Demystifying Why Local Aggregation Helps: Convergence Analysis of Hierarchical SGD

Jiayi Wang[1], **Shiqiang Wang**[2], **Rong-Rong Chen**[1], **Mingyue Ji**[1]

[1] Department of Electrical & Computer Engineering, University of Utah, Salt Lake City, UT, USA
[2] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
jiayi.wang@utah.edu, wangshiq@us.ibm.com, rchen@ece.utah.edu, mingyue.ji@utah.edu

Abstract

Hierarchical SGD (H-SGD) has emerged as a new distributed SGD algorithm for multi-level communication networks. In H-SGD, before each global aggregation, workers send their updated local models to local servers for aggregations. Despite recent research efforts, the effect of local aggregation on global convergence still lacks theoretical understanding. In this work, we first introduce a new notion of "upward" and "downward" divergences. We then use it to conduct a novel analysis to obtain a worst-case convergence upper bound for two-level H-SGD with non-IID data, non-convex objective function, and stochastic gradient. By extending this result to the case with random grouping, we observe that this convergence upper bound of H-SGD is between the upper bounds of two single-level local SGD settings, with the number of local iterations equal to the local and global update periods in H-SGD, respectively. We refer to this as the "sandwich behavior". Furthermore, we extend our analytical approach based on "upward" and "downward" divergences to study the convergence for the general case of H-SGD with more than two levels, where the "sandwich behavior" still holds. Our theoretical results provide key insights of why local aggregation can be beneficial in improving the convergence of H-SGD.

Introduction

Stochastic gradient descent (SGD) is a widely used optimization technique in machine learning applications. In distributed SGD, all n workers collaboratively learn a global model $\mathbf{w}$ by minimizing the empirical loss with their local data:

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) := \frac{1}{n}\sum_{j=1}^{n} F_j(\mathbf{w}), \tag{1}$$

where $F_j(\cdot)$ is the local loss function of worker j. Traditionally, workers send their models or gradients to the central server after one local iteration, which is inefficient due to frequent model aggregations. This may cost high communication latency in practice. Recently, a form of distributed SGD was proposed to reduce communication cost by allowing multiple local iterations during one communication round (McMahan et al. 2017). This is referred to as local SGD (Stich 2019; Lin et al. 2020). However, since data on workers can be non-IID, to avoid model divergence, the number of local iterations P cannot be too large, which limits its advantage of reducing communication cost.

In practical scenarios, networks often have a hierarchical structure in nature, such as edge computing systems (Li, Ota, and Dong 2018) and software defined networks (SDN) (Kim and Feamster 2013). In these networks, workers often directly communicate with their local server rather than global server. Motivated by this hierarchical structure, a few works (Castiglia, Das, and Patterson 2021; Liu et al. 2020) proposed hierarchical SGD (H-SGD). In H-SGD, workers are partitioned into N groups, where each group has a local server. To reduce communication cost, workers send their models to local servers to do several local aggregations before communicating with global server. They perform I local iterations (referred to as local period) between local aggregations, and G (referred to as global period) local iterations ($G > I$) between global aggregations. In this paper, we call the connection between local servers and workers as "downward" network and the connection between local servers and the global server as "upward" network.

Recently, there have been a few works analyzing the convergence behavior of H-SGD. A structure where the upward network is a peer-to-peer network while downward network is a server-worker network was considered by Castiglia, Das, and Patterson (2021). However, it only considers IID data. The work by Liu et al. (2020) considers non-IID data but it uses full-batch (non-stochastic) gradient descent and the convergence bound is an exponential function of G. In both works, a comprehensive comparison between H-SGD and local SGD is missing and none of them analyzes the effect of local aggregation on overcoming data heterogeneity, which is the key merit of H-SGD. Therefore, a more general and tighter analysis for H-SGD with non-convex objective function, non-IID data and stochastic gradient descent is needed. The effect of local aggregation on overcoming data heterogeneity lacks theoretical understanding.

In this paper, to provide a better theoretical understanding to H-SGD, we devise a novel characterization of the data heterogeneity of H-SGD by "upward" divergence and "downward" divergence, respectively. Furthermore, we show that the global divergence can be partitioned into upward and downward divergences. H-SGD can be seen as performing distributed SGD both within a group and across groups.

Table 1: A summary of convergence bounds in the literature.

Paper	Convergence Bound	Non-IID	SGD	Type	Assumption
Yu, Jin, and Yang (2019)	$O\left(\frac{1+\sigma^2}{\sqrt{nT}} + \frac{n}{T}(P\sigma^2 + P^2\tilde{\epsilon}^2)\right)$	✔	✔	Local SGD	$N = 1$
Liu et al. (2020)	$O\left(\frac{1+B^G\tilde{\epsilon}^2}{\sqrt{nT}}\right)$	✔	✗	H-SGD	$\sigma^2 = 0$
Castiglia, Das, and Patterson (2021)	$O\left(\frac{1+\sigma^2}{\sqrt{nT}} + \frac{n}{T}\frac{G^2}{I}\sigma^2\right)$	✗	✔	H-SGD	$\tilde{\epsilon}^2 = 0$
Ours	$O\left(\frac{1+\sigma^2}{\sqrt{nT}} + \frac{(N-1)(G\sigma^2+G^2\tilde{\epsilon}^2)+(n-N)(I\sigma^2+I^2\tilde{\epsilon}^2)}{T}\right)$	✔	✔	H-SGD	None

[1] P: aggregation period in local SGD; G: global aggregation period in H-SGD; I: local aggregation period in H-SGD; N: number of groups in H-SGD; $\tilde{\epsilon}^2$: global divergence (Assumption 2); σ^2: stochastic gradient noise
[2] B is a constant and $B > 2$. n is the number of nodes and N is the number of groups. T is the total number of local iterations.
[3] Our bound can reduce to the IID case by setting $\tilde{\epsilon}^2 = 0$.

