1}{n}) \\ &\leq (n-1)\eta_t^2\sigma^2 + (1 - \frac{1}{n})\eta_t^2 c\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^t)\|^2,\end{aligned}$$

where in the last equality we have used conditional independence of ϵ_i^t to conclude $\mathbb{E}[\|\bar{\epsilon}^t\|^2|\mathscr{F}^t] = (1/n^2)\sum_{i=1}^n \mathbb{E}[\|\epsilon_i^t\|^2|\mathscr{F}^t]$. If we take expectation from

the two relations above and combine them with (9) and (10), we get

$$
\begin{aligned}
&nV_{t+1} \le \rho_t^2 \mathbb{E}[\|(\mathbf{X}^t - \eta_t \mathbf{G}(\mathbf{X}^t))(\mathbf{I}-Q)\|_F^2] \\
&= \rho_t^2 \sum_{i=1}^n \|\mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}]\|^2 \\
&\quad + \rho_t^2 \sum_{i=1}^n \mathbb{E}\left[\|\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2} - \mathbb{E}[\mathbf{x}_i^{t+1/2} - \bar{\mathbf{x}}^{t+1/2}]\|^2\right] \\
&\le \rho_t^2 \mathbb{E}[\sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2] \left(1 - 2\eta_t\mu + \eta_t^2 L^2\right) \\
&\quad + \rho_t^2((n-1)\eta_t^2\sigma^2 + (1-\frac{1}{n})\eta_t^2 c\mathbb{E}[\sum_{i=1}^n \|\nabla F(\mathbf{x}_i^t)\|^2]) \\
&= \rho_t^2(nV_t\left(1-2\eta_t\mu+\eta_t^2L^2\right)+(n-1)\eta_t^2\sigma^2+(n-1)\eta_t^2 cG_t).
\end{aligned}
$$

That completes the proof. ∎

Now we are ready to prove Lemma 2.

Proof of Lemma 2: Define $\Delta_k = (1 - 2\eta_k\mu + \eta_k^2 L^2)$ for $k \ge 0$. Using Lemma 4, recursively, we can write

$$
\begin{aligned}
&V_t \le \rho_{t-1}^2\Big(\Delta_{t-1}V_{t-1} + \frac{\eta_{t-1}^2(n-1)}{n}(\sigma^2 + cG_{t-1})\Big) \\
&\le \rho_{t-1}^2\Delta_{t-1}\rho_{t-2}^2\Delta_{t-2}V_{t-2} \\
&+ \rho_{t-1}^2\rho_{t-2}^2\frac{\eta_{t-2}^2(n-1)}{n}(\sigma^2 + cG_{t-2}) \\
&+ \rho_{t-1}^2\frac{\eta_{t-1}^2(n-1)}{n}(\sigma^2 + cG_{t-1}) \le \dots \\
&\le \prod_{k=0}^{t-1}\rho_k^2\Delta_k V_0 + \frac{n-1}{n}\sum_{k=0}^{t-1}\eta_k^2(\sigma^2 + c\mathbb{E}[G^k])\prod_{i=k+1}^{t-1}\Delta_i\prod_{i=k}^{t-1}\rho_i^2 \\
&= \frac{n-1}{n}\sum_{k=0}^{t-1}\eta_k^2(\sigma^2 + c\mathbb{E}[G^k])\prod_{i=k+1}^{t-1}\Delta_i\prod_{i=k}^{t-1}\rho_i^2,
\end{aligned}
$$

where in the last equality we have used $V_0 = 0$. By the choice of stepsize and $\beta \ge 2\kappa^2$, we have

$$
\begin{aligned}
\Delta_k &= 1 - \frac{4}{(k+\beta)} + \frac{4L^2}{\mu^2(k+\beta)^2} \\
&\le 1 - \frac{4}{k+\beta} + \frac{4\kappa^2}{(k+\beta)\beta} \\
&\le 1 - \frac{4}{k+\beta} + \frac{2}{(k+\beta)} = 1 - \frac{2}{k+\beta}.
\end{aligned}
$$

