Revisiting LocalSGD and SCAFFOLD: Improved Rates and Missing Analysis

Ruichen Luo
IST Austria
rluo@ist.ac.at

Sebastian U. Stich
CISPA Helmholtz Center
stich@cispa.de

Samuel Horváth
MBZUAI
samuel.horvath@mbzuai.ac.ae

Martin Takáč
MBZUAI
Takac.MT@gmail.com

Abstract

LocalSGD and SCAFFOLD are widely used methods in distributed stochastic optimization, with numerous applications in machine learning, large-scale data processing, and federated learning. However, rigorously establishing their theoretical advantages over simpler methods, such as minibatch SGD (MbSGD), has proven challenging, as existing analyses often rely on strong assumptions, unrealistic premises, or overly restrictive scenarios.

In this work, we revisit the convergence properties of LocalSGD and SCAFFOLD under a variety of existing or weaker conditions, including gradient similarity, Hessian similarity, weak convexity, and Lipschitz continuity of the Hessian. Our analysis shows that (i) LocalSGD achieves faster convergence compared to MbSGD for weakly convex functions without requiring stronger gradient similarity assumptions; (ii) LocalSGD benefits significantly from higher-order similarity and smoothness; and (iii) SCAFFOLD demonstrates faster convergence than MbSGD for a broader class of non-quadratic functions. These theoretical insights provide a clearer understanding of the conditions under which LocalSGD and SCAFFOLD outperform MbSGD.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

This paper focuses on solving the following *distributed non-convex optimization* problem:

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

where $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable, f is the global objective function, n is the number of workers, and f_i is the local objective function distributed to the ith worker, $i \in [n] := \{1, \cdots, n\}$. The local objective functions are *heterogeneous* in general, *i.e.*, $f_i \neq f$.

Distributed non-convex optimization. Large scale optimization problems arise in various fields of science and engineering, such as management science (Magnanti, 2021), signal processing (So et al., 2020), machine learning (Bottou et al., 2018), among others. Given the scale of the problems, it is essential to consider distributed settings, where the optimization tasks are divided among multiple workers who work in parallel to solve the problem. In many real-world applications, the objective functions are often non-convex, making the optimization process more challenging compared to convex problems.

Stochastic gradient methods with intermittent communication. For non-convex optimization problems, stochastic gradient descent (SGD) (Robbins and Monro, 1951) is a basic and strong method widely used, *e.g.*, in training machine learning models (Bottou, 2010). When applying SGD in distributed settings, the communication between the workers often becomes the bottleneck (Jaggi et al., 2014; McMahan et al., 2017). To adapt SGD to distributed setting, many communication-efficient algorithms are designed so that the workers do more local computations and communicate only intermittently.

MbSGD vs. LocalSGD/SCAFFOLD. Among the communication-efficient methods, we focus on three. Mini-batch SGD (or MbSGD) (Woodworth et al., 2020b; Takáč et al., 2013) is the simplest baseline and is well-understood, while LocalSGD (McMahan et al., 2017; Stich, 2019) and SCAFFOLD (Karimireddy et al., 2020) are another two methods gaining great popularity recently. At their core, all three methods aim to reduce communication by performing more local computations between communication rounds. In MbSGD, each worker computes a more accurate stochastic gradient using a larger mini-batch; in LocalSGD, each worker takes several local gradient steps before communicating; and in SCAFFOLD, those local steps are enhanced with SAGA-like variance reduction techniques (Defazio et al., 2014). All these methods have been widely adopted in distributed optimization, with their original papers amassing thousands of citations as of 2024. Their importance is reflected in their integration into open-source federated learning frameworks such as FedML (He et al., 2020), NVIDIA Flare (Roth et al., 2022), and Flower (Beutel et al., 2020). While in practice, LocalSGD and SCAFFOLD often achieve faster convergence than MbSGD, there remains a notable gap between theory and practice. From a theoretical point of view, MbSGD is a formidable baseline, especially in the presence of heterogeneity, where it dominates most of the existing analyses.

Open questions and missing analysis. Despite the growing interest in distributed non-convex optimization, popular methods like LocalSGD and SCAFFOLD remain not well understood even under the classic assumptions, such as gradient similarity (*a.k.a.* bounded gradient dissimilarity), Hessian similarity (*a.k.a.* bounded Hessian dissimilarity), weak convexity, and Lipschitz continuous Hessian. Several theoretical questions remain open. For instance, is there any speedup analysis of LocalSGD for non-convex functions? Can LocalSGD converge faster without the assumption of uniform gradient similarity? Can LocalSGD benefit from higher-order conditions? Is there any speedup analysis of SCAFFOLD for general non-convex functions, potentially without the stringent assumption of uniform Hessian similarity? And to what extent can we relax the assumptions for quadratic functions? Given the growing body of work in this field and the significant impact of these two algorithms, it is crucial to properly understand their convergence behaviors, at least under the classic assumptions. Moreover, since these two algorithms have become standard baselines for communication-efficient distributed optimization, ensuring that their convergence rates are cited correctly is essential for fair algorithmic comparisons and future advancements in the field.

Contributions. To address the above issues, we revisit the convergence rates of LocalSGD and SCAFFOLD under gradient similarity, Hessian similarity, weak convexity, and Lipschitz Hessian. Our key contributions are as follows:

- We show that LocalSGD can converge faster than MbSGD for weakly convex functions, marking the first time such a speedup has been extended to non-convex analysis.
- We show that standard gradient similarity is sufficient for LocalSGD to achieve a speedup, making the stronger condition of uniform gradient similarity (as used in Woodworth et al. (2020b); Patel et al. (2024)) redundant.
- We show that LocalSGD, a basic gradient method without variance reduction or quasi-Newton tricks, can also benefit from higher-order conditions.
- We show that SCAFFOLD can converge faster than MbSGD under standard Hessian similarity and weak convexity. This is the first analysis demonstrating such a speedup for general non-quadratic functions without relying on uniform Hessian similarity.
- We introduce a weaker assumption: the existence of a Lipschitz continuous function within the convex hull of the distributed functions. This assumption underpins our proof that LocalSGD benefits from higher-order conditions and also allows us to obtain a similar speedup result for SCAFFOLD as seen in Karimireddy et al. (2020).

We summarize our results and compare them with the existing analyses of different methods in Table 1.

Structure of the paper. We discuss the related work in Section 2. Then, we formally define the problem, algorithms, and assumptions in Section 3. We review the current rates of MbSGD, LocalSGD, and SCAFFOLD in Section 4. Our new analysis is presented in Section 5, with key implications and technical innovations highlighted in the main text, while detailed proofs are deferred to Appendix A. In Section 6, we present suitable synthetic experiments that validate our theoretical findings. We conclude with a discussion of limitations and future work.

2 Related Work

Classic assumptions and existing analysis. Early analysis focuses on the more restrictive, homogeneous setting (Zhou and Cong, 2018; Yu et al., 2019; Stich, 2019; Woodworth et al., 2020a). This paper, instead, considers the more general, heterogeneous settings. Recent works on the heterogeneous settings often consider the following classic assumptions: gradient similarity, Hessian similarity, weak convexity, and Lipschitz continuous Hessian. Below, we briefly summarize the existing analyses based on these assumptions:

- Koloskova et al. (2020); Karimireddy et al. (2020) show that LocalSGD can only match MbSGD under standard gradient similarity. Moreover, the

Table 1: Summary of the asymptotic non-convex rates of the relevant gradient-based distributed optimization algorithms. Parameters include ζ (or $\bar{\zeta}$) for gradient similarity, δ (or $\bar{\delta}$) for Hessian similarity, ρ for weak convexity, and $\mathcal{M}$ for the Lipschitz continuity of the Hessian. "Speedup" refers to whether the algorithm can converge faster than the MbSGD baseline. "General" indicates there is no additional restriction on the class of functions. "Deterministic" specifies if the algorithm's randomness only comes from the stochastic oracle. "Oracle" refers to the type of gradient oracle used in the algorithm. Detailed definitions are provided in Section 3.

Algorithm	Analysis	Suboptimality	Speedup	General	Deterministic	Oracle
MbSGD	Dekel et al. (2012)	$\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}}$ [a]	baseline	✓	✓	fully stochastic
LocalSGD	Koloskova et al. (2020)	$\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{L\Delta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}$ [a]	✗	✓	✓	fully stochastic
	Woodworth et al. (2020b)	$\frac{LD^2}{\tau R} + \frac{\sigma D}{\sqrt{n\tau R}} + \left(\frac{L\bar{\zeta}^2 D^4}{R^2}\right)^{\frac{1}{3}} + \left(\frac{L\sigma^2 D^4}{\tau R^2}\right)^{\frac{1}{3}}$ [b]	✓	convex	✓	fully stochastic
SCAFFOLD	Karimireddy et al. (2020)	$\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}}$ [c]	✗	✓	✓	fully stochastic
	Karimireddy et al. (2020)	$\left(\frac{L}{\tau} + \bar{\delta} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}}$ [a]	✓	quadratic	✓	exact [d]
ScaffNew	Mishchenko et al. (2022)	$(1-\mu/L)^T D^2 + (1-\theta)^T D^2 + \eta^2\frac{\sigma^2}{\theta}$ [e]	✓	μ-strongly convex	✗	fully stochastic
CE-LSGD	Patel et al. (2022)	$\left(\frac{\bar{L}}{\sqrt{\tau}} + \bar{\delta}\right)\frac{\Delta}{R} + \frac{\sigma^2}{n\tau R} + \left(\frac{\bar{L}\Delta\sigma}{n\tau R}\right)^{\frac{2}{3}}$ [a]	✓	✓	✗	stochastic multi-point[f]
LocalSGD	**Theorem 1**	$\left(\frac{L}{\tau} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{L\Delta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}$ [a]	✓	✓	✓	fully stochastic
	Theorem 2	$\frac{LD^2}{\tau R} + \frac{\sigma D}{\sqrt{n\tau R}} + \left(\frac{L\zeta^2 D^4}{R^2}\right)^{\frac{1}{3}} + \left(\frac{L\sigma^2 D^4}{\tau R^2}\right)^{\frac{1}{3}}$ [b]	✓	convex	✓	fully stochastic
	Theorem 1 & 3 [g]	$\left(\frac{L}{\tau} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{\bar{\delta}\Delta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}} + \left(\frac{\mathcal{M}^2\Delta^4\zeta^4}{R^4}\right)^{\frac{1}{5}}$ [a]	✓	✓	✓	fully stochastic
SCAFFOLD	**Theorem 4**	$\left(\frac{L}{\tau} + \sqrt{L\delta} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}$ [a]	✓	✓	✓	fully stochastic
	Theorem 5	$\left(\frac{L}{\tau} + \sqrt{\bar{\delta}\delta} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \frac{(\bar{\delta}\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}$ [a]	✓	Assumption 4 ($\mathcal{M}=0$)	✓	fully stochastic

[a]The suboptimality of $\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla f(\bar{\mathbf{x}}_t)\|_2^2$. This is standard in the analysis of non-convex functions (Nesterov, 2003).

[b]The suboptimality of $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - f^*$. This is standard in the analysis of convex functions (Nesterov, 2003).

[c]The suboptimality of $\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_{2r\tau})\|_2^2$. In their analysis, they combine all the iterates between two communication rounds into one big step and analyze the descent. Moreover, for theoretical purposes, they use a different global stepsize η_g to aggregate the workers' updates.

[d]They use exact gradients in the variance reduction term.

[e]The suboptimality of $\mathbb{E}\|\bar{\mathbf{x}}_T - \mathbf{x}^*\|_2^2$. This is standard in the analysis of strongly-convex functions (Nesterov, 2003). In this rate, $\theta := \min\{\mu\eta, p^2\}$, where p is the communication probability.

[f]In CE-LSGD, they use a stochastic multi-point gradient oracle, where each worker can query $G_i(\cdot, \xi_t^i)$ again at a different point in the next iteration. Moreover, they need to assume a stronger $\bar{L}$-mean smoothness in their analysis, *i.e.*, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\mathbb{E}_{\xi_t^i}\|G_i(\mathbf{x}, \xi_t^i) - G_i(\mathbf{y}, \xi_t^i)\|_2 \leq \bar{L}\|\mathbf{x} - \mathbf{y}\|_2$.

[g]This rate can be obtained by combining the proofs of Theorems 1 and 3.

speedup of LocalSGD is only proven for convex functions, and this relies on a stronger condition of uniform gradient similarity (Woodworth et al., 2020b).

- For quadratic functions, Karimireddy et al. (2020) show the speedup of SCAFFOLD under uniform Hessian similarity and weak convexity. However, for more general non-quadratic functions, there is no theoretical proof of a speedup for SCAFFOLD.