Within a group, workers perform multiple local iterations (on each worker) before local aggregation (within each group). Across groups, each group performs multiple local aggregations before global aggregation. With this characterization, we conduct a novel convergence analysis for H-SGD with non-IID data, non-convex objective function and stochastic gradients. Furthermore, we show that although data can be highly non-IID, local aggregation can help global convergence even when grouping is random. Then by a detailed comparison with local SGD, we show that our convergence upper bound for H-SGD lies in between the convergence upper bounds of two single-level local SGD settings with aggregation periods of I and G, respectively. This is referred to as the "sandwich" behavior, which reveals the fundamental impact of local aggregation.

Our convergence analysis shows that better convergence of H-SGD can be achieved with local aggregation when the number of groups, together with the global and local periods, are chosen appropriately to control the combined effect of upward and downward divergences. In general, we show that since local aggregation is more frequent, grouping strategies with a smaller upward divergence can strengthen the benefit of local aggregation. To reduce the communication cost while maintaining similar or better convergence, it can be beneficial to increase the global period G and decrease the local period I. We also extend our results to multi-level cases where there are multiple levels of local servers, where our characterization with upward and downward divergences can be applied to each level, from which we derive the convergence bound for general multi-level cases.

A comparison of our result with existing results is shown in Table 1. When setting $N = 1$ and $P = I = G$, our result recovers the well-known result for single-level local SGD by Yu, Jin, and Yang (2019). We can also see that choosing $I < G = P$ for H-SGD gives a smaller convergence upper bound, which shows the benefit of the hierarchical structure. The result by Liu et al. (2020) only considers full gradient descent. In this case, the stochastic noise $\sigma^2 = 0$. Even when setting $\sigma^2 = 0$, our result is tighter. The result by Castiglia, Das, and Patterson (2021) only considers IID case where the global divergence $\tilde{\epsilon}^2 = 0$. Setting $\tilde{\epsilon}^2 = 0$, our result is still tighter than theirs since $I < G$. It can be seen that our result is the most general and tightest.

Main Contributions.

- We introduce the new notion of "upward" and "downward" divergences to characterize data heterogeneity of H-SGD. We show that it can be extended to multi-level cases.
- We derive a general convergence bound for two-level H-SGD with non-IID data, non-convex objective functions and stochastic gradient descent.
- We provide a novel convergence analysis for random grouping and show how local aggregation helps global convergence by a "sandwich" behavior.
- We extend our analysis to multi-level H-SGD and the result shows similar properties as the two-level case. To our knowledge, this is the first analysis which can be extended to multi-level cases.

We also conduct experiments on CIFAR-10, FEMNIST, and CelebA datasets. The results of experiments validate our theoretical results.

Related Works

Traditional distributed SGD is introduced by Zinkevich et al. (2010), which can be seen as a special case of local SGD with only one local iteration during one communication round. There have been a large volume of works analyzing the convergence of local SGD, for convex objective functions (Li et al. 2020; Wang et al. 2019), non-convex objective functions (Haddadpour, Farzin et al. 2019; Yu, Yang, and Zhu 2019), and their variants (Karimireddy et al. 2020; Li et al. 2019; Reddi et al. 2020; Wang and Joshi 2019; Yu, Jin, and Yang 2019). Local SGD can be regarded as a special case of H-SGD with only one level. Our theoretical results for H-SGD recover results for both local SGD and traditional distributed SGD. There have been a few works analyzing the convergence of H-SGD, including Castiglia, Das, and Patterson (2021); Zhou and Cong (2019) for IID data, and Liu et al. (2020) for non-IID data with full-batch gradient descent, as described earlier. There are also works on system design for H-SGD without theoretical guarantees (Abad et al. 2020; Luo et al. 2020). In addition, there are works on decentralized SGD (Wang and Joshi 2019; Bellet, Kermarrec, and

Lavoie 2021), where workers exchange their models based on a doubly stochastic mixing matrix. However, the analysis for decentralized SGD cannot be applied to H-SGD, since the second largest eigenvalue of the mixing matrix in the case of H-SGD is one and the decentralized SGD analysis requires this second largest eigenvalue to be strictly less than one.

There are also some works focusing on practical aspects such as model compression and sparsification (Han, Wang, and Leung 2020; Jiang and Agrawal 2018; Jiang et al. 2020; Konecny et al. 2016) and partial worker participation (Bonawitz et al. 2019; Chen et al. 2020). These algorithms and techniques are orthogonal to our work and may be applied together with H-SGD.

H-SGD Setup

In two-level H-SGD, all workers are partitioned into N groups $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_N$. The number of workers in each group is denoted by $n_i := |\mathcal{V}_i|$ $(i = 1, 2, \ldots, N)$. Then we have $n = \sum_{i=1}^{N} n_i$. With this grouping, the objective function (1) is equivalent to

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) := \sum_{i=1}^{N} \frac{n_i}{n} f_i(\mathbf{w}), \tag{2}$$

where $f_i(\cdot)$ is the averaged loss function of workers in group i that is defined as follows:

$$f_i(\mathbf{w}) := \frac{1}{n_i} \sum_{j\in\mathcal{V}_i} F_j(\mathbf{w}). \tag{3}$$

During each local iteration t, each worker j updates its own model using SGD:

$$\mathbf{w}_j^{t+1} = \mathbf{w}_j^t - \gamma \mathbf{g}(\mathbf{w}_j^t, \zeta_j^t), \tag{4}$$

where γ is the learning rate, $\mathbf{g}(\mathbf{w}_j^t, \zeta_j^t)$ is the stochastic gradient of $F_j(\mathbf{w})$, and ζ_j^t represents random data samples from the local dataset $\mathcal{D}_j$ at worker j. We assume that $\mathbb{E}_{\zeta_j^t\sim\mathcal{D}_j}[\mathbf{g}(\mathbf{w}_j^t, \zeta_j^t)] = \nabla F_j(\mathbf{w}_j^t)$.