Therefore, we have,

$$
\begin{aligned}
V_t &\le \frac{n-1}{n}\sum_{k=0}^{t-1}\frac{4(\sigma^2 + c\mathbb{E}[G^k])}{\mu^2(k+\beta)^2}\frac{(k+\beta+1)^2}{(t+\beta)^2}\prod_{i=k}^{t-1}\rho_i^2 \\
&\le \frac{n-1}{n}\sum_{k=0}^{t-1}\frac{9(\sigma^2 + c\mathbb{E}[G^k])}{\mu^2(t+\beta)^2}\prod_{i=k}^{t-1}\rho_i^2,
\end{aligned}
$$

where in the first inequality we have used the valid inequality $\prod_{i=a}^{b}\left(1-\frac{2}{i}\right) \le \left(\frac{a}{b+1}\right)^2$, and in the second inequality we have used $(k+\beta+1)/(k+\beta) \le (\beta+1)/\beta \le 3/2$ since $\beta \ge 2\kappa^2 \ge 2$.

REFERENCES

[1] T. Qin, S. R. Etesami, and C. A. Uribe, "Communication-efficient decentralized local sgd over undirected networks," *arXiv preprint arXiv:2011.03255*, 2020.

[2] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang *et al.*, "Large scale distributed deep networks," in *Advances in neural information processing systems*, 2012, pp. 1223–1231.

[3] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[4] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging. corr abs/1602.05629 (2016)," *arXiv preprint arXiv:1602.05629*, 2016.

[5] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[6] 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," in *Advances in Neural Information Processing Systems*, 2017, pp. 5330–5340.

[7] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, "Decentralized training over decentralized data," *arXiv preprint arXiv:1803.07068*, 2018.

[8] A. Koloskova, S. U. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," *arXiv preprint arXiv:1902.00340*, 2019.

[9] M. Assran and M. Rabbat, "Asynchronous subgradient-push," *arXiv preprint arXiv:1803.08950*, 2018.

[10] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization," *arXiv preprint arXiv:1905.03817*, 2019.

[11] P. Nazari, D. A. Tarzanagh, and G. Michailidis, "Dadam: A consensus-based distributed adaptive gradient method for online optimization," *arXiv preprint arXiv:1901.09109*, 2019.

[12] Y. Lu and C. De Sa, "Moniqua: Modulo quantized communication in decentralized sgd," *arXiv preprint arXiv:2002.11787*, 2020.

[13] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6155–6165.

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

[15] M. Zinkevich, M. Weimer, L. Li, and A. J. Smola, "Parallelized stochastic gradient descent," in *Advances in neural information processing systems*, 2010, pp. 2595–2603.

[16] 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.

[17] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich, "A unified theory of decentralized sgd with changing topology and local updates," *arXiv preprint arXiv:2003.10422*, 2020.

[18] Y. Lu, J. Nash, and C. De Sa, "Mixml: A unified analysis of weakly consistent parallel learning," *arXiv preprint arXiv:2005.06706*, 2020.

[19] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Local sgd with a communication overhead depending only on the number of workers," *arXiv preprint arXiv:2006.02582*, 2020.

[20] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local sgd on identical and heterogeneous data," *arXiv*, pp. arXiv–1909, 2019.

[21] S. U. Stich and S. P. Karimireddy, "The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication," *arXiv preprint arXiv:1909.05350*, 2019.

[22] X. Li, W. Yang, S. Wang, and Z. Zhang, "Communication efficient decentralized training with multiple local updates," *arXiv preprint arXiv:1910.09126*, 2019.

[23] A. Nedić, A. Olshevsky, and C. A. Uribe, "Graph-theoretic analysis of belief system dynamics under logic constraints," *Scientific reports*, vol. 9, no. 1, pp. 1–16, 2019.