Related work on new algorithms. Distributed non-convex optimization continues to be an active research area. Perhaps motivated by the limitations of the current theories of LocalSGD and SCAFFOLD, many recent works focus on designing new algorithms. Examples include ProxSkip (or ScaffNew) (Mishchenko et al., 2022), BVR-L-SGD (Murata and Suzuki, 2021), CE-LSGD (Patel et al., 2022), MimeMVR (Karimireddy et al., 2021), DANE+ and FedRed (Jiang et al., 2024). Each of these works considers somehow different settings. Mishchenko et al. (2022) proposes ScaffNew, a variant of the SCAFFOLD algorithm with randomized communication rounds, and shows acceleration from local steps, but their result holds only in expectation for strongly convex functions. Murata and Suzuki (2021); Patel et al. (2022) study algorithms with stochastic multi-point oracle and randomized worker sampling, while the algorithms considered in this paper are deterministic and only use the fully stochastic gradient oracle, which is more general. Karimireddy et al. (2021) assumes each local function f_i is a finite sum of components with similar Hessians across workers, making the setting more restrictive. Jiang et al. (2024) studies variance-reduced proximal methods, which are more difficult to implement, and it is not clear how to handle stochastic noise. Thus far, we are not aware of any algorithm that uses a fully stochastic oracle and converges provably better than MbSGD/LocalSGD/SCAFFOLD for general non-convex functions.

Related work on new assumptions. A parallel line of research focuses on analyzing LocalSGD under new assumptions. Wang et al. (2024) introduces a novel assumption that the (so-called) average pseudo-gradient is bounded at optimum. However, this analysis is problematic: if the assumed bound is near zero, LocalSGD converges in constant communication rounds; otherwise, it diverges. Moreover, Patel et al. (2023) proves the lower bound, showing that this assumption fails for non-convex functions. Zindari et al. (2023) considers quadratic functions under the assumption of identical Hessians or identical optima, but these are overly restrictive and do not apply to general, non-convex functions. Ultimately, none of these works provide a reasonable explanation for the fast convergence observed in practice for LocalSGD.

Related work on generalization. In the context of distributed learning, our optimization problem (1) corresponds to empirical risk minimization. Some more recent works, under different set of assumptions, focus on bounding the generalization gap (Sefidgaran et al., 2024; Sun et al., 2024; Gu et al., 2023). These works are orthogonal or somewhat complementary to this paper, as the true risk in distributed learning is bounded by the sum of the empirical risk and the generalization gap. We include a more detailed discussion in Appendix C.

3 Preliminary

We start with the definition of the optimization problem. Then, we describe the algorithms of interest and, finally, the technical assumptions.

3.1 Distributed Stochastic Non-Convex Optimization with Intermittent Communication

Distributed non-convex optimization. We recall the distributed non-convex optimization problem:

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

and we assume throughout the paper that f is bounded below, and $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is possibly non-convex and has L-Lipschitz continuous gradient ($L>0$), for all $i\in[n]$.

Stochastic oracle. We focus on solving problem (1) by iterative algorithms via subsequent queries to a *fully stochastic oracle* $\mathcal{SO}$. Let T be the total number of successive queries to $\mathcal{SO}$. At iteration t, where $0\le t\le T-1$, the workers' inputs being $\left(\mathbf{x}_t^1,\cdots,\mathbf{x}_t^n\right)\in\mathbb{R}^{d\times n}$, the $\mathcal{SO}$ outputs vectors

$$\left(\mathbf{g}_t^1,\cdots,\mathbf{g}_t^n\right) := \left(G_1(\mathbf{x}_t^1,\xi_t^1),\cdots,G_n(\mathbf{x}_t^n,\xi_t^n)\right)\in\mathbb{R}^{d\times n},$$

where $\left\{\xi_t^i : 0\le t\le T-1\right\}$ are i.i.d. random variables for each $i\in[n]$. We assume the following conditions on the Borel functions $G_i(\mathbf{x},\xi_t^i)$:

$$\begin{aligned}\mathbb{E}_{\xi_t^i}\big[G_i(\mathbf{x},\xi_t^i)\big] &= \nabla f_i(\mathbf{x}),\\ \mathbb{E}_{\xi_t^i}\big\|G_i(\mathbf{x},\xi_t^i)-\nabla f_i(\mathbf{x})\big\|_2^2 &\le \sigma^2.\end{aligned} \tag{2}$$

Intermittent communication. In distributed optimization, each worker $i\in[n]$ has access only to its local function f_i, which is not visible to the other workers. The optimization process proceeds through *intermittent communication*, where workers communicate periodically after every *interval* of $\tau\in\mathbb{Z}_{\ge 2}$ iterations. We assume w.l.o.g. that T is a multiple of τ. That is, after the r-th communication round ($r\in[0,T/\tau-1]$), each worker $i\in[n]$ can access the gradients $\{\mathbf{g}_{r\tau}^i,\cdots,\mathbf{g}_{r\tau+\tau-1}^i\}$, and only then communicates with the other workers at the subsequent $(r+1)$-th communication round.

To simplify the notations, let $\bar{\mathbf{x}}_t=\frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^i$, $f^*=\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, and $\Delta=f(\bar{\mathbf{x}}_0)-f^*$.

3.2 Algorithms

In the context of distributed non-convex optimization with intermittent communication, we focus on the analysis of three algorithms, MbSGD, LocalSGD, and SCAFFOLD. Below, we give their formal descriptions. In these algorithms, we assume $\mathbf{x}_0^1=\cdots=\mathbf{x}_0^n$.

MbSGD and LocalSGD. Both MbSGD and LocalSGD are two well-recognized, basic methods in distributed optimization. Let R be the total number of communication rounds, *i.e.*, $R=T/\tau$.

- **MbSGD.** Between two communication rounds, each worker computes its gradient estimate on a τ-times larger mini-batch and communicates at the next communication round. Formally, for each $i\in[n]$ and $t\in[0,T-1]$,

$$\mathbf{x}_{t+1}^i = \begin{cases}\bar{\mathbf{x}}_{t-\tau+1}-\frac{\eta}{n}\sum_{j=1}^n\sum_{k=0}^{\tau-1}\mathbf{g}_{t-k}^j, & \text{if } t+1 \text{ is a multiple of } \tau,\\ \mathbf{x}_t^i, & \text{otherwise.}\end{cases} \tag{3}$$

- **LocalSGD.** Between two communication rounds, each worker takes τ local steps using its gradient estimates. Then, at the next communication round, all workers average their solutions globally. Formally, for each $i\in[n]$ and $t\in[0,T-1]$,

$$\mathbf{x}_{t+1}^i = \begin{cases}\bar{\mathbf{x}}_{t-\tau+1}-\frac{\eta}{n}\sum_{j=1}^n\sum_{k=0}^{\tau-1}\mathbf{g}_{t-k}^j, & \text{if } t+1 \text{ is a multiple of } \tau,\\ \mathbf{x}_t^i-\eta\mathbf{g}_t^i, & \text{otherwise.}\end{cases} \tag{4}$$

SCAFFOLD. SCAFFOLD, introduced by Karimireddy et al. (2020), incorporates variance reduction into the local steps to address the issue of client drift. Suppose there are $2R$ communication rounds, *i.e.*, $2R = T/\tau$. The algorithm considered in this paper is a simplified variant of the original SCAFFOLD algorithm. As described in Algorithm 1, across two consecutive communication rounds, in the first round, each worker computes a stochastic gradient on a τ-times larger mini-batch; and in the second round, they take variance-reduced gradient steps using the gradients computed in the first round. The original algorithm in Karimireddy et al. (2020) is a bit more complex, as it allows partial worker participation, the use of the last update for variance reduction, and different local/global stepsizes. We drop these features because none of them are supported in their speedup analysis (*cf.* Section 6 of Karimireddy et al. (2020)). Also, the use of a different global stepsize is mainly for theoretical purposes and is not a common practice in real-world applications (*cf.* Section 7.1 of Karimireddy et al. (2020)). Hence, our simplified version retains the core variance reduction mechanism while streamlining the analysis.

Algorithm 1 SCAFFOLD

for $r = 0, 1, \cdots, R-1$ **do**
  for $i \in [n]$ **do in parallel**
    for $k = 0, 1, \cdots, \tau - 1$ **do**
      $\mathbf{x}^i_{2r\tau+k+1} = \mathbf{x}^i_{2r\tau+k}$
    end for
    $\hat{\mathbf{g}}^i_{(r\tau)} = \frac{1}{\tau}\sum_{k=0}^{\tau-1} \mathbf{g}^i_{2r\tau+k}$
  end for
  Compute and broadcast: $\hat{\mathbf{g}}_{(r\tau)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}^i_{(r\tau)}$
  for $i \in [n]$ **do in parallel**
    for $k = \tau, \tau+1, \cdots, 2\tau - 2$ **do**

$$\mathbf{x}^i_{2r\tau+k+1} = \mathbf{x}^i_{2r\tau+k} - \eta\left(\mathbf{g}^i_{2r\tau+k} - \hat{\mathbf{g}}^i_{(r\tau)} + \hat{\mathbf{g}}_{(r\tau)}\right) \tag{5}$$

    end for
  end for
  Compute: $\bar{\mathbf{x}}_{2(r+1)\tau} = \bar{\mathbf{x}}_{2r\tau} - \frac{\eta}{n}\sum_{j=1}^{n}\sum_{l=\tau}^{2\tau-1} \mathbf{g}^j_{2r\tau+l}$
  Broadcast: $\mathbf{x}^i_{2(r+1)\tau} = \bar{\mathbf{x}}_{2(r+1)\tau}$, for each $i \in [n]$
end for

3.3 Assumptions

We present the technical assumptions in this section and discuss their relevance and implications.

Standard gradient similarity is required to prove the convergence of LocalSGD. Similar assumptions are used in many previous works (Koloskova et al., 2020; Karimireddy et al., 2020).

Assumption 1 (Standard gradient similarity)**.** *For some $\zeta \geq 0$, we have*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \leq \zeta^2. \tag{6}$$

Some recent works, such as Woodworth et al. (2020b); Patel et al. (2024), rely on the stronger assumption of uniform gradient similarity.

Assumption 1+ (Uniform gradient similarity)**.** *For some $\bar{\zeta} \geq 0$, we have*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \sup_{i\in[n]} \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \leq \bar{\zeta}^2 \tag{7}$$

Remark 1. Based on the definitions, ζ is always less than or equal to $\bar{\zeta}$. In cases where the heterogeneity is very imbalanced between the workers, ζ could be significantly smaller–up to a factor of $n^{1/2}$. Our speedup analysis of LocalSGD avoids the need for this stronger assumption, which makes it applicable to scenarios with more heterogeneous data. □

To study the speedup of taking local steps in LocalSGD and SCAFFOLD, following Karimireddy et al. (2020, 2021), we introduce the assumptions of standard/uniform Hessian similarity and weak convexity:

Assumption 2 (Standard Hessian similarity)**.** *For some $\delta \in [0, L]$, we have*

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x}) - \nabla f_i(\mathbf{y}) + \nabla f(\mathbf{y})\|_2^2 \\ \leq \delta^2 \|\mathbf{x} - \mathbf{y}\|_2^2, \end{aligned} \tag{8}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Assumption 2+ (Uniform Hessian similarity)**.** *For some $\bar{\delta} \in [0, 2L]$, we have*

$$\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x}) - \nabla f_i(\mathbf{y}) + \nabla f(\mathbf{y})\|_2 \leq \bar{\delta}\|\mathbf{x} - \mathbf{y}\|_2, \tag{9}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and for all $i \in [n]$.

Assumption 3 (Weak convexity)**.** *For some $\rho \in [0, L]$, we have*

$$f_i(\mathbf{x}) + \frac{\rho}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x} \text{ is convex,}$$

for all $i \in [n]$.

Remark 2. We remark on the above assumptions.

- Hessian similarity has been widely used in the literature of distributed optimization (Shamir et al., 2014; Karimireddy et al., 2020, 2021). Despite the name, we do not necessitate that f_i is twice continuously differentiable. Indeed, all functions with L-Lipschitz continuous gradient satisfy L-standard Hessian similarity and $2L$-uniform Hessian similarity.
- Weak convexity is often used in distributed optimization to show the effectiveness of the local gradient methods (Karimireddy et al., 2020, 2021). As a matter of fact, all the functions with L-Lipschitz continuous gradients are L-weakly convex, and the convex functions are 0-weakly convex.

In this paper, when adopting these assumptions, we always assume $\delta \ll L$, $\bar{\delta} \ll 2L$, and $\rho \ll L$. □

Finally, we introduce a new variant of Lipschitz continuous Hessian.

Assumption 4 (Lipschitz continuous Hessian)**.** *For some $\mathcal{M} \geq 0$, there exists (at least) one function $\hat{f}$ such that: $\hat{f} \in \mathbf{conv}\{f_1, \cdots, f_n\}$,* [1] *and*

$$\left\|\nabla^2 \hat{f}(\mathbf{x}) - \nabla^2 \hat{f}(\mathbf{y})\right\|_2 \leq \mathcal{M} \left\|\mathbf{x} - \mathbf{y}\right\|_2, \tag{10}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Remark 3. Quadratic or Lipschitz continuous Hessian has been widely assumed in the literature of distributed optimization (Shamir et al., 2014; Hendrikx et al., 2020; Karimireddy et al., 2020). Nonetheless, our Assumption 4 might represent the weakest assumption among similar ones made in the literature. It is significantly weaker than requiring all f_i functions to be quadratic, *i.e.*, 0-Lipschitz continuous Hessian. Moreover, it is also weaker than the assumption that "there exists $i \in [n]$ such that $\nabla^2 f_i(x)$ is $\mathcal{M}$-Lipschitz continuous." [2] □

4 Existing Convergence Analyses

In this section, we formally state the preliminary results of MbSGD, LocalSGD and SCAFFOLD in literature, and discuss the open questions and missing analysis.