During one communication round, local models are first averaged within group i $(i = 1, 2, \ldots, N)$ after every I_i local iterations. In particular, at local iteration $t \in \{I_i, 2I_i, 3I_i, \ldots\}$, we compute $\bar{\mathbf{w}}_i^t := \frac{1}{n_i} \sum_{j\in\mathcal{V}_i} \mathbf{w}_j^t$. This can be done at a *local server* (e.g., a computational component in close proximity) for group i. After several rounds of "intra-group" aggregations, models from the N groups are averaged globally. Let a global aggregation be performed for every G local iterations. Then, at local iteration t, we have $\bar{\mathbf{w}}^t := \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{w}}_i^t$ for $t = G, 2G, 3G, \ldots$. Note that we assume that all workers perform synchronous updates and let G be a common multiple of $\{I_1, \ldots, I_N\}$. Therefore, distributed SGD is conducted both at the local level within a group and the global level across groups. We summarize the algorithm in Algorithm 1, where $a \mid b$ (or $a \nmid b$) denotes that a divides (or does not divide) b, i.e., b is (or is not) an integer multiple of a.

In order to understand the fundamental convergence behavior of H-SGD, we will mainly focus on the two-level model introduced above. We will extend our analysis to more than two levels in a later section.

Algorithm 1: Hierarchical SGD (H-SGD)

Input: γ, $\bar{\mathbf{w}}^0$, G, $\{\mathcal{V}_i : i \in \{1, 2, \ldots, N\}\}$, $\{I_i : i \in \{1, 2, \ldots, N\}\}$
Output: Global aggregated model $\bar{\mathbf{w}}^T$
for $t = 0$ *to* $T - 1$ **do**
  for *Each group* $i \in \{1, 2, \ldots, N\}$, *in parallel* **do**
    for *Each worker* $j \in \mathcal{V}_i$, *in parallel* **do**
      Compute $\mathbf{g}(\mathbf{w}_j^t, \zeta_j^t)$;
      $\mathbf{w}_j^{t+1} \leftarrow \mathbf{w}_j^t - \gamma \mathbf{g}(\mathbf{w}_j^t, \zeta_j^t)$;
    if $I_i \mid t + 1$ **then**
      Local aggregate: $\bar{\mathbf{w}}_i^{t+1} \leftarrow \frac{1}{n_i} \sum_{j\in\mathcal{V}_i} \mathbf{w}_j^{t+1}$;
      if $G \nmid t + 1$ **then**
        Distribute: $\mathbf{w}_j^{t+1} \leftarrow \bar{\mathbf{w}}_i^{t+1}, \forall j \in \mathcal{V}_i$;
  if $G \mid t + 1$ **then**
    Global aggregate: $\bar{\mathbf{w}}^{t+1} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \bar{\mathbf{w}}_i^{t+1}$;
    Distribute: $\mathbf{w}_j^{t+1} \leftarrow \bar{\mathbf{w}}^{t+1}, \forall j \in \mathcal{V}$;

Convergence Analysis for Two-level H-SGD

We begin with a description of the assumptions made in our convergence analysis. Then, we present our results for general two-level H-SGD with fixed groupings. The number of workers and the number of local period during one communication round can vary among groups.

Assumptions

We make the following minimal set of assumptions that are common in the literature.

Assumption 1. *For H-SGD, we assume the following.*

a) ***Lipschitz gradient***

$$\|\nabla F_j(\mathbf{w}) - \nabla F_j(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \forall j, \mathbf{w}, \mathbf{w}' \tag{5}$$

where L is some positive constant.

b) ***Bounded variance***

$$\mathbb{E}_{\zeta_j^t\sim\mathcal{D}_j} \|\mathbf{g}(\mathbf{w}; \zeta_j^t) - \nabla F_j(\mathbf{w})\|^2 \leq \sigma^2, \forall j, \mathbf{w} \tag{6}$$

where $\mathcal{D}_j$ is the dataset at worker $j \in \mathcal{V}$.

c) ***Bounded upward divergence (H-SGD)***

$$\frac{n_i}{n} \sum_{i=1}^{N} \|\nabla f_i(\mathbf{w}) - \nabla f(\mathbf{w})\|^2 \leq \epsilon^2, \forall \mathbf{w} \tag{7}$$

d) ***Bounded downward divergence (H-SGD)***

$$\frac{1}{n_i} \sum_{j\in\mathcal{V}_i} \|\nabla F_j(\mathbf{w}) - \nabla f_i(\mathbf{w})\|^2 \leq \epsilon_i^2, \forall i, \mathbf{w} \tag{8}$$

Note that the Lipschitz gradient assumption also applies to the group objective $f_i(\mathbf{w})$ and overall objective $f(\mathbf{w})$. For example, $\|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}')\| = \|\frac{1}{n_i} \sum_{j\in\mathcal{V}_i} \nabla F_j(\mathbf{w}) - \frac{1}{n_i} \sum_{j\in\mathcal{V}_i} \nabla F_j(\mathbf{w}')\| \leq \frac{1}{n_i} \sum_{j\in\mathcal{V}_i} \|\nabla F_j(\mathbf{w}) - \nabla F_j(\mathbf{w}')\| \leq L\|\mathbf{w} - \mathbf{w}'\|$. Global divergence is often used to describe the data heterogeneity of single-level local SGD (Yu, Jin, and Yang 2019), which is as the following.

Assumption 2. ***Bounded global divergence***

$$\frac{1}{n}\sum_{j=1}^{n}\|\nabla F_j(\mathbf{w})-\nabla f(\mathbf{w})\|^2\leq\tilde{\epsilon}^2,\forall\mathbf{w}. \tag{9}$$

Note that when $N=1$, H-SGD reduces to local SGD and downward divergence becomes the global divergence while upward divergence becomes zero. When $N=n$, H-SGD also reduces to local SGD but upward divergence becomes the global divergence while downward divergence becomes zero. Here, we will use global divergence to show our results for H-SGD can encompass the results for local SGD.

Note for the global divergence, it is easy to show that

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{N}\|\nabla F_j(\mathbf{w})-\nabla f(\mathbf{w})\|^2&=\sum_{i=1}^{N}\frac{n_i}{n}\|\nabla f_i(\mathbf{w})-\nabla f(\mathbf{w})\|^2\\&+\sum_{i=1}^{N}\frac{n_i}{n}\frac{1}{n_i}\sum_{j\in\mathcal{V}_i}\|\nabla F_j(\mathbf{w})-\nabla f_i(\mathbf{w})\|^2.\end{aligned} \tag{10}$$

In (10), we see that the upward and downward divergences are in fact a partition of the global divergence, which implies that upward and downward divergences do not increase at the same time. It is the hierarchical structure that makes this partition possible. Later we will show that this partition can exactly explain the benefits of local aggregation. We will discuss more about the relationship between upward/downward and global divergences in a later section.

Convergence Analysis

Technical Challenge. The main challenge of the analysis is that workers only perform local iterations before global aggregation in local SGD, whereas in H-SGD, workers not only perform local iterations but also do aggregations with other workers in the same group. If we directly apply the analysis for local SGD in upward part, the effect of local aggregations would be neglected and the resulted bound would be loose. To address this, our key idea is to construct the local parameter drift $\|\mathbf{w}_j-\bar{\mathbf{w}}_i\|, j\in\mathcal{V}_i$ in the analysis of global parameter drift $\|\bar{\mathbf{w}}_i-\bar{\mathbf{w}}\|$, so that the analysis for the downward part can be incorporated in the analysis for the upward part.

Theorem 1. *Consider the problem in (2). For any fixed worker grouping that satisfies Assumption 1, if the learning rate in Algorithm* 1 *satisfies* $\gamma<\frac{1}{2\sqrt{6}GL}$*, then for any* $T\geq 1$*, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{w}}^t)\|^2\leq\frac{2(f^0-f^*)}{\gamma T}+\gamma L\frac{1}{n}\sigma^2 \tag{11a}$$

$$+2C\gamma^2G\frac{N-1}{n}\sigma^2+3C\gamma^2G^2\epsilon^2 \tag{11b}$$

$$+2C\gamma^2\sigma^2\sum_{i=1}^{N}\frac{(n_i-1)I_i}{n}+3C\gamma^2\sum_{i=1}^{N}\frac{n_i}{n}I_i^2\epsilon_i^2, \tag{11c}$$

where $C=40/3$.

Remark 1. Note that the bound (11a)–(11c) can be partitioned into three parts. The terms in (11a) are the original SGD part. If we set $N=n_i=1$ (ϵ and ϵ_i become zero), then only (11a) will remain, which is the same as convergence bound for non-convex SGD in (Bottou, Curtis, and Nocedal 2018). The terms in (11b) are the upward part, which consists of noise and upward divergence associated with G. The terms in (11c) are the downward part, which has a similar form as (11b) associated with I_i. Divergence plays a more important role than noise since the coefficients in front of the SGD noise σ^2 include G and I_i, while the coefficients in front of the divergences ϵ and ϵ_i include G^2 and I_i^2. Since $G>I_i,\forall i$, the upward part has a stronger influence on convergence.

Remark 2. Note that all the divergences of H-SGD can be written as $O(\gamma^2G^2\epsilon^2+\gamma^2\sum_{i=1}^{N}\frac{n_i}{n}I_i^2\epsilon_i^2)$ while the corresponding part in local SGD is $O(\gamma^2G^2\tilde{\epsilon}^2)$ (Yu, Jin, and Yang 2019). This exactly shows how local aggregation overcomes data heterogeneity: global divergence is partitioned into two parts and local aggregation weakens the effects of the downward part. This also brings us some insights on how to group workers. The grouping strategy should have the smallest upward divergence since this can make full use of benefits of local aggregation. We will validate this property in the experiments.

Remark 3. Let $\gamma=\sqrt{\frac{n}{T}}$ with $T\geq\frac{1}{24G^2L^2\sqrt{n}}$, when T is sufficiently large, $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{w}}^t)\|^2=O\left(\sqrt{\frac{1}{nT}}\right)+O\left(\frac{1}{T}\right)$, which achieves a linear speedup in n.

In the following corollary, we show that our results can encompass local SGD with local period P.

Corollary 1. ***(Degenerate to local SGD)*** *Let* $N=1$ *and* $\gamma\leq\frac{1}{2\sqrt{6}PL}$*, from Theorem 1, we obtain*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{w}}^t)\|^2&\leq\frac{2(f^0-f^*)}{\gamma T}+\frac{\gamma L\sigma^2}{n}\\&+2C\gamma^2L^2\sigma^2\left(1-\frac{1}{n}\right)P+3C\gamma^2L^2P^2\tilde{\epsilon}^2\\&=O\left(\frac{1}{\gamma T}\right)+O\left(\frac{\gamma\sigma^2}{n}\right)\\&+O\left(\gamma^2P\sigma^2\left(1-\frac{1}{n}\right)\right)+O\left(\gamma^2P^2\tilde{\epsilon}^2\right).\end{aligned} \tag{12}$$

While (12) is similar to the bound by Yu, Jin, and Yang (2019), we note that the third term $O\left(\gamma^2P\sigma^2\left(1-\frac{1}{n}\right)\right)$ in the last equality in (12) has an additional term of $\left(1-\frac{1}{n}\right)$ compared to the bound by Yu, Jin, and Yang (2019). This term can potentially make our bound tighter. Another important observation is that the techniques used to obtain this term is the key to make our H-SGD bound in Theorem 1 encompass original SGD cases.

H-SGD with Random Grouping

We now state our convergence results for H-SGD with random grouping. We will show that the convergence upper bound of H-SGD with random grouping takes a value that is between the convergence upper bounds of two single-level

local SGD settings with local and global periods of I and G, respectively. This will provide insights on when and why local aggregation helps.

For worker grouping, we consider all possible grouping strategies with the constraint that $n_i = n/N, \forall i$. Then, we uniformly select one grouping strategy at random. Let the random variable S denote the uniformly random grouping strategy. This means that each realization of S corresponds to one grouping realization. First, we introduce two key lemmas for our convergence analysis.