We first state the well known rate of MbSGD. This rate remains tight under all of Assumptions 1 to 4.

Lemma 1 (Dekel et al. (2012))**.** *There exists $\eta > 0$ such that* MbSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2$:*

$$\mathcal{O}\left(\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}}\right). \tag{11}$$

Remark 4. As a simplification of (11), for the case $L = \Delta = \sigma = 1$ and $\tau = \infty$, we get the asymptotic convergence rate of $\mathcal{O}\left(\frac{1}{R}\right)$ for MbSGD. □

Now, we state the best-known rate of LocalSGD for non-convex functions.

Lemma 2 (Koloskova et al. (2020))**.** *Under Assumption 1, there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2$:*

$$\mathcal{O}\left(\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{L\Delta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}\right). \tag{12}$$

Remark 5. We see that the rate of LocalSGD in (12) is no faster than that of MbSGD in (11). As a simplification of (12), for the case $L = \Delta = \sigma = 1$ and $\tau = \infty$, the asymptotic rate of LocalSGD is $\mathcal{O}\left(\frac{1}{R} + \left(\frac{\zeta}{R}\right)^{2/3}\right)$. Hence, the conditioning of $\zeta^2 = \mathcal{O}\left(1/R\right)$ is required for LocalSGD to match the rate of MbSGD. □

We also present the only known speedup result of LocalSGD for convex functions, as established in Woodworth et al. (2020b).

Lemma 3 (Woodworth et al. (2020b))**.** *Under Assumption 1+, if all the local functions f_i are convex, $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, and there exists some $D \geq 0$ such that $\left\|\bar{\mathbf{x}}_0 - \mathbf{x}^*\right\|_2 \leq D$, then there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[f\left(\bar{\mathbf{x}}_t\right)\right] - f^*$:*

$$\mathcal{O}\left(\frac{LD^2}{\tau R} + \frac{\sigma D}{\sqrt{n\tau R}} + \left(\frac{L\bar{\zeta}^2 D^4}{R^2}\right)^{\frac{1}{3}} + \left(\frac{L\sigma^2 D^4}{\tau R^2}\right)^{\frac{1}{3}}\right). \tag{13}$$

Remark 6. Lemma 3 shows that LocalSGD can converge faster than MbSGD. The speedup is reflected in the optimization term, the first term in (13), from $\mathcal{O}\left(\frac{1}{R}\right)$ to $\mathcal{O}\left(\frac{1}{\tau R}\right)$. In this speedup analysis, however, the heterogeneity term, the third term in (13), gets worse as compared to (12). Suppose we assume only ζ-standard gradient similarity in Lemma 3. Then, the same analysis requires replacing $\bar{\zeta}^2$ with $n\zeta^2$ in (13). Therefore, when the heterogeneity is very imbalanced between the workers, this bound can have a dependence of $n^{1/3}$ with respect to the number of workers, which is counterintuitive, as we typically expect faster convergence with the increasing number of workers. □

Open questions for LocalSGD. The following two questions remain open: (i) *whether for non-convex functions there can be a speedup analysis*, and (ii) *whether the stronger condition of uniform gradient similarity can be avoided.*

Then, we state the convergence rates of SCAFFOLD.

Lemma 4 (Karimireddy et al. (2020))**.** *Suppose in Line 14 of Algorithm 1, a different global stepsize η_g can be used when aggregating the updates of the workers. There exists $\eta_g \geq \eta > 0$ such that* SCAFFOLD *ensures the following upper bound on $\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{2r\tau})\right\|_2^2$:*

$$\mathcal{O}\left(\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}}\right). \tag{14}$$

[1] The convex hull $\mathbf{conv}\{f_1, \cdots, f_n\}$ denotes $\left\{\sum_{i\in[n]} a_i f_i \mid \sum_{i\in[n]} a_i = 1 \text{ and } a_i \geq 0, \text{ for all } i \in [n]\right\}$.

[2] For instance, for $f_1(x) := \frac{1}{2}(x+2)^2 + 2\log(1+x^2)$ and $f_2(x) := \frac{1}{2}(x-1)^2 - \log(1+x^2)$, while f_1 has 6-Lipschitz continuous Hessian and f_2 has 3-Lipschitz continuous Hessian, we have $\hat{f}(x) = \frac{1}{3}f_1(x) + \frac{2}{3}f_2(x)$ with 0-Lipschitz continuous Hessian.

Lemma 5 (Karimireddy et al. (2020)). *Suppose exact gradients can be used for variance reduction,* i.e., $\hat{\mathbf{g}}^i_{(r\tau)} = \nabla f_i(\bar{\mathbf{x}}_{2r\tau})$ *in Line 6 of Algorithm 1. Under Assumptions 2+ and 3, if all the local functions f_i are quadratic, then there exists $\eta > 0$ such that* SCAFFOLD *ensures the following upper bound on* $\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{2r\tau+\tau+k})\right\|_2^2$*:*

$$\mathcal{O}\left(\left(\frac{L}{\tau}+\bar{\delta}+\rho\right)\frac{\Delta}{R}+\sqrt{\frac{L\Delta\sigma^2}{n\tau R}}\right). \tag{15}$$

Remark 7. Lemma 4 shows that SCAFFOLD, with the rescaling of local and global stepsizes, converges at the same rate as MbSGD for general non-convex functions. Further, Lemma 5 shows that, under Hessian similarity and weak convexity, SCAFFOLD converges faster than MbSGD for quadratic functions. The speedup is reflected in the optimization term, from $\mathcal{O}\left(\frac{1}{R}\right)$ to $\mathcal{O}\left((\frac{1}{\tau}+\bar{\delta}+\rho)\frac{1}{R}\right)$. □

Open question for SCAFFOLD. One key question remains open: *whether any speedup can be proved for general non-quadratic functions and/or without uniform Hessian similarity.*

Recently, there has been growing interest in designing new algorithms under the various assumptions of Hessian similarity, weak convexity, and/or Lipschitz continuous Hessian. Several works have referred to the rates of LocalSGD and/or SCAFFOLD as baselines for comparisons, including Karimireddy et al. (2020, 2021); Shen et al. (2021); Khaled and Jin (2023), to name a few. However, it remains unclear whether these cited rates are tight and correctly interpreted across different settings. Therefore, the following issue is subtle but crucial for fair algorithmic comparisons in the field:

Missing analysis. The following analysis is missing: *whether the rates of* LocalSGD *and* SCAFFOLD *can be improved under the various classic assumptions.*

5 New Convergence Analysis

To address the aforementioned open problems and missing issues, we revisit the analysis of LocalSGD and SCAFFOLD. We highlight the technical novelty in our analysis, and defer the proofs to Appendix A.

5.1 New Analysis of LocalSGD

We show that the classic rates of LocalSGD can be improved in the various settings: (i) Theorem 1 shows the speedup under weak convexity and Theorem 2 shows the improvement for convex functions, both without uniform gradient similarity; (ii) Theorem 3 shows a possible improvement under Hessian similarity and Lipschitz continuous Hessian.

Theorem 1. *Under Assumptions 1 and 3, there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2$*:*

$$\mathcal{O}\left(\left(\frac{L}{\tau}+\rho\right)\frac{\Delta}{R}+\sqrt{\frac{L\Delta\sigma^2}{n\tau R}}+\left(\frac{L\Delta\zeta}{R}\right)^{\frac{2}{3}}+\frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}\right). \tag{16}$$

Remark 8. As compared to the rate of MbSGD (Lemma 1) and the previous rate of LocalSGD (Lemma 2), our new rate of LocalSGD in Theorem 1 has a faster optimization term of $\mathcal{O}\left((\frac{L}{\tau}+\rho)\frac{\Delta}{R}\right)$. As a simplification of (16), for the case $L=\Delta=\sigma=1$ and $\tau=\infty$, the asymptotic rate of LocalSGD is $\mathcal{O}\left(\frac{\rho}{R}+\left(\frac{\zeta}{R}\right)^{2/3}\right)$. We see that the conditioning of $\zeta^2=\mathcal{O}(1/R)$ ensures that LocalSGD converges faster than MbSGD. To the best of our knowledge, this is the first speedup analysis of LocalSGD for non-convex functions. Moreover, this result should not be considered a trivial extension of the existing convex speedup analysis (cf. Lemma 3) because we do not rely on uniform gradient similarity here. □

As a small detour, we would like to briefly present a convex result here (although the focus of this work is non-convex analysis). We show that, for convex functions, following similar techniques used in our Theorem 1, we can strictly improve the previous results of Lemma 3.

Theorem 2. *Under Assumption 1, if all the local functions f_i are convex, $\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, and there exists some $D \ge 0$ such that $\left\|\bar{\mathbf{x}}_0-\mathbf{x}^*\right\|_2 \le D$, then there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on* $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right]-f^*$*:*

$$\mathcal{O}\left(\frac{LD^2}{\tau R}+\frac{\sigma D}{\sqrt{n\tau R}}+\left(\frac{L\zeta^2D^4}{R^2}\right)^{\frac{1}{3}}+\left(\frac{L\sigma^2D^4}{\tau R^2}\right)^{\frac{1}{3}}\right). \tag{17}$$

Remark 9. As compared to the previous speedup result (13), the heterogeneity term in (17) replaces the dependence on uniform gradient similarity with *standard* gradient similarity, which is always tighter, and can be substantially tighter up to a factor of $n^{\frac{1}{3}}$ with imbalanced heterogeneity. Moreover, Theorem 2 shows that standard gradient similarity $\zeta^2=\mathcal{O}(1/R)$ is sufficient to achieve speedup for LocalSGD. In contrast, the stronger condition of uniform gradient similarity $\bar{\zeta}^2=\mathcal{O}(1/R)$ in Woodworth et al. (2020b); Patel et al. (2024) is *redundant.* □

Novelty. The key technique in our proof of Theorems 1 and 2 is a "variance trick". A crucial term in the analysis is the variance between the workers, *i.e.*,

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_t^i-\bar{\mathbf{x}}_t\right\|_2^2 := \varXi_t, \text{ where } t=r\tau+k,\ k\in[\tau-1].$$

Essentially, many prior works either upper bound $\varXi_t$ by $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_t^i-\bar{\mathbf{x}}_{r\tau}\right\|_2^2$ (Koloskova et al., 2020; Karimireddy et al., 2020), which ignores the worker similarity; or upper bound $\varXi_t$ by $\sup_{i,j}\left\|\mathbf{x}_t^i-\mathbf{x}_t^j\right\|_2^2$ (Woodworth et al., 2020b; Patel et al., 2024), which loses the symmetry. We instead upper bound $\varXi_t$ by $\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_t^i-\bar{\mathbf{x}}_{t-1}+\eta\nabla f(\bar{\mathbf{x}}_{t-1})\right\|_2^2$ and then unroll the recursion. This small trick simplifies the analysis and leads to a strict improvement over previous analyses. □

Theorem 3. *Under Assumptions 1, 2+ and 4, there exists $\eta>0$ such that* LocalSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2$:*

$$\mathcal{O}\Bigg(\frac{L\varDelta}{R}+\sqrt{\frac{L\varDelta\sigma^2}{n\tau R}}+\left(\frac{\bar{\delta}\varDelta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(L\varDelta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}+\left(\frac{\mathcal{M}^2\varDelta^4\zeta^4}{R^4}\right)^{\frac{1}{5}}\Bigg). \tag{18}$$

Remark 10. As compared to the previous non-convex analysis (Lemma 2), Theorem 3 shows that, under the assumptions of uniform Hessian similarity and Lipschitz continuous Hessian, it is possible to improve the conditioning in the heterogeneity term. As a simplification of (18), for the case $L=\varDelta=\sigma=1$ and $\tau=\infty$, the asymptotic rate of LocalSGD is

$$\mathcal{O}\left(\frac{1}{R}+\left(\frac{\bar{\delta}\zeta}{R}\right)^{2/3}+\left(\frac{\mathcal{M}^2\zeta^4}{R^4}\right)^{1/5}\right).$$

We see that the above rate of LocalSGD is still no faster than that of MbSGD, but the conditioning required to match MbSGD has been relaxed. Asymptotically, this rate of (18) can match that of MbSGD under the conditioning of $\bar{\delta}^2\zeta^2+\mathcal{M}^2\zeta^4=\mathcal{O}\left(1/R\right)$. For $\mathcal{M}^2=\mathcal{O}\left(R\right)$, this is weaker than the conditioning of $\zeta^2=\mathcal{O}(1/R)$ in the classic analysis (*cf.* Remark 5).