Lemma 1. *Using the uniformly random grouping strategy* S*, for any* $\mathbf{w}$*, the average upward divergence is*

$$\mathbb{E}_{\mathsf{S}}\left[\frac{1}{N}\sum_{i=1}^{N}\|\nabla f(\mathbf{w})-\nabla f_i(\mathbf{w})\|^2\right] \leq \left(\frac{N-1}{n-1}\right)\tilde{\epsilon}^2, \quad (13)$$

where $\tilde{\epsilon}$ *is given in (9).*

Lemma 2. *Using the uniformly random grouping strategy* S*, for any* $\mathbf{w}$*, the average downward divergence is*

$$\mathbb{E}_{\mathsf{S}}\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{\mathsf{k}\in\mathcal{V}_i}\|\nabla f_i(\mathbf{w})-\nabla F_{\mathsf{k}}(\mathbf{w})\|^2\right] \leq \left(1-\frac{N-1}{n-1}\right)\tilde{\epsilon}^2. \quad (14)$$

Similar to (10), Lemma 1 and Lemma 2 show that the sum of upper bounds for upward and downward divergences is equal to the global divergence. Furthermore, when the number of groups increases, upward divergence becomes larger while downward divergence becomes smaller. When grouping is random, N is the key parameter that determines how global divergence is partitioned. For simplicity, we let each group have the same local period $I_i = I, \forall i$. Then, we can obtain the following theorem.

Theorem 2. *Using the uniformly random grouping strategy* S*, let* $\gamma \leq \frac{1}{2\sqrt{6}GL}$*, then we have*

$$\begin{aligned}\mathbb{E}_{\mathsf{S}}\left[\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{w}}^t)\|^2\right] &\leq \frac{2(f^0-f^*)}{\gamma T}+\frac{\gamma L\sigma^2}{n}\\ &+2C\gamma^2L^2\left[\left(\frac{N-1}{n}\right)G+\left(1-\frac{N}{n}\right)I\right]\sigma^2\\ &+3C\gamma^2L^2\left[\left(\frac{N-1}{n-1}\right)G^2+\left(1-\frac{N-1}{n-1}\right)I^2\right]\tilde{\epsilon}^2, \quad (15)\end{aligned}$$

where $C = 40/3$.

Remark 4. Note that both multiplicative factors of the noise term with σ^2 and the divergence term with $\tilde{\epsilon}^2$ are composed of two parts, where the upward part is "modulated" by G while the downward part is "modulated" by I. As N becomes larger, the upward part has a stronger influence on the convergence bound since both $\frac{N-1}{n}$ and $\frac{N-1}{n-1}$ become larger. Note that the convergence upper bound of H-SGD can be sandwiched by the convergence upper bounds of two local SGD. To see this, we consider the following three scenarios: 1) local SGD with aggregation period $P = G$, 2) local SGD with aggregation period $P = I$, and 3) H-SGD with local aggregation period I and global aggregation period G. We let them all start with the same $\bar{\mathbf{w}}^0$ and choose learning rate γ such that $\gamma \leq \frac{1}{2\sqrt{6}GL}$. We can see that all these three cases have the same first two terms in (15). However, for the third and fourth terms in (15), we have

$$\left(1-\frac{1}{n}\right)I \leq \left(\frac{N-1}{n}\right)G+\left(1-\frac{N}{n}\right)I \leq \left(1-\frac{1}{n}\right)G, \quad (16)$$

$$I^2 \leq \left(\frac{N-1}{n-1}\right)G^2+\left(1-\frac{N-1}{n-1}\right)I^2 \leq G^2. \quad (17)$$

Equations (16) and (17) show that the convergence upper bound of H-SGD has a value between the convergence upper bounds of local SGD with aggregation periods of $P = I$ and $P = G$, respectively. The grouping approach can explicitly characterize how much the convergence bound moves towards the best case, i.e., local SGD with $P = I$. However, this case incurs the highest global communication cost, so grouping can adjust the trade-off between convergence and communication cost.

Remark 5. Even with a large G, if we make I sufficiently small, the convergence bound of H-SGD can be improved. To see this, consider the last two terms of (15). Suppose $G = mI, m = 1, 2, \cdots$. For a non-trivial worker grouping, i.e., $1 < N < n$, if we increase G to $G' = lG, 1 < l < \sqrt{\frac{1}{m^2}\frac{n-N}{N}+1}$ and decrease I to $I' = qI, q \leq \sqrt{1-m^2(l^2-1)\frac{N}{n-N}}$, then one can show that the bound (15) using G' and I' can be lower than that using G and I. A similar behavior can also be seen for fixed grouping, as shown empirically in Figure 3b in the experiments.

H-SGD with More Than Two Levels

In the two-level H-SGD setting discussed in the previous sections, there is only one level of local servers between the global server and workers. In this section, we extend our analysis to H-SGD with more than one level of local servers. Specifically, as shown in Figure 1, we consider a total of $M \geq 2$ levels. For $\ell = 1, 2, \ldots, M-1$, each server at level $\ell-1$ is connected to N_ℓ servers at level ℓ, where we assume that the global server is at a "dummy" level $\ell = 0$ for convenience. For $\ell = M$, each server at level $M-1$ directly connects to N_M workers. As a result, we have a total of $n = \Pi_{\ell=1}^{M} N_\ell$ workers. With this notation, a sequence of indices $(k_1, k_2, \ldots, k_\ell)$ denotes a "path" from the global server at level 0 to a local server/worker at level ℓ, where this path traverses the k_1-th server at level 1, the k_2-th server at level 2 connected to the k_1-th server at level 1, and so on. Hence, the index sequence $(k_1, k_2, \ldots, k_\ell)$ uniquely determines servers (and workers if $\ell = M$) down to level ℓ.