We also see that when $\zeta^2=\varOmega\left(1/R\right)$, the heterogeneity term dominates the asymptotic rate in the classic analysis (*cf.* Remark 5). Meanwhile, if $\mathcal{M}^3\zeta=\mathcal{O}\left(R\right)$, with the new analysis, our rate in (18) demonstrates a strict improvement over the previous analysis. This is a perhaps surprising result showing that the local gradient steps can benefit from higher-order conditions, even without variance reduction. □

Novelty. We outline the proof techniques for Theorem 3. At each iteration t, the desired step is $\nabla f(\bar{\mathbf{x}}_t)$, but the actual step is approximately $\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_t^i)$. Thus, a crucial part of the analysis is bounding the discrepancy between these two, *i.e.*,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_t^i)-\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 := Q_t.$$

Previous approaches, such as Koloskova et al. (2020); Karimireddy et al. (2020), simply upper bound Q_t by $L^2\varXi_t$. Instead, under Assumptions 2+ and 4, we derive a tighter upper bound of $8\bar{\delta}^2\varXi_t+\frac{\mathcal{M}^2}{2}{\varXi_t}^2$ on Q_t. This technique alone does not suffice due to the stochastic noise in ${\varXi_t}^2$. To this end, we combine it with the construction of a "noiseless sequence", which helps to decouple the stochasticity from the gradient discrepancy, and the desired result finally follows. □

5.2 New Analysis of SCAFFOLD

Currently, the only speedup analysis of SCAFFOLD requires all f_i functions to be quadratic. In this section, we show the speedup of SCAFFOLD, while relaxing the "all-quadratic" restriction.

Theorem 4. *Under Assumptions 2 and 3, there exists $\eta>0$ such that* SCAFFOLD *ensures the following upper bound on $\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{2r\tau+\tau+k})\right\|_2^2$:*

$$\mathcal{O}\left(\left(\frac{L}{\tau}+\sqrt{L\delta}+\rho\right)\frac{\varDelta}{R}+\sqrt{\frac{L\varDelta\sigma^2}{n\tau R}}+\frac{(L\varDelta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}\right). \tag{19}$$

Remark 11. As compared to the rate of MbSGD (Lemma 1) and the previous rate of SCAFFOLD (Lemma 4), our new rate of SCAFFOLD in Theorem 4 has a faster optimization term of $\mathcal{O}\left((\frac{L}{\tau}+\sqrt{L\delta}+\rho)\frac{\varDelta}{R}\right)$. As a simplification of (19), for the case $L=\varDelta=\sigma=1$ and $\tau=\infty$, the asymptotic rate of SCAFFOLD is $\mathcal{O}\left(\frac{\sqrt{\delta}+\rho}{R}\right)$, which is faster than that of MbSGD. □

Novelty. This result extends the speedup analysis of SCAFFOLD beyond quadratic cases, addressing a challenging open problem stated in Karimireddy et al. (2020). We also weaken the uniform Hessian similarity required in prior work to the standard one. □

Finally, for theoretical purposes, we show that, under Assumption 4 with $\mathcal{M}=0$, we can get more benefit from Hessian similarity, thus improving the optimization term from $\mathcal{O}\left((\frac{L}{\tau}+\sqrt{L\delta}+\rho)\frac{\varDelta}{R}\right)$ to $\mathcal{O}\left((\frac{L}{\tau}+\sqrt{\bar{\delta}\delta}+\rho)\frac{\varDelta}{R}\right)$.

Theorem 5. *Under Assumptions 2, 2+, 3 and 4 with $\mathcal{M}=0$, there exists $\eta>0$ s.t.* SCAFFOLD *ensures the following upper bound on*

$\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{2r\tau+\tau+k})\right\|_2^2$:

$$\mathcal{O}\left(\left(\frac{L}{\tau}+\sqrt{\bar{\delta}\delta}+\rho\right)\frac{\Delta}{R}+\sqrt{\frac{L\Delta\sigma^2}{n\tau R}}+\frac{(\bar{\delta}\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}\right).$$

6 Synthetic Experiments

Experiment setup. We consider a distributed setting with n workers, where the i-th worker holds a dataset $(\mathbf{A}_i, \mathbf{y}_i) \in \mathbb{R}^{m_i \times d} \times \mathbb{R}^{m_i}$, for each $i \in [n]$. The total number of data points across all workers is denoted by $m = \sum_{i=1}^{n} m_i$. We focus on solving the following regression problem with smoothed Huber loss and non-convex regularizer:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathcal{L}(\mathbf{x};\mathbf{A}_i,\mathbf{y}_i)+\lambda\cdot\sum_{l=1}^{d}\frac{\mathbf{x}_l^2}{1+\mathbf{x}_l^2}\right], \tag{20}$$

where $\mathcal{L}(\mathbf{x};\mathbf{A}_i,\mathbf{y}_i) = \frac{n}{m}\cdot\sum_{j=1}^{m_i} h\left(\mathbf{A}_i(j)\cdot\mathbf{x}-\mathbf{y}_i(j)\right)$, in which $h(\cdot)$ is defined as follows:

$$h(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le 1, \\ -\frac{1}{6}(|u|-1)^3+\frac{1}{2}u^2, & 1 < |u| \le 2, \\ \frac{3}{2}|u|-\frac{7}{6}, & |u| > 2. \end{cases}$$

In all experiments, we set the dimension $d = 100$ and the number of workers $n = 10$. We fix the regularization parameter $\lambda = 0.01$ and set the following conditionings: $L \approx 1$, $\Delta \approx 1$ and $\sigma \approx 0.01$. We tune the stepsizes over $\{0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3\}$.

In this section, we present the experiments that validate the dependence on Hessian similarity δ. Further details and additional experiments on the other parameters can be found in Appendix B.

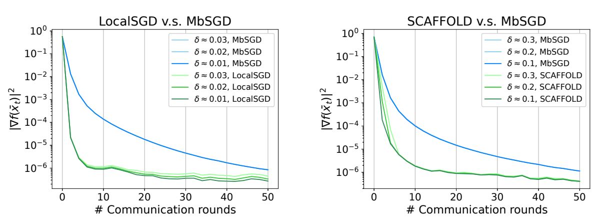

Figure 1: Comparisons between the convergence rates.

LocalSGD v.s. MbSGD. The theoretical results in Theorems 1 and 3 imply that LocalSGD can achieve a speedup over MbSGD for weakly convex functions and benefit from higher order conditions. To validate these results, we generated data such that $\zeta \approx 0.03$ and $\delta \approx 0.01, 0.02, 0.03$, and set the communication interval $\tau = 50$. As shown in Figure 1 (left), MbSGD converges at the same rate despite variations in δ. In contrast, LocalSGD converges faster as δ decreases. The most pronounced differences between the LocalSGD runs are observed in later communication rounds (≥ 10), where the heterogeneity term dominates the rate.

SCAFFOLD v.s. MbSGD. Theorem 4 shows the speedup of SCAFFOLD from Hessian similarity and weak convexity. To validate this result, we generated data such that $\delta \approx 0.1, 0.2, 0.3$, and set the communication interval $\tau = 50$. As shown in Figure 1 (right), MbSGD converges at the same rate despite variations in δ. In contrast, SCAFFOLD converges faster as δ decreases. The most pronounced difference between the SCAFFOLD runs is observed in early communication rounds (≤ 10), where the optimization term dominates the rate.

7 Limitations and Future Work

This paper revisits the convergence rates of LocalSGD and SCAFFOLD under gradient similarity, Hessian similarity, weak convexity, and Lipschitz Hessian. Our extensive studies demonstrate improved rates for both algorithms across various settings, suggesting that our new analysis can better capture their effectiveness as observed in real-world applications.

While our analysis focuses on the standard algorithms in the intermittent communication setting, many orthogonal techniques have been applied to the framework to further reduce communication cost. These include the use of previous updates for variance reduction (Karimireddy et al., 2020), quantization and sparsification of the updates (Alistarh et al., 2017; Stich et al., 2018), decentralized communication network (Lian et al., 2017; Sadiev et al., 2022), partial participation (McMahan et al., 2017), and asynchronous communication (Recht et al., 2011; Nguyen et al., 2018). We do not cover these techniques in our analysis and leave them as future work.

Acknowledgements

The authors thank for the helpful discussions with Eduard Gorbunov, Kumar Kshitij Patel, Anton Rodomanov, and Ali Zindari during the preparation of this work.

This work was partially done during the first author's stays at CISPA and at MBZUAI. The first author also acknowledges ERC CoG 863818 (ForM-SMArt) and Austrian Science Fund (FWF) 10.55776/COE12.

References

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30.

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Kwing, H. L., Parcollet, T., Gusmão, P. P. d., and Lane, N. D. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer.

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

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. (2012). Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1).

Gu, X., Lyu, K., Huang, L., and Arora, S. (2023). Why (and when) does local SGD generalize better than SGD? In *The Eleventh International Conference on Learning Representations*.

He, C., Li, S., So, J., Zeng, X., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H., et al. (2020). Fedml: A research library and benchmark for federated machine learning. *arXiv preprint arXiv:2007.13518*.

Hendrikx, H., Xiao, L., Bubeck, S., Bach, F., and Massoulie, L. (2020). Statistically preconditioned accelerated gradient method for distributed optimization. In *International conference on machine learning*, pages 4203–4227. PMLR.

Jaggi, M., Smith, V., Takác, M., Terhorst, J., Krishnan, S., Hofmann, T., and Jordan, M. I. (2014). Communication-efficient distributed dual coordinate ascent. *Advances in neural information processing systems*, 27.

Jiang, X., Rodomanov, A., and Stich, S. U. (2024). Federated optimization with doubly regularized drift correction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21912–21945. PMLR.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S., Stich, S. U., and Suresh, A. T. (2021). Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676.

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*, pages 5132–5143. PMLR.

Khaled, A. and Jin, C. (2023). Faster federated optimization under second-order similarity. In *The Eleventh International Conference on Learning Representations*.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. (2020). A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30.

Magnanti, T. L. (2021). Optimization: From its inception. *Management Science*, 67(9):5349–5363.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtarik, P. (2022). ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15750–15769. PMLR.

Murata, T. and Suzuki, T. (2021). Bias-variance reduced local sgd for less heterogeneous federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7872–7881. PMLR.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Nguyen, L. M., Nguyen, P. H., van Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. (2018). Sgd and hogwild! convergence without the bounded gradients assumption. In *In 34th International Conference on Machine Learning, ICML 2018*.

Patel, K. K., Glasgow, M., Wang, L., Joshi, N., and Srebro, N. (2023). On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*.

Patel, K. K., Glasgow, M., Zindari, A., Wang, L., Stich, S. U., Cheng, Z., Joshi, N., and Srebro, N. (2024). The limits and potentials of local sgd for distributed heterogeneous learning with intermittent communication. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4115–4157. PMLR.

Patel, K. K., Wang, L., Woodworth, B., Bullins, B., and Srebro, N. (2022). Towards optimal communication complexity in distributed non-convex optimization. In *Advances in Neural Information Processing Systems*.

Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407.

Roth, H. R., Cheng, Y., Wen, Y., Yang, I., Xu, Z., Hsieh, Y.-T., Kersten, K., Harouni, A., Zhao, C., Lu, K., et al. (2022). Nvidia flare: Federated learning from simulation to real-world. *arXiv preprint arXiv:2210.13291*.

Sadiev, A., Borodich, E., Beznosikov, A., Dvinskikh, D., Chezhegov, S., Tappenden, R., Takáč, M., and Gasnikov, A. (2022). Decentralized personalized federated learning: Lower bounds and optimal algorithm for all personalization modes. *EURO Journal on Computational Optimization*, 10:100041.

Sefidgaran, M., Chor, R., Zaidi, A., and Wan, Y. (2024). Lessons from generalization error analysis of federated learning: You may communicate less often! In *Forty-first International Conference on Machine Learning*.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1000–II–1008. JMLR.org.

Shen, S., Cheng, Y., Liu, J., and Xu, L. (2021). Stl-sgd: Speeding up local sgd with stagewise communication period. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9576–9584.

So, A. M.-C., Jain, P., Ma, W.-K., and Scutari, G. (2020). Nonconvex optimization for signal processing and machine learning. *IEEE Signal Processing Magazine*, 37(5):15–17.

Stich, S. U. (2019). Local SGD converges fast and communicates little. In *International Conference on Learning Representations*.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31.

Sun, Z., Niu, X., and Wei, E. (2024). Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR.

Takáč, M., Bijral, A., Richtárik, P., and Srebro, N. (2013). Mini-batch primal and dual methods for svms. In *In 30th International Conference on Machine Learning, ICML 2013*.

Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., and Zhang, T. (2024). On the unreasonable effectiveness of federated averaging with heterogeneous data. *Transactions on Machine Learning Research*.

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. (2020a). Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR.

Woodworth, B. E., Patel, K. K., and Srebro, N. (2020b). Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700.

Zhou, F. and Cong, G. (2018). On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3219–3227. International Joint Conferences on Artificial Intelligence Organization.

Zindari, A., Luo, R., and Stich, S. U. (2023). On the convergence of local SGD under third-order smoothness and hessian similarity. In *NeurIPS Workshop OPT 2023: Optimization for Machine Learning*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials for Revisiting LocalSGD and SCAFFOLD

A Proof Details

A.1 Technical Lemmas

In this section, we present several technical lemmas and remark on their usages in our analysis.

For simplicity, we denote $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}^i$ and $\Xi = \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}^i - \bar{\mathbf{x}}\right\|_2^2$. We assume all the vectors are in $\mathbb{R}^d$.

Lemma 6 (variance trick). *For any* $\mathbf{y} \in \mathbb{R}^d$,

$$\Xi \leq \sum_{i=1}^n \left\|\mathbf{x}^i - \mathbf{y}\right\|_2^2. \tag{21}$$

Proof. Since $\sum_{i=1}^n \left\|\mathbf{x}^i - \mathbf{y}\right\|_2^2$ is strongly convex in $\mathbf{y}$, let $\mathbf{y}^* = \arg\min_{\mathbf{y}\in\mathbb{R}^d} \sum_{i=1}^n \left\|\mathbf{x}^i - \mathbf{y}\right\|_2^2$. Then,

$$\left.\frac{d\left(\sum_{i=1}^n \left\|\mathbf{x}^i - \mathbf{y}\right\|_2^2\right)}{d\mathbf{y}}\right|_{\mathbf{y}^*} = 2(\mathbf{y} - \bar{\mathbf{x}})\big|_{\mathbf{y}^*} = \mathbf{0},$$

and we have $\mathbf{y}^* = \bar{\mathbf{x}}$. □

Remark 12. We will use this variance trick frequently in our analysis. For instance, when upper bounding Ξ_{t+1} recursively, we can use the following inequality:

$$\Xi_{t+1} \leq \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}_{t+1}^i - \bar{\mathbf{x}}_t + \eta\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2.$$

□

Lemma 7. *Under Assumptions 2+ and 4, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i) - \nabla f(\bar{\mathbf{x}})\right\|_2^2 \leq 8\bar{\delta}^2\Xi + \frac{\mathcal{M}^2}{2}\Xi^2. \tag{22}$$

Proof. Let $\hat{f} \in \mathbf{conv}\{f_1, \cdots, f_n\}$ *s.t.* $\hat{f}$ has $\mathcal{M}$-Lipschitz continuous Hessian. We have

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i) - \nabla f(\bar{\mathbf{x}})\right\|_2^2 \\
&= \left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}^i) - \nabla f_i(\bar{\mathbf{x}})\right)\right\|_2^2 \\
&\leq 2\left(\left\|\frac{1}{n}\sum_{i=1}^n \nabla \hat{f}(\mathbf{x}^i) - \nabla \hat{f}(\bar{\mathbf{x}})\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(\mathbf{x}^i) - \nabla \hat{f}(\mathbf{x}^i) - \nabla f_i(\bar{\mathbf{x}}) + \nabla \hat{f}(\bar{\mathbf{x}})\right)\right\|_2^2\right).
\end{aligned} \tag{23}$$

We then upper bound the above two terms respectively:

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla\hat{f}(\mathbf{x}^i)-\nabla\hat{f}(\bar{\mathbf{x}})\right\|_2^2 \\
=&\left\|\frac{1}{n}\sum_{i=1}^{n}\int_0^1\nabla^2\hat{f}\left(\bar{\mathbf{x}}+u(\mathbf{x}^i-\bar{\mathbf{x}})\right)(\mathbf{x}^i-\bar{\mathbf{x}})du\right\|_2^2 \\
=&\left\|\frac{1}{n}\sum_{i=1}^{n}\int_0^1\left[\nabla^2\hat{f}\left(\bar{\mathbf{x}}+u(\mathbf{x}^i-\bar{\mathbf{x}})\right)-\nabla^2\hat{f}(\bar{\mathbf{x}})\right](\mathbf{x}^i-\bar{\mathbf{x}})du\right\|_2^2 \\
\le&\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^1\left\|\nabla^2\hat{f}\left(\bar{\mathbf{x}}+u(\mathbf{x}^i-\bar{\mathbf{x}})\right)-\nabla^2\hat{f}(\bar{\mathbf{x}})\right\|_2\cdot\left\|\mathbf{x}^i-\bar{\mathbf{x}}\right\|_2du\right]^2 \\
\overset{(10)}{\le}&\left[\frac{1}{n}\sum_{i=1}^{n}\int_0^1\mathcal{M}u\left\|\mathbf{x}^i-\bar{\mathbf{x}}\right\|_2^2du\right]^2 \\
=&\frac{\mathcal{M}^2}{4}\varXi^2,
\end{aligned}
\tag{24}
$$

and

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla f_i(\mathbf{x}^i)-\nabla\hat{f}(\mathbf{x}^i)-\nabla f_i(\bar{\mathbf{x}})+\nabla\hat{f}(\bar{\mathbf{x}})\right)\right\|_2^2 \\
\le&\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}^i)-\nabla\hat{f}(\mathbf{x}^i)-\nabla f_i(\bar{\mathbf{x}})+\nabla\hat{f}(\bar{\mathbf{x}})\right\|_2^2 \\
\le&\frac{1}{n}\sum_{i=1}^{n}\left(2\bar{\delta}\left\|\mathbf{x}^i-\bar{\mathbf{x}}\right\|_2\right)^2 \\
=&4\bar{\delta}^2\varXi.
\end{aligned}
\tag{25}
$$

Finally, (22) follows from plugging (24) and (25) into (23). □

Remark 13. Lemma 7 is the key to the benefit from higher order conditions in Theorem 3 and Theorem 5. Instead, without higher order conditions, we can only upper bound the term as follows:

$$
\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}^i)-\nabla f(\bar{\mathbf{x}})\right\|_2^2\le\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}^i)-\nabla f_i(\bar{\mathbf{x}})\right\|_2^2\le L^2\varXi. \tag{26}
$$

□

Lemma 8. *Under Assumption 3, we have*

$$
(L-\rho)\left\langle\mathbf{x}-\mathbf{y},\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\right\rangle\ge\left\|\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\right\|_2^2-\rho L\left\|\mathbf{x}-\mathbf{y}\right\|_2^2. \tag{27}
$$

Proof. $f_i(\mathbf{x})+\frac{\rho}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}$ is $(L+\rho)$-smooth and convex. By (2.1.8) in Nesterov (2003), we have

$$
\frac{1}{L+\rho}\left\|\nabla f_i(\mathbf{x})+\rho\mathbf{x}-\nabla f_i(\mathbf{y})-\rho\mathbf{y}\right\|_2^2\le\left\langle\nabla f_i(\mathbf{x})+\rho\mathbf{x}-\nabla f_i(\mathbf{y})-\rho\mathbf{y},\mathbf{x}-\mathbf{y}\right\rangle.
$$

Then, after multiplying both sides by $(L+\rho)$, a rearrangement will yield (27).

□

Lemma 9. *Under Assumption 3, for $\eta\le\frac{2}{L-\rho}$, we have*

$$
\left\|(\mathbf{x}-\eta\nabla f_i(\mathbf{x}))-(\mathbf{y}-\eta\nabla f_i(\mathbf{y}))\right\|_2^2\le\left(1+\frac{2L}{L-\rho}\cdot\rho\eta\right)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2. \tag{28}
$$

Proof. For $\eta \leq \frac{2}{L-\rho}$,

$$\begin{aligned}
&\left\|\mathbf{x}-\eta\nabla f_i(\mathbf{x})-(\mathbf{y}-\eta\nabla f_i(\mathbf{y}))\right\|_2^2 \\
&= \|\mathbf{x}-\mathbf{y}\|_2^2 - 2\eta\left\langle \mathbf{x}-\mathbf{y}, \nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\right\rangle + \eta^2\left\|\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\right\|_2^2 \\
&\overset{(27)}{\leq} \left(1+\frac{2L}{L-\rho}\cdot\rho\eta\right)\|\mathbf{x}-\mathbf{y}\|_2^2 - \left(\frac{2\eta}{L-\rho}-\eta^2\right)\left\|\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\right\|_2^2 \\
&\leq \left(1+\frac{2L}{L-\rho}\cdot\rho\eta\right)\|\mathbf{x}-\mathbf{y}\|_2^2 .
\end{aligned}$$

$\square$

Remark 14. Lemma 9 is the key to prove the speedup from weak convexity. Instead, without weak convexity, we can only upper bound the term as follows:

$$\begin{aligned}
\left\|\mathbf{x}-\eta\nabla f_i(\mathbf{x})-(\mathbf{y}-\eta\nabla f_i(\mathbf{y}))\right\|_2^2 &\leq \left(\|\mathbf{x}-\mathbf{y}\|_2+\eta\left\|\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\right\|_2\right)^2 \\
&\leq (1+L\eta)^2\|\mathbf{x}-\mathbf{y}\|_2^2 .
\end{aligned} \tag{29}$$

This upper bound in Equation (29) is looser compared to the one in Equation (28). $\square$

A.2 The Analysis of LocalSGD

We recall the following notations: for $t = r\tau + k$, $r \in [0, R-1]$, $k \in [0, \tau-1]$, we denote

$$r(t) = r\tau, \quad \mathit{\Xi}_t = \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t\right\|_2^2, \quad Q_t = \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 .$$

Define the following *noiseless sequence*: for $t \in [0, T-1]$,

$$\hat{\mathbf{x}}_t^i = \begin{cases} \bar{\mathbf{x}}_t, & \text{if } t \text{ is a multiple of } \tau, \\ \hat{\mathbf{x}}_{t-1}^i - \eta\nabla f_i(\hat{\mathbf{x}}_{t-1}^i), & \text{otherwise,} \end{cases}$$

and

$$\hat{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \hat{\mathbf{x}}_t^i, \quad \hat{Q}_t = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\hat{\mathbf{x}}_t^i) - \nabla f(\hat{\mathbf{x}}_t)\right\|_2^2, \quad \hat{\mathit{\Xi}}_t = \frac{1}{n}\sum_{i=1}^n \left\|\hat{\mathbf{x}}_t^i - \hat{\mathbf{x}}_t\right\|_2^2 .$$

Lemma 10. *For $\eta \leq \frac{1}{L}$,* LocalSGD *ensures*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \leq \frac{2\Delta}{\eta T} + \eta\frac{L\sigma^2}{n} + \frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\left(6L^2\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\left\|\mathbf{x}_{r\tau+k}^i - \hat{\mathbf{x}}_{r\tau+k}^i\right\|_2^2\right] + 3\,\mathbb{E}\left[\hat{Q}_{r\tau+k}\right]\right). \tag{30}$$

Proof. For $\eta \leq \frac{1}{L}$,

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{\mathbf{x}}_{t+1})\right] \\
&\overset{(2)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \eta\,\mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}_t), \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t^i)\right\rangle + \eta^2\frac{L}{2}\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t^i)\right\|_2^2 + \eta^2\frac{L\sigma^2}{2n} \\
&= \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \frac{\eta}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 - \left(\frac{\eta}{2}-\eta^2\frac{L}{2}\right)\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_t^i)\right\|_2^2 + \eta^2\frac{L\sigma^2}{2n} + \frac{\eta}{2}\mathbb{E}\left[Q_t\right] \\
&\leq \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \frac{\eta}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 + \eta^2\frac{L\sigma^2}{2n} + \frac{\eta}{2}\mathbb{E}\left[Q_t\right].
\end{aligned}$$

Summing over t from 0 to $T-1$ and multiplying both sides by $\frac{2}{\eta T}$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 &\leq \frac{2}{\eta T}\mathbb{E}\left[f(\bar{\mathbf{x}}_0)-f(\bar{\mathbf{x}}_T)\right]+\eta\frac{L\sigma^2}{n}+\frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{r\tau+k}\right] \\ &\leq \frac{2\Delta}{\eta T}+\eta\frac{L\sigma^2}{n}+\frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{r\tau+k}\right].\end{aligned}$$

Finally, Equation (30) follows from the following inequality:

$$\begin{aligned}\mathbb{E}\left[Q_t\right] &\leq 3\,\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left[\nabla f_i(\mathbf{x}_t^i)-\nabla f_i(\hat{\mathbf{x}}_t^i)\right]\right\|_2^2\right]+3\,\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}_t)-\nabla f(\hat{\mathbf{x}}_t)\right\|_2^2\right]+3\,\mathbb{E}\left[\hat{Q}_t\right] \\ &\leq \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\nabla f_i(\mathbf{x}_t^i)-\nabla f_i(\hat{\mathbf{x}}_t^i)\right\|_2^2\right]+3\,\mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}_t)-\nabla f(\hat{\mathbf{x}}_t)\right\|_2^2\right]+3\,\mathbb{E}\left[\hat{Q}_t\right] \\ &\leq 3L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2\right]+3L^2\left\|\bar{\mathbf{x}}_t-\hat{\mathbf{x}}_t\right\|_2^2+3\,\mathbb{E}\left[\hat{Q}_t\right] \\ &\leq 6L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2\right]+3\,\mathbb{E}\left[\hat{Q}_t\right].\end{aligned}$$

□

Lemma 11 (variance trick)**.** *For any $\gamma>0$, when $(t+1)$ is not a multiple of τ, we have*

$$\hat{\Xi}_{t+1}\leq\frac{1}{n}\sum_{i=1}^{n}(1+\gamma)\left\|\hat{\mathbf{x}}_t^i-\eta\nabla f_i(\hat{\mathbf{x}}_t^i)-(\hat{\mathbf{x}}_t-\eta\nabla f_i(\hat{\mathbf{x}}_t))\right\|_2^2+(1+\gamma^{-1})\eta^2\zeta^2. \tag{31}$$

Proof. For any $\gamma>0$,

$$\begin{aligned}&\hat{\Xi}_{t+1} \\ &\overset{(21)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{x}}_{t+1}^i-(\hat{\mathbf{x}}_t-\eta\nabla f(\hat{\mathbf{x}}_t))\right\|_2^2 \\ &=\frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{x}}_t^i-\eta\nabla f_i(\hat{\mathbf{x}}_t^i)-(\hat{\mathbf{x}}_t-\eta\nabla f(\hat{\mathbf{x}}_t))\right\|_2^2 \\ &\overset{(6)}{\leq}\frac{1}{n}\sum_{i=1}^{n}(1+\gamma)\left\|\hat{\mathbf{x}}_t^i-\eta\nabla f_i(\hat{\mathbf{x}}_t^i)-(\hat{\mathbf{x}}_t-\eta\nabla f_i(\hat{\mathbf{x}}_t))\right\|_2^2+(1+\gamma^{-1})\eta^2\zeta^2.\end{aligned} \tag{32}$$

□

A.2.1 Proof of Theorem 1

Lemma 12. *Under Assumption 3, for $\eta\leq\min\left\{\frac{2}{L-\rho},\frac{1-\rho/L}{4\rho(\tau-1)}\right\}$,* LocalSGD *ensures*

$$\mathbb{E}\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2\leq 3k\eta^2\sigma^2, \tag{33}$$

where $k=t-r(t)$.