We assume in the multi-level case that each server at level $\ell-1$ connects to the same number (N_ℓ) of servers/workers in the next level ℓ, for ease of presentation. Our results can be extended to the general case with different group sizes.

Let $F_{k_1\ldots k_M}(\mathbf{w})$ denote the objective function of worker $k_1 \ldots k_M$ and $f_{k_1\ldots k_\ell}(\mathbf{w})$ denote the averaged objective func-

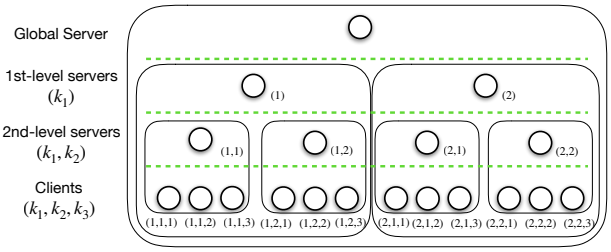


Figure 1: Three-level example with $N_1 = N_2 = 2, N_3 = 3$.

tion of workers connected to server $k_1 \dots k_\ell$,

$$f_{k_1 \dots k_\ell}(\mathbf{w}) := \frac{1}{\prod_{j=\ell+1}^{M} N_j} \sum_{k_{\ell+1}=1}^{N_{\ell+1}} \cdots \sum_{k_M=1}^{N_M} F_{k_1 \dots k_M}(\mathbf{w}). \quad (18)$$

Then the global objective function can be rewritten as

$$\begin{aligned} f(\mathbf{w}) &= \frac{1}{\prod_{j=1}^{M} N_j} \sum_{k_1=1}^{N_1} \cdots \sum_{k_M=1}^{N_M} F_{k_1 \dots k_M}(\mathbf{w}) \\ &= \frac{1}{\prod_{j=1}^{\ell} N_j} \sum_{k_1=1}^{N_1} \cdots \sum_{k_\ell=1}^{N_\ell} f_{k_1 \dots k_\ell}(\mathbf{w}). \end{aligned} \quad (19)$$

Let P_ℓ ($\ell = 1, 2, \dots, M$) denote the period (expressed as the number of local iterations) that the parameters at servers at level ℓ are aggregated by their parent server at level $\ell - 1$. We require that $P_1 > P_2 > \dots > P_M$ and P_ℓ is an integer multiple of $P_{\ell-1}$. The algorithm is in the supplementary material.

In the following, we provide results for the random grouping case. Results for the fixed grouping case can be found in the supplementary material. We apply the divergence partition idea to each level, so we extend Lemma 1 and Lemma 2 to obtain the following lemma.

Lemma 3. *Using the uniformly random grouping strategy* S*, the* ℓ*-th level averaged upward and downward divergences are given by*

$$\mathbb{E}_{\mathsf{S}}\left[\frac{1}{n_\ell} \sum_{k_1=1}^{N_1} \cdots \sum_{k_\ell=1}^{N_\ell} \|\nabla f_{k_1 \dots k_\ell}(\mathbf{w}) - \nabla f(\mathbf{w})\|^2\right] \leq \left(\frac{n_\ell - 1}{n - 1}\right)\tilde{\epsilon}, \quad (20)$$

$$\begin{aligned} &\mathbb{E}_{\mathsf{S}}\left[\frac{n_\ell}{n} \sum_{k_{\ell+1}=1}^{N_{\ell+1}} \cdots \sum_{k_M=1}^{N_M} \|\nabla f_{k_1 \dots k_\ell}(\mathbf{w}) - \nabla F_{k_1 \dots k_M}(\mathbf{w})\|^2\right] \\ &\leq \left(1 - \frac{n_\ell - 1}{n - 1}\right)\tilde{\epsilon}^2 \end{aligned} \quad (21)$$

respectively, $\forall \mathbf{w}$*,* $\forall k_1, \dots, k_\ell$ *in (21), where* $n_\ell = \Pi_{j=1}^{\ell} N_j$ *and* $\tilde{\epsilon}$ *is the global divergence.*

When the number of servers in the ℓth-level n_ℓ increases, the ℓth-level's upward divergence becomes larger while its downward divergence becomes smaller. The sum of the ℓth-level upward and downward divergences are upper bounded by the global divergence. Based on this lemma, we derive the convergence bound for multi-level case as the following.

Theorem 3. *Consider uniform random grouping strategy* S*, if* $\gamma \leq \frac{1}{2\sqrt{6}P_1 L}$*, then for multi-level case,*

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{\mathsf{S}}\left\|\nabla f(\bar{\mathbf{w}}^t)\right\|^2 &\leq \frac{2}{\gamma T}\left(f^0 - f^*\right) + \frac{\gamma L \sigma^2}{n} \\ &+ C\gamma^2 L^2 \frac{1}{M-1} \cdot \sum_{\ell=1}^{M-1} \left[2A_1(\ell)\sigma^2 + 3A_2(\ell)\tilde{\epsilon}^2\right] \end{aligned} \quad (22)$$

where, $A_1(\ell) := P_1\left(\frac{1}{\Pi_{j=\ell}^{M} N_j} - \frac{1}{n}\right) + P_\ell\left(1 - \frac{1}{\Pi_{j=\ell}^{M} N_j}\right)$, $A_2(\ell) := P_1^2\left(\frac{n_\ell - 1}{n-1}\right) + P_\ell^2\left(1 - \frac{n_\ell - 1}{n-1}\right)$.

Remark 6. Note that when $M = 2$, this bound reduces to the bound given in Theorem 2. The effect of the ℓ-th level can be represented as $2A_1(\ell)\sigma^2 + 3A_2(\ell)\tilde{\epsilon}^2$, where each $A_i(\ell), i = 1, 2$ is composed of an upward part "modulated" by P_1 and a downward part "modulated" by P_ℓ. Similar to (16) and (17), we can obtain the sandwich behavior as:

$$\left(1 - \frac{1}{n}\right)P_M \leq \frac{1}{M-1}\sum_{\ell=1}^{M-1} A_1(\ell) \leq \left(1 - \frac{1}{n}\right)P_1, \quad (23)$$

$$P_M^2 \leq \frac{1}{M-1}\sum_{\ell=1}^{M-1} A_2(\ell) \leq P_1^2. \quad (24)$$

It can be seen that the convergence upper bound of H-SGD with more than two levels also takes a value that is between the convergence upper bounds of local SGD with local periods P_1 and P_M, respectively. Compared to the two-level case, this provides greater freedom to choose parameters in $A_1(\ell)$ and $A_2(\ell)$, for $\ell = 1, \dots, M-1$.