The proof of the above lemma is similar to the proof of Lemma 16. Lemma 9 is used in this proof when bounding the distance recursively.

Lemma 13. *Under Assumptions 1 and 3, for $\eta\leq\min\left\{\frac{2}{L-\rho},\frac{1-\rho/L}{6\rho(\tau-1)}\right\}$,* LocalSGD *ensures*

$$\hat{\Xi}_t\leq 9k(\tau-1)\eta^2\zeta^2, \tag{34}$$

where $k=t-r(t)$.

The proof of the above lemma is similar to the proof of Lemma 17. Lemma 9 is used in this proof when bounding the distance recursively.

Now we are ready to prove Theorem 1.

Proof of Theorem 1. For $\eta \leq \min\left\{\frac{1}{L}, \frac{1-\rho/L}{6\rho(\tau-1)}\right\}$,

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \\
&\overset{(30)}{\leq} \frac{2\Delta}{\eta T} + \eta\frac{L\sigma^2}{n} + \frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\left(6L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{x}_{r\tau+k}^i - \hat{\mathbf{x}}_{r\tau+k}^i\right\|_2^2\right] + 3\,\mathbb{E}\left[\hat{Q}_{r\tau+k}\right]\right) \\
&\overset{(33)(26)}{\leq} \frac{2\Delta}{\eta T} + \eta\frac{L\sigma^2}{n} + 9(\tau-1)L^2\eta^2\sigma^2 + \frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}3L^2\,\mathbb{E}\left[\hat{\Xi}_{r\tau+k}\right] \\
&\overset{(34)}{\leq} \frac{2\Delta}{\eta T} + \eta\frac{L\sigma^2}{n} + 9(\tau-1)L^2\eta^2\sigma^2 + \frac{27}{2}(\tau-1)^2L^2\eta^2\zeta^2.
\end{aligned}
$$

Finally, (18) follows from the following assignment

$$
\eta := \min\left\{\frac{1}{L}, \frac{1-\rho/L}{6\rho(\tau-1)}, \sqrt{\frac{2\Delta n}{L\sigma^2 T}}, \left(\frac{4\Delta}{27L^2(\tau-1)^2\zeta^2 T}\right)^{\frac{1}{3}}, \left(\frac{2\Delta}{9L^2(\tau-1)\sigma^2 T}\right)^{\frac{1}{3}}\right\}.
$$

□

A.2.2 Proof of Theorem 2

Lemma 14 (*cf.* Lemma 7 in Woodworth et al. (2020b))**.** *If all f_i's are convex, for $\eta \leq \frac{1}{2L}$,* LocalSGD *ensures*

$$
\mathbb{E}\left[f(\bar{\mathbf{x}}_t) - f^*\right] \leq \frac{1}{\eta}\mathbb{E}\left\|\bar{\mathbf{x}}_t - \mathbf{x}^*\right\|_2^2 - \frac{1}{\eta}\mathbb{E}\left\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}^*\right\|_2^2 + 3\frac{\eta\sigma^2}{n} + 2L\,\mathbb{E}\left[\Xi_t\right]. \tag{35}
$$

Lemma 15. *Under Assumptions 1 and 3, if all f_i's are convex, for $\eta \leq \frac{2}{L}$,* LocalSGD *ensures*

$$
\mathbb{E}\left[\Xi_t\right] \leq 3k \cdot \left[(\tau-1)\eta^2\zeta^2 + \eta^2\sigma^2\right], \tag{36}
$$

where $k = t - r(t)$.

Proof. Let $\gamma := \frac{1}{\tau-1}$. We prove

$$
\mathbb{E}\left[\Xi_t\right] \leq (1+\gamma)^k \cdot k \cdot \left[(\tau-1)\eta^2\zeta^2 + \eta^2\sigma^2\right] \tag{37}
$$

by induction on k.

1. (37) obviously holds for $k = 0$.
2. We assume (37) holds for k, where $k \leq \tau - 2$. Then for $\eta \leq \frac{2}{L}$,

$$
\begin{aligned}
&\mathbb{E}\left[\Xi_{t+1}\right] \\
&\overset{(21)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_t^i - \eta\mathbf{g}_t^i - (\bar{\mathbf{x}}_t - \eta\nabla f(\bar{\mathbf{x}}_t))\right\|_2^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_t^i - \eta\nabla f_i(\mathbf{x}_t^i) - (\bar{\mathbf{x}}_t - \eta\nabla f(\bar{\mathbf{x}}_t))\right\|_2^2 + \eta^2\sigma^2 \\
&\overset{(6)}{\leq} \frac{1}{n}\sum_{i=1}^{n}(1+\gamma)\,\mathbb{E}\left\|\mathbf{x}_t^i - \eta\nabla f_i(\mathbf{x}_t^i) - (\bar{\mathbf{x}}_t - \eta\nabla f_i(\bar{\mathbf{x}}_t))\right\|_2^2 + (1+\gamma^{-1})\eta^2\zeta^2 + \eta^2\sigma^2 \\
&\overset{(28)}{\leq} (1+\gamma)\,\mathbb{E}\left[\Xi_t\right] + (1+\gamma^{-1})\eta^2\zeta^2 + \eta^2\sigma^2,
\end{aligned} \tag{38}
$$

and by induction hypothesis, we have

$$\begin{aligned}
&\mathbb{E}\left[\Xi_{t+1}\right] \\
&\overset{(37)}{\le} (1+\gamma)^{k+1} \cdot k\left[(\tau-1)\eta^2\zeta^2+\eta^2\sigma^2\right]+(1+\gamma^{-1})\eta^2\zeta^2+\eta^2\sigma^2 \\
&= (1+\gamma)^{k+1} \cdot k\left[(\tau-1)\eta^2\zeta^2+\eta^2\sigma^2\right]+\tau\eta^2\zeta^2+\eta^2\sigma^2 \\
&\le (1+\gamma)^{k+1} \cdot (k+1)\left[(\tau-1)\eta^2\zeta^2+\eta^2\sigma^2\right].
\end{aligned}$$

Now, we have proven (37), and then (36) follows directly from the fact that

$$(1+\gamma)^k \le (1+\gamma)^{\tau-1} < e < 3.$$

□

Now we are ready to prove Theorem 2.

Proof of Theorem 2. For $\eta \le \frac{1}{2L}$, multiplying both sides of Equation (35) by $\frac{1}{T}$ and summing over t from 0 to $T-1$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right]-f^* &\le \frac{D^2}{\eta T}+3\eta\frac{\sigma^2}{n}+\frac{2L}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[\Xi_{r\tau+k}\right] \\
&\overset{(36)}{\le} \frac{D^2}{\eta T}+3\eta\frac{\sigma^2}{n}+3(\tau-1)^2L\eta^2\zeta^2+3(\tau-1)L\eta^2\sigma^2 .
\end{aligned}$$

Finally, (17) follows from the following assignment:

$$\eta := \min\left\{\frac{1}{2L}, \sqrt{\frac{nD^2}{3\sigma^2 T}}, \left(\frac{D^2}{3L(\tau-1)^2\zeta^2 T}\right)^{\frac{1}{3}}, \left(\frac{D^2}{3L(\tau-1)\sigma^2 T}\right)^{\frac{1}{3}}\right\}.$$

□

A.2.3 Proof of Theorem 3

Lemma 16. *For $\eta \le \frac{1}{2L(\tau-1)}$,* LocalSGD *ensures*

$$\mathbb{E}\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2 \le 3k\eta^2\sigma^2, \tag{39}$$

where $k = t - r(t)$.

Proof. Let $\gamma := \frac{1}{2(\tau-1)}$. We prove

$$\mathbb{E}\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2 \le (1+\gamma)^{2\cdot k}\cdot k\cdot\eta^2\sigma^2 \tag{40}$$

by induction on k.

1. (40) obviously holds for $k = 0$.
2. We assume (40) holds for k, where $k \le \tau-2$. Then, since $\eta \le \frac{1}{2L(\tau-1)}$,

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{t+1}^i-\hat{\mathbf{x}}_{t+1}^i\right\|_2^2 &\le \mathbb{E}\left\|\mathbf{x}_t^i-\eta\nabla f_i(\mathbf{x}_t^i)-\hat{\mathbf{x}}_t^i+\eta\nabla f_i(\hat{\mathbf{x}}_t^i)\right\|_2^2+\eta^2\sigma^2 \\
&\overset{(29)}{\le} (1+L\eta)^2\,\mathbb{E}\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2+\eta^2\sigma^2 \\
&\le (1+\gamma)^2\cdot\mathbb{E}\left\|\mathbf{x}_t^i-\hat{\mathbf{x}}_t^i\right\|_2^2+\eta^2\sigma^2,
\end{aligned} \tag{41}$$

and by induction hypothesis, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{x}_{t+1}^i-\hat{\mathbf{x}}_{t+1}^i\right\|_2^2 &\overset{(40)}{\le} (1+\gamma)^{2(k+1)}\cdot k\cdot\eta^2\sigma^2+\eta^2\sigma^2 \\
&\le (1+\gamma)^{2(k+1)}\cdot(k+1)\eta^2\sigma^2.
\end{aligned}$$

Now, we have proven (40), and then (39) follows directly from the fact that

$$(1+\gamma)^{2\cdot k} \le (1+\gamma)^{2(\tau-1)} < e < 3.$$

□

Lemma 17. *Under Assumption 1, for* $\eta \le \frac{1}{3L(\tau-1)}$, LocalSGD *ensures*

$$\hat{\varXi}_t \le 9k(\tau-1)\eta^2\zeta^2, \tag{42}$$

where $k = t - r(t)$.

Proof. Let $\gamma := \frac{1}{3(\tau-1)}$. We prove

$$\hat{\varXi}_t \le (1+\gamma)^{3\cdot k} \cdot k \cdot 3(\tau-1)\eta^2\zeta^2 \tag{43}$$

by induction on k.

1. (43) obviously holds for $k = 0$.
2. We assume (43) holds for k, where $k \le \tau - 2$. Then, since $\eta \le \frac{1}{3L(\tau-1)}$,

$$\begin{aligned}
\hat{\varXi}_{t+1} &\overset{(31)}{\le} \frac{1}{n}\sum_{i=1}^{n}(1+\gamma)\left\|\hat{\mathbf{x}}_t^i - \eta\nabla f_i(\hat{\mathbf{x}}_t^i) - (\hat{\mathbf{x}}_t - \eta\nabla f_i(\hat{\mathbf{x}}_t))\right\|_2^2 + (1+\gamma^{-1})\eta^2\zeta^2 \\
&\overset{(29)}{\le} (1+\gamma)\cdot(1+L\eta)^2\,\hat{\varXi}_t + (1+\gamma^{-1})\eta^2\zeta^2 \\
&\le (1+\gamma)^3\cdot\hat{\varXi}_t + (1+\gamma^{-1})\eta^2\zeta^2.
\end{aligned} \tag{44}$$

   and by induction hypothesis, we have

$$\begin{aligned}
\hat{\varXi}_{t+1} &\overset{(43)}{\le} (1+\gamma)^{3(k+1)}\cdot k\cdot 3(\tau-1)\eta^2\zeta^2 + (1+\gamma^{-1})\eta^2\zeta^2 \\
&= (1+\gamma)^{3(k+1)}\cdot k\cdot 3(\tau-1)\eta^2\zeta^2 + (3\tau-2)\eta^2\zeta^2 \\
&\le (1+\gamma)^{3(k+1)}\cdot(k+1)\cdot 3(\tau-1)\eta^2\zeta^2.
\end{aligned}$$

Now, we have proven (43), and then (42) follows directly from the fact that

$$(1+\gamma)^{3\cdot k} \le (1+\gamma)^{3(\tau-1)} < e < 3.$$

□

Now we are ready to prove Theorem 3.