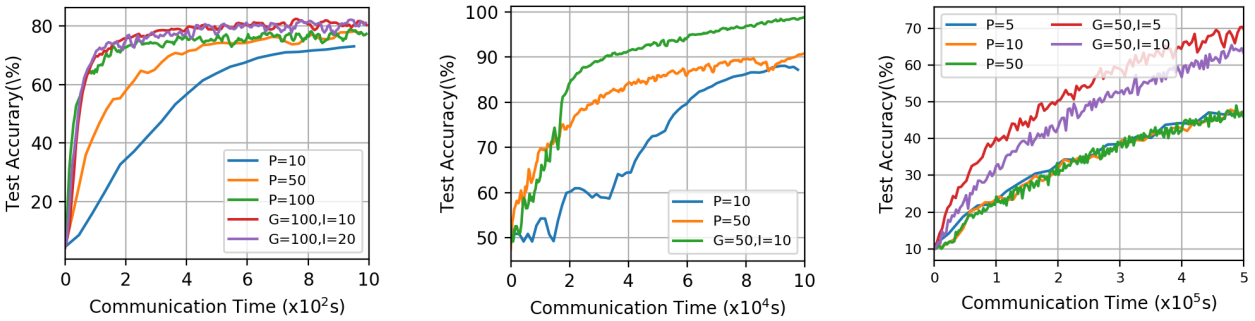
Experiments

In this section, we validate our theoretical results with experiments on training the VGG-11 model over CIFAR-10 (Krizhevsky, Hinton et al. 2009) and CelebA (Liu et al. 2015), and training a convolutional neural network (CNN) over FEMNIST (Cohen et al. 2017), all with non-IID data partitioning across workers. The communication time is emulated by measuring the round-trip time of transmitting the model between a device (in a home) and local (near) / global (far) Amazon EC2 instances. For the computation time, we measured the averaged computation time needed for VGG-11 during one iteration, where we ran SGD with VGG-11 on a single GPU for 100 times and then computed the averaged computation time, which is approximately 4 ms per iteration on each worker. The experiments presented in this section are for *two-level* H-SGD. Additional details of the setup and experiments with *more levels* and *partial worker participation* can be found in the supplementary material.

In all plots, curves with P (denoting the aggregation period) are for local SGD and curves with G (global period), I (local period), and N (number of groups) are for H-SGD. Figure 2 shows that for all the datasets and models evaluated in the experiments, H-SGD can achieve a better accuracy than local SGD given the same communication time. In other words, for any given accuracy, H-SGD can achieve this accuracy using shorter or similar communication time. In Table 2,

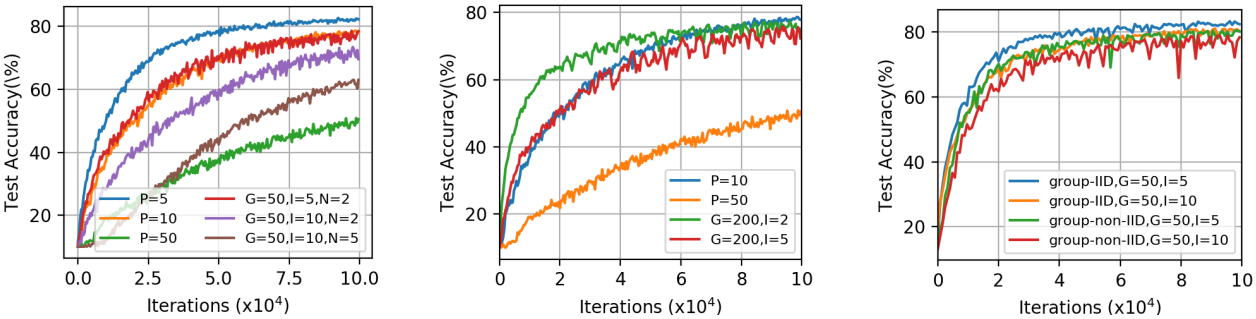
Table 2: Total time (s) needed to achieve 50% test accuracy for VGG-11 with CIFAR-10.

Case	$P = 5$	$P = 10$	$P = 50$	$G = 50, I = 5$	$G = 50, I = 10$
Time (s)	673.5	690.2	944.3	160.3	381.1



(a) FEMNIST with CNN. (b) CelebA with VGG-11. (c) CIFAR-10 with VGG-11.

Figure 2: Test accuracy v.s. communication time ($N = 2$).



(a) Non-IID ($G = 50$ for two-level case). (b) Non-IID ($G = 200$ for two-level case). (c) Group-IID/Group-non-IID.

Figure 3: Results with CIFAR-10. Test accuracy v.s. local iterations. By default, $N = 2$.

we provide the total time (including both communication and computation) needed for VGG-11 with CIFAR-10 to achieve 50% test accuracy (since 50% is nearly the best accuracy for $P = 50$). We can see H-SGD cases ($G = 50, I = 5$ and $G = 50, I = 10$) perform much better than single-level local SGD cases. This is because communication time plays a much more important role and H-SGD can achieve 50% accuracy within less amount of time in total.

To show the "sandwich" behavior, we examine the test accuracy as a function of the total number of local iterations for the non-IID scenario in Figure 3a. We can observe that the performance of H-SGD with global period G and local period I is between that of local SGD with $P = G$ and with $P = I$. We can also observe that as N becomes larger, the performance becomes worse since the upward divergence becomes larger, which is consistent with our analysis in Remark 4. Comparing Figures 3a and 3b, we see that decreasing I while increasing G can even improve the performance, which is consistent with Theorem 2. For example, $G = 200, I = 2$ gives a better performance than $G = 50, I = 5$, while $G = 50, I = 5$ is similar to $P = 10$. This shows that by allowing more local aggregations, H-SGD can reduce the number of global aggregations by 95% ($G = 200$ vs. $P = 10$) while maintaining a similar performance to local SGD. In Figure 3c, we show the effects of grouping corresponding to our analysis for Theorem 1. Group-IID is a grouping strategy which makes upward divergence nearly zero while group-non-IID is a grouping strategy with a large upward divergence. We can see that group-IID performs as well as group-non-IID after reducing I by half.

Conclusion

We have studied H-SGD with multi-level model aggregations. In particular, we have provided a thorough theoretical analysis on the convergence of H-SGD over non-IID data, under both fixed and random worker grouping. We have successfully answered the important question on how local aggregation affects convergence of H-SGD. Based on our novel analysis of the local and global divergences, we established explicit comparisons of the convergence rate for H-SGD and local SGD. Our theoretical analysis provides valuable insights into the design of practical H-SGD systems, including the choice of global and local aggregation periods. Different grouping strategies are considered to best utilize the divergences to reduce communication costs while accelerating learning convergence. In addition, we have extended the analysis approach to H-SGD with arbitrary number of levels. Future work could theoretically analyze the effect of partial worker participation in H-SGD.

Acknowledgment

This research was partly sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

Abad, M. S. H.; Ozfatura, E.; Gunduz, D.; and Ercetin, O. 2020. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8866–8870. IEEE.

Bellet, A.; Kermarrec, A.-M.; and Lavoie, E. 2021. D-Cliques: Compensating NonIIDness in Decentralized Federated Learning with Topology. *arXiv preprint arXiv:2104.07365*.

Bonawitz, K.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C.; Konečnỳ, J.; Mazzocchi, S.; McMahan, H. B.; et al. 2019. Towards Federated Learning at Scale: System Design. In *SysML 2019*.

Bottou, L.; Curtis, F. E.; and Nocedal, J. 2018. Optimization methods for large-scale machine learning. *SAIM Review*, 60(2): 223–311.

Castiglia, T.; Das, A.; and Patterson, S. 2021. Multi-Level Local SGD: Distributed SGD for Heterogeneous Hierarchical Networks. In *International Conference on Learning Representations*.

Chen, C.; Chen, Z.; Zhou, Y.; and Kailkhura, B. 2020. FedCluster: Boosting the Convergence of Federated Learning via Cluster-Cycling. *arXiv preprint arXiv:2009.10748*.

Cohen, G.; Afshar, S.; Tapson, J.; and Van Schaik, A. 2017. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, 2921–2926. IEEE.

Haddadpour, Farzin; et al. 2019. Local SGD with Periodic Averaging: Tighter Analysis and Adaptive Synchronization. In *Advances in Neural Information Processing Systems*.

Han, P.; Wang, S.; and Leung, K. K. 2020. Adaptive Gradient Sparsification for Efficient Federated Learning: An Online Learning Approach. In *IEEE International Conference on Distributed Computing Systems (ICDCS)*.

Jiang, P.; and Agrawal, G. 2018. A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication. In *Advances in Neural Information Processing Systems*.

Jiang, Y.; Wang, S.; Valls, V.; Ko, B. J.; Lee, W.-H.; Leung, K. K.; and Tassiulas, L. 2020. Model Pruning Enables Efficient Federated Learning on Edge Devices. In *NeurIPS Workshop on Scalability, Privacy, and Security in Federated Learning (SpicyFL)*.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.

Kim, H.; and Feamster, N. 2013. Improving network management with software defined networking. *IEEE Communications Magazine*, 51(2): 114–119.

Konecny, J.; McMahan, H. B.; Yu, F. X.; Richtarik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated Learning: Strategies for Improving Communication Efficiency. In *NeurIPS Workshop on Private Multi-Party Machine Learning*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, H.; Ota, K.; and Dong, M. 2018. Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing. *IEEE Network*, 32(1): 96–101.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2019. Federated learning: Challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.

Lin, T.; Stich, S. U.; Patel, K. K.; and Jaggi, M. 2020. Don't Use Large Mini-batches, Use Local SGD. In *International Conference on Learning Representations*.

Liu, L.; Zhang, J.; Song, S.; and Letaief, K. B. 2020. Client-edge-cloud hierarchical federated learning. In *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, 1–6. IEEE.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Luo, S.; Chen, X.; Wu, Q.; Zhou, Z.; and Yu, S. 2020. HFEL: Joint Edge Association and Resource Allocation for Cost-Efficient Hierarchical Federated Edge Learning. *IEEE Transactions on Wireless Communications*, 19(10): 6535–6548.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 1273–1282.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive Federated Optimization. arXiv:2003.00295.

Stich, S. U. 2019. Local SGD Converges Fast and Communicates Little. In *International Conference on Learning Representations*.

Wang, J.; and Joshi, G. 2019. Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. In *ICML 2019 Workshop on Coding Theory for Large-Scale ML*.

Wang, S.; Tuor, T.; Salonidis, T.; Leung, K. K.; Makaya, C.; He, T.; and Chan, K. 2019. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. *IEEE Journal on Selected Areas in Communications*, 37(6): 1205–1221.

Yu, H.; Jin, R.; and Yang, S. 2019. On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization. In *ICML*, 7184–7193.

Yu, H.; Yang, S.; and Zhu, S. 2019. Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning. In *AAAI*.

Zhou, F.; and Cong, G. 2019. A distributed hierarchical SGD algorithm with sparse global reduction. *arXiv preprint arXiv:1903.05133*.

Zinkevich, M.; Weimer, M.; Li, L.; and Smola, A. J. 2010. Parallelized stochastic gradient descent. In *NeurIPS*, 2595–2603.