Proof of Theorem 3. From Lemma 7, we have

$$\hat{Q}_t \le 8\bar{\delta}^2\hat{\varXi}_t + \frac{\mathcal{M}^2}{2}\hat{\varXi}_t^2. \tag{45}$$

Hence, for $\eta \le \frac{1}{3L(\tau-1)}$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \\
&\overset{(30)}{\le} \frac{2\varDelta}{\eta T} + \eta\frac{L\sigma^2}{n} + \frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\left(6L^2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left\|\mathbf{x}_{r\tau+k}^i - \hat{\mathbf{x}}_{r\tau+k}^i\right\|_2^2\right] + 3\,\mathbb{E}\left[\hat{Q}_{r\tau+k}\right]\right) \\
&\overset{(39)(45)}{\le} \frac{2\varDelta}{\eta T} + \eta\frac{L\sigma^2}{n} + 18(\tau-1)L^2\eta^2\sigma^2 + \frac{1}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\left(24\bar{\delta}^2\,\mathbb{E}\left[\hat{\varXi}_{r\tau+k}\right] + \frac{3}{2}\mathcal{M}^2\,\mathbb{E}\left[\hat{\varXi}_{r\tau+k}^2\right]\right) \\
&\overset{(42)}{\le} \frac{2\varDelta}{\eta T} + \eta\frac{L\sigma^2}{n} + 18(\tau-1)L^2\eta^2\sigma^2 + 108(\tau-1)^2\bar{\delta}^2\eta^2\zeta^2 + \frac{81}{4}\mathcal{M}^2(\tau-1)^3(2\tau-1)\eta^4\zeta^4.
\end{aligned}$$

Note that in the last inequality, we can upper bound $\mathbb{E}\left[\hat{\Xi}^2_{r\tau+k}\right]$ by $(9k(\tau-1)\eta^2\zeta^2)^2$, because the bound in (42) is deterministic. (This is why it is necessary to construct the noiseless sequence in our proof. Otherwise, we cannot upper bound $\mathbb{E}\left[\Xi^2_{r\tau+k}\right]$ likewise.)

Finally, (18) follows from the following assignment

$$\eta := \min\left\{\frac{1}{L}, \frac{1}{3L(\tau-1)}, \sqrt{\frac{2\Delta n}{L\sigma^2 T}}, \left(\frac{\Delta}{54\bar{\delta}^2(\tau-1)^2\zeta^2 T}\right)^{\frac{1}{3}}, \left(\frac{\Delta}{9L^2(\tau-1)\sigma^2 T}\right)^{\frac{1}{3}}, \left(\frac{8\Delta}{81\mathcal{M}^2(\tau-1)^3(2\tau-1)\zeta^4 T}\right)^{\frac{1}{5}}\right\}.$$

□

A.3 The Analysis of SCAFFOLD

In section A.3, to simplify the notations, for $t = r\tau + k$, $r \in [0, R-1]$, $k \in [0, \tau-1]$, we denote

$$r(t) = r\tau, \quad \mathbf{x}^i_{(t)} = \mathbf{x}^i_{2r\tau+\tau+k}, \quad \mathbf{g}^i_{(t)} = \mathbf{g}^i_{2r\tau+\tau+k},$$

and

$$\bar{\mathbf{x}}^i_{(t)} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}^i_{(t)},$$
$$\Xi_{(t)} = \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}^i_{(t)} - \bar{\mathbf{x}}_{(t)}\right\|_2^2,$$
$$Q_{(t)} = \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t)}) - \nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2.$$

Also, we denote $\bar{\mathbf{x}}_{(R\tau)} = \bar{\mathbf{x}}_T$.

Lemma 18. *For $\eta \le \frac{1}{2L}$,* SCAFFOLD *ensures*

$$\begin{aligned}
&\frac{2}{T}\sum_{t=0}^{R\tau-1}\left[\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t)})\right\|_2^2\right] \\
&\le \frac{4\Delta}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{(r\tau+k)}\right].
\end{aligned} \tag{46}$$

Proof. For $t = r\tau + k$, $r \in [0, R-1]$, $k \in [0, \tau-1]$, for $\eta \le \frac{1}{2L}$, we have

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{\mathbf{x}}_{(t+1)})\right] \\
&\le \mathbb{E}\left[f(\bar{\mathbf{x}}_{(t)})\right] - \eta\,\mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}_{(t)}), \frac{1}{n}\sum_{i=1}^n\left[\mathbf{g}^i_{(t)} - \hat{\mathbf{g}}^i_{(r(t))} + \hat{\mathbf{g}}_{(r(t))}\right]\right\rangle \\
&\quad + \eta^2\frac{L}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n\left[\mathbf{g}^i_{(t)} - \hat{\mathbf{g}}^i_{(r(t))} + \hat{\mathbf{g}}_{(r(t))}\right]\right\|_2^2 \\
&= \mathbb{E}\left[f(\bar{\mathbf{x}}_{(t)})\right] - \eta\,\mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}_{(t)}), \frac{1}{n}\sum_{i=1}^n \mathbf{g}^i_{(t)}\right\rangle + \frac{L}{2}\eta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{g}^i_{(t)}\right\|_2^2 \\
&\overset{(2)}{\le} \mathbb{E}\left[f(\bar{\mathbf{x}}_{(t)})\right] - \eta\,\mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}_{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t)})\right\rangle + \frac{L}{2}\eta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t)})\right\|_2^2 + \frac{\eta^2}{2}L\frac{\sigma^2}{n} \\
&= \mathbb{E}\left[f(\bar{\mathbf{x}}_{(t)})\right] - \frac{\eta}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 - \frac{\eta}{4}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t)})\right\|_2^2 + \frac{\eta}{2}\mathbb{E}\left[Q_{(t)}\right] + \frac{\eta^2}{2}L\frac{\sigma^2}{n}.
\end{aligned}$$

Summing over k from 0 to $\tau-1$ and over r from 0 to $R-1$, and multiplying both sides by $\frac{4}{\eta T}$, we have

$$\begin{aligned}
&\frac{2}{T}\sum_{t=0}^{R\tau-1}\left[\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2+\frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}^i_{(t)})\right\|_2^2\right]\\
&\leq \frac{4}{\eta T}\mathbb{E}\left[f(\bar{\mathbf{x}}_{(0)})-f(\bar{\mathbf{x}}_{(T)})\right]+2\eta\frac{L\sigma^2}{n}+\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{(r\tau+k)}\right]\\
&\leq \frac{4\Delta}{\eta T}+2\eta\frac{L\sigma^2}{n}+\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{(r\tau+k)}\right].
\end{aligned}$$

□

Lemma 19. *Under Assumptions 2 and 3, for $\gamma=\frac{1}{3(\tau-1)}$ and $\eta\leq\frac{(1-\rho/L)}{2\rho}\gamma$, we have*

$$\begin{aligned}
\mathbb{E}\left[\Xi_{(t)}\right]\leq\ & 3\gamma^{-2}(1+\gamma^{-1})\eta^4\delta^2\sum_{l=0}^{k-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}^i_{(r(t)+l)})\right\|_2^2\\
&+3\left(\frac{1+\gamma^{-1}}{\tau}+1\right)\eta^2k\sigma^2+3\gamma^{-2}\eta^4k\delta^2\frac{\sigma^2}{n},
\end{aligned} \tag{47}$$

where $k=t-r(t)$.

Proof. For $\gamma>0$, when $t+1>r(t+1)$, in view of

$$\begin{aligned}
&\mathbb{E}\left[\Xi_{(t+1)}\right]\\
&\overset{(21)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}^i_{(t+1)}-\left(\bar{\mathbf{x}}_{(t)}-\eta\nabla f(\bar{\mathbf{x}}_{(t)})\right)\right\|_2^2\\
&\overset{(5)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}^i_{(t)}-\eta\left(\mathbf{g}^i_{(t)}-\hat{\mathbf{g}}^i_{(r(t))}+\hat{\mathbf{g}}_{(r(t))}\right)-\bar{\mathbf{x}}_{(t)}+\eta\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2\\
&\overset{(2)}{\leq}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}^i_{(t)}-\eta\left(\nabla f_i(\mathbf{x}^i_{(t)})-\hat{\mathbf{g}}^i_{(r(t))}+\hat{\mathbf{g}}_{(r(t))}\right)-\bar{\mathbf{x}}_{(t)}+\eta\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2+\eta^2\sigma^2\\
&\leq(1+\gamma)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}^i_{(t)}-\eta\left(\nabla f_i(\mathbf{x}^i_{(t)})-\nabla f_i(\bar{\mathbf{x}}_{(r(t))})+\nabla f(\bar{\mathbf{x}}_{(r(t))})\right)-\bar{\mathbf{x}}_{(t)}+\eta\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2\\
&\quad+\left(1+\gamma^{-1}\right)\eta^2\mathbb{E}\left\|\hat{\mathbf{g}}^i_{(r(t))}-\hat{\mathbf{g}}_{(r(t))}-\nabla f_i(\bar{\mathbf{x}}_{(r(t))})+\nabla f(\bar{\mathbf{x}}_{(r(t))})\right\|_2^2+\eta^2\sigma^2\\
&\overset{(2)}{\leq}(1+\gamma)\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}^i_{(t)}-\eta\left(\nabla f_i(\mathbf{x}^i_{(t)})-\nabla f_i(\bar{\mathbf{x}}_{(r(t))})+\nabla f(\bar{\mathbf{x}}_{(r(t))})\right)-\bar{\mathbf{x}}_{(t)}+\eta\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2\\
&\quad+\left(\frac{1+\gamma^{-1}}{\tau}+1\right)\eta^2\sigma^2,
\end{aligned} \tag{48}$$

and, for $\eta \le \frac{(1-\rho/L)}{2\rho}\gamma$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_{(t)}^{i}-\eta\left(\nabla f_i(\mathbf{x}_{(t)}^{i})-\nabla f_i(\bar{\mathbf{x}}_{(r(t))})+\nabla f(\bar{\mathbf{x}}_{(r(t))})\right)-\bar{\mathbf{x}}_{(t)}+\eta\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(1+\gamma)\left\|\mathbf{x}_{(t)}^{i}-\eta\nabla f_i(\mathbf{x}_{(t)}^{i})-\bar{\mathbf{x}}_{(t)}+\eta\nabla f_i(\bar{\mathbf{x}}_{(t)})\right\|_2^2\right. \\
&\qquad \left.+(1+\gamma^{-1})\eta^2\left\|\nabla f_i(\bar{\mathbf{x}}_{(r(t))})-\nabla f(\bar{\mathbf{x}}_{(r(t))})-\nabla f_i(\bar{\mathbf{x}}_{(t)})+\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2\right] \\
&\overset{(8)}{\le} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(1+\gamma)\left\|\mathbf{x}_{(t)}^{i}-\eta\nabla f_i(\mathbf{x}_{(t)}^{i})-\bar{\mathbf{x}}_{(t)}+\eta\nabla f_i(\bar{\mathbf{x}}_{(t)})\right\|_2^2+(1+\gamma^{-1})\eta^2\delta^2\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2\right] \\
&\overset{(28)}{\le} (1+\gamma)\left(1+\frac{2L}{L-\rho}\cdot\rho\eta\right)\mathbb{E}\left[\Xi_{(t)}\right]+(1+\gamma^{-1})\eta^2\delta^2\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2 \\
&\le (1+\gamma)^2\,\mathbb{E}\left[\Xi_{(t)}\right]+(1+\gamma^{-1})\eta^2\delta^2\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2,
\end{aligned} \tag{49}$$

we have

$$\begin{aligned}
&\mathbb{E}\left[\Xi_{(t+1)}\right] \\
&\le (1+\gamma)^3\,\mathbb{E}\left[\Xi_t\right]+(1+\gamma)(1+\gamma^{-1})\eta^2\delta^2\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2+\left(\frac{1+\gamma^{-1}}{\tau}+1\right)\eta^2\sigma^2.
\end{aligned}$$

Also, when $t+1 > r(t+1)$,

$$\begin{aligned}
&\mathbb{E}\left\|\bar{\mathbf{x}}_{(t+1)}-\bar{\mathbf{x}}_{(r(t+1))}\right\|_2^2 \\
&\overset{(5)}{=} \mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\eta\cdot\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_{(t)}^{i}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2 \\
&\overset{(2)}{\le} \mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\eta\cdot\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_{(t)}^{i})-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2+\eta^2\frac{\sigma^2}{n} \\
&\le (1+\gamma)\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2+(1+\gamma^{-1})\eta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_{(t)}^{i})\right\|_2^2+\eta^2\frac{\sigma^2}{n}.
\end{aligned} \tag{50}$$

Now, when $t+1 > r(t+1)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\Xi_{(t+1)}+\gamma^{-2}\eta^2\delta^2\left\|\bar{\mathbf{x}}_{(t+1)}-\bar{\mathbf{x}}_{(r(t+1))}\right\|_2^2\right] \\
&\overset{(48)}{\le} (1+\gamma)^3\cdot\mathbb{E}\left[\Xi_{(t)}\right]+(1+\gamma)(1+\gamma^{-1})\eta^2\delta^2\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2 \\
&\qquad+\gamma^{-2}\eta^2\delta^2\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t+1)}-\bar{\mathbf{x}}_{(r(t+1))}\right\|_2^2+\left(\frac{1+\gamma^{-1}}{\tau}+1\right)\eta^2\sigma^2 \\
&\overset{(50)}{\le} (1+\gamma)^3\cdot\mathbb{E}\left[\Xi_{(t)}\right]+\left[(1+\gamma)(1+\gamma^{-1})+(1+\gamma)\gamma^{-2}\right]\eta^2\delta^2\,\mathbb{E}\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2 \\
&\qquad+\gamma^{-2}(1+\gamma^{-1})\eta^4\delta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_{(t)}^{i})\right\|_2^2+\left(\frac{1+\gamma^{-1}}{\tau}+1\right)\eta^2\sigma^2+\gamma^{-2}\eta^4\delta^2\frac{\sigma^2}{n} \\
&\le (1+\gamma)^3\cdot\mathbb{E}\left[\Xi_{(t)}+\gamma^{-2}\eta^2\delta^2\left\|\bar{\mathbf{x}}_{(t)}-\bar{\mathbf{x}}_{(r(t))}\right\|_2^2\right] \\
&\qquad+\gamma^{-2}(1+\gamma^{-1})\eta^4\delta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_{(t)}^{i})\right\|_2^2+\left(\frac{1+\gamma^{-1}}{\tau}+1\right)\eta^2\sigma^2+\gamma^{-2}\eta^4\delta^2\frac{\sigma^2}{n}.
\end{aligned}$$

Finally, for $\gamma \le \frac{1}{3(\tau-1)}$,

$$
\begin{aligned}
&\mathbb{E}\left[\Xi_{(t)} + \gamma^{-2}\eta^2\delta^2 \left\|\bar{\mathbf{x}}_{(t)} - \bar{\mathbf{x}}_{(r(t))}\right\|_2^2\right] \\
&\le (1+\gamma)^3 \cdot \mathbb{E}\left[\Xi_{(t-1)} + \gamma^{-2}\eta^2\delta^2 \left\|\bar{\mathbf{x}}_{(t-1)} - \bar{\mathbf{x}}_{(r(t-1))}\right\|_2^2\right] \\
&\quad + \gamma^{-2}(1+\gamma^{-1})\eta^4\delta^2\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t-1)})\right\|_2^2 + \left(\frac{1+\gamma^{-1}}{\tau} + 1\right)\eta^2\sigma^2 + \gamma^{-2}\eta^4\delta^2\frac{\sigma^2}{n} \\
&\le \cdots \\
&\le 0 + e\gamma^{-2}(1+\gamma^{-1})\eta^4\delta^2 \sum_{l=0}^{k-1} \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(r(t)+l)})\right\|_2^2 \\
&\quad + e\left(\frac{1+\gamma^{-1}}{\tau} + 1\right)\eta^2 k\sigma^2 + e\gamma^{-2}\eta^4 k\delta^2\frac{\sigma^2}{n},
\end{aligned}
$$

where $k = t - r(t)$. □

A.3.1 Proof of Theorem 4

Proof of Theorem 4. For $\eta \le \min\left\{\frac{1}{2L}, \frac{1-\rho/L}{6\rho(\tau-1)}\right\}$,

$$
\begin{aligned}
&\frac{2}{T}\sum_{t=0}^{R\tau-1}\left[\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(t)})\right\|_2^2\right] \\
&\overset{(46)}{\le} \frac{4\Delta}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{(r\tau+k)}\right] \\
&\le \frac{4\Delta}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{2L^2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[\Xi_{(r\tau+k)}\right] \\
&\overset{(47)}{\le} \frac{4\Delta}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{162}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\eta^4L^2\delta^2(\tau-1)^3\tau\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i_{(r\tau+k)})\right\|_2^2 \\
&\quad + 6\eta^2L^2(\tau-1)\sigma^2 + \frac{27}{2}\eta^4L^2\delta^2(\tau-1)^3\frac{\sigma^2}{n}.
\end{aligned}
$$

Then, for $\eta \le \frac{1}{4\sqrt{L\delta}\tau}$,

$$
\frac{2}{T}\sum_{t=0}^{R\tau-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 \le \frac{4\Delta}{\eta T} + 2\eta\frac{L\sigma^2}{n} + 6\eta^2L^2(\tau-1)\sigma^2 + \frac{27}{2}\eta^4L^2\delta^2(\tau-1)^3\frac{\sigma^2}{n}.
$$

Finally, (19) follows from the following assignment

$$
\eta := \min\left\{\frac{1}{2L}, \frac{1-\rho/L}{6\rho(\tau-1)}, \frac{1}{4\sqrt{L\delta}\tau}, \sqrt{\frac{\Delta n}{2L\sigma^2 T}}, \left(\frac{2\Delta}{3L^2(\tau-1)\sigma^2 T}\right)^{\frac{1}{3}}\right\}.
$$

□

A.3.2 Proof of Theorem 5

Proof of Theorem 5. For $\eta \leq \min\left\{\frac{1}{2L}, \frac{1-\rho/L}{6\rho(\tau-1)}\right\}$,

$$\begin{aligned}
&\frac{2}{T}\sum_{t=0}^{R\tau-1}\left[\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 + \frac{1}{2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_{(t)}^i)\right\|_2^2\right] \\
\overset{(46)}{\leq}\ & \frac{4\mathit{\Delta}}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[Q_{(r\tau+k)}\right] \\
\overset{(22)}{\leq}\ & \frac{4\mathit{\Delta}}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{16\bar{\delta}^2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\left[\mathit{\Xi}_{(r\tau+k)}\right] + 0 \\
\overset{(47)}{\leq}\ & \frac{4\mathit{\Delta}}{\eta T} + 2\eta\frac{L\sigma^2}{n} + \frac{1296}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\eta^4\bar{\delta}^2\delta^2(\tau-1)^3\tau\,\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_{(r\tau+k)}^i)\right\|_2^2 \\
&+ 48\eta^2\bar{\delta}^2(\tau-1)\sigma^2 + 108\eta^4\bar{\delta}^2\delta^2(\tau-1)^3\frac{\sigma^2}{n}.
\end{aligned}$$

Then, for $\eta \leq \frac{1}{6\sqrt{\bar{\delta}\delta}\tau}$,

$$\frac{2}{T}\sum_{t=0}^{R\tau-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_{(t)})\right\|_2^2 \leq \frac{4\mathit{\Delta}}{\eta T} + 2\eta\frac{L\sigma^2}{n} + 48\eta^2\bar{\delta}^2(\tau-1)\sigma^2 + 108\eta^4\bar{\delta}^2\delta^2(\tau-1)^3\frac{\sigma^2}{n}.$$

Finally, (15) follows from the following assignment

$$\eta := \min\left\{\frac{1}{2L}, \frac{1-\rho/L}{6\rho(\tau-1)}, \frac{1}{6\sqrt{\bar{\delta}\delta}\tau}, \sqrt{\frac{2\mathit{\Delta} n}{L\sigma^2 T}}, \left(\frac{\mathit{\Delta}}{12\bar{\delta}^2(\tau-1)\sigma^2 T}\right)^{\frac{1}{3}}\right\}.$$

□

B Details and Further Experiments

B.1 Details of the Experiments

We use a smoothed variant of Huber loss $h(\cdot)$ in all our experiments. We recall its definition:

$$h(u) = \begin{cases} \frac{1}{2}u^2, & |u| \leq 1, \\ -\frac{1}{6}(|u|-1)^3 + \frac{1}{2}u^2, & 1 < |u| \leq 2, \\ \frac{3}{2}|u| - \frac{7}{6}, & |u| > 2. \end{cases}$$

In the original Huber loss function, the quadratic loss is used for small errors and linear loss for large errors. While in the smoothed variant, a cubic term is added in the middle range to provide a smooth transition between the quadratic and linear sections, thus ensuring the Lipschitz continuity of the Hessian.

The data is generated as follows. We first generate a diagonal matrix Λ with values ranging from 0 to a maximal value L. Next, we generate a random orthonormal matrix $\mathbf{Q}$, and compute $\mathbf{A} = \sqrt{\Lambda}\mathbf{Q}^{\mathrm{T}}$. With $\mathbf{H} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$, the quadratic function $\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{H}\mathbf{x}$ has L-Lipschitz continuous gradient. Then, we assign the function for each worker: $f_i(x) = \sum_{j=1}^{d} h\left(\mathbf{A}_i(j)\cdot(\mathbf{x}-\mathbf{x}_i^*)\right)$, where $\mathbf{A}_i$ introduces random noise to $\mathbf{A}$ and $\mathbf{x}_i^*$ represents a unique optimal point for each worker. The function $h(u)$ behaves quadratically when $|u|$ is small, specifically $f_i(x) = \sum_{j=1}^{d} h\left(\mathbf{A}_i(j)\cdot(\mathbf{x}-\mathbf{x}_i^*)\right) \approx \frac{1}{2}(\mathbf{x}-\mathbf{x}_i^*)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}(\mathbf{x}-\mathbf{x}_i^*)$. This allows us to approximately control the behavior of the functions. We control the Hessian similarity δ by adjusting the noise in $\mathbf{A}_i$, and the gradient similarity ζ by the variation in $\mathbf{x}_i^*$.

We run all the experiments on Intel Xeon CPU E5-2680v4. We use three different random seeds $\{111, 222, 333\}$ for all experiments and report the average results across the runs.

Codes are available here: `https://github.com/riekenluo/Pub_Code_LocalSGD_and_SCAFFOLD`.

B.2 Further Experiments

We run additional experiments to validate the impact of the other parameters on the convergence rates. Following our previous experiment setup, we set $L \approx 1.0$, $\Delta \approx 1.0$, $\sigma \approx 0.01$, and $\tau = 50$. To highlight the optimization term, we consider the 10-th communication round; to highlight the heterogeneity term, we consider the 50-th communication round.

The level of non-convexity is controlled by λ in our experiments. The regularizer $\frac{x^2}{1+x^2}$ exhibits non-convexity only when $|x| > \frac{\sqrt{3}}{3}$. Therefore, to validate the dependence on different levels of non-convexity, we need to increase the scales of $\mathbf{x}_i^*$ in these experiments.

For LocalSGD, in addition to the experiments in Section 6, we perform further tests by varying the parameters of gradient similarity, Hessian similarity, and weak convexity. Specifically, from the basic setting $(\zeta, \delta, \lambda) = (0.03, 0.01, 0.01)$, we change ζ over $\{0.03, 0.04, 0.05, 0.06, 0.07\}$, δ over $\{0.01, 0.015, 0.02, 0.025, 0.03\}$, and λ over $\{0.01, 0.015, 0.02, 0.025, 0.03\}$. The results of these experiments are plotted in Figure 2.

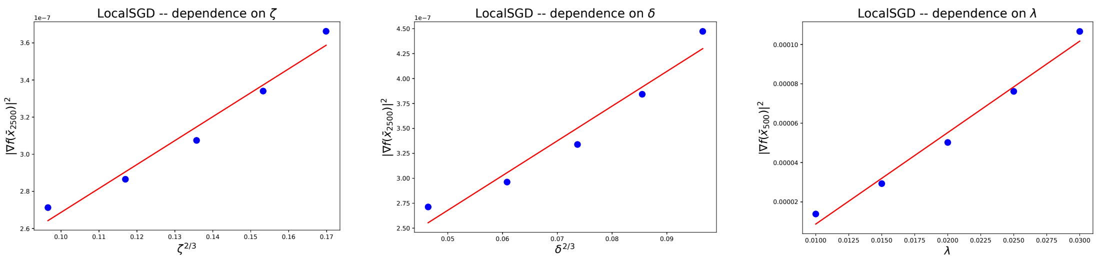

Figure 2: Convergence rates of LocalSGD with changing parameters.

For SCAFFOLD, in addition to the experiments in Section 6, we perform further tests by varying the parameters of Hessian similarity and weak convexity. Specifically, from the basic setting $(\delta, \lambda) = (0.1, 0.01)$, we change δ over $\{0.1, 0.15, 0.2, 0.25, 0.3\}$ and λ over $\{0.01, 0.015, 0.02, 0.025, 0.03\}$. The results of these experiments are plotted in Figure 3.

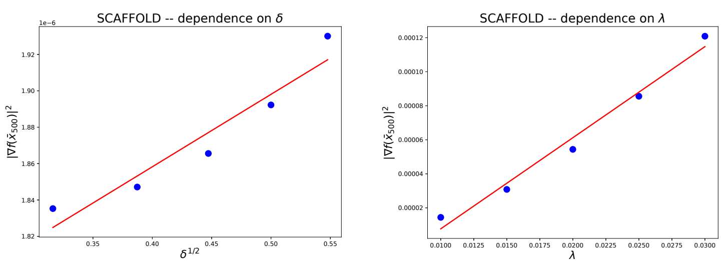

Figure 3: Convergence rates of SCAFFOLD with changing parameters.

The blue dots represent the convergence rates corresponding to each parameter value, and the red lines show the linear fit of these data points. From the figures, it is evident that the convergence rates are positively correlated with these parameters. All these experiments are qualitative in nature, as it is not possible to completely isolate the effects of other additive terms on the convergence rates.

C Discussions on the Related Work on Generalization

Sefidgaran et al. (2024) studies LocalSGD when T, the number of queries to $\mathcal{SO}$, is fixed. Under the assumption that the loss functions are Σ-subgaussian, they show that the generalization gap is bounded by $\mathcal{O}(\Sigma/\sqrt{\tau})$, which decreases as τ increases. Therefore, for this intermittent communication model, combining the result of Sefidgaran et al. (2024) and the optimization bounds (in this paper or in other related works) yields a preliminary insight for minimizing the true risk, *i.e.* to choose the communication interval τ balancing the empirical risk and the generalization gap. Please refer to Sefidgaran et al. (2024, section 6) for more discussions.