On the Convergence of Local SGD Under Third-Order Smoothness and Hessian Similarity

Ali Zindari ALI.ZINDARI@CISPA.DE
CISPA, Germany

Ruichen Luo LUORUICHEN@OUTLOOK.COM
Zhejiang University, China

Sebastian U. Stich STICH@CISPA.DE
CISPA, Germany

Abstract

Local SGD (i.e. Federated Averaging without client sampling) is widely used for solving federated optimization problems in the presence of heterogeneous data. However, there is a gap between the existing convergence rates for Local SGD and its observed performance on real-world problems. It seems that current rates do not correctly capture the effectiveness Local SGD. We first show that the existing rates for Local SGD in a heterogeneous setting cannot recover the correct rate when the global function is quadratic. Then we first derive a new rate for the case that the global function is a general strongly convex function depending on third-order smoothness and Hessian similarity. These additional parameters allow us to capture the problem in a more refined way and to overcome some of the limitations of the previous worst-case results derived under the standard assumptions. We further extend our analysis to the case when all clients have non-convex quadratic functions with identical Hessians.

1. Introduction

Machine learning (ML) models are getting bigger and bigger every day, along with the huge amount of data needed to train such large models. Data confidentiality is also becoming more important for security reasons, making the training process even more challenging. All this motivates us to use Federated Learning (FL) for training our models in a distributed environment [7, 9]. Federated Averaging (FedAvg) and Local SGD [11] allows a shared ML model to be trained across multiple clients without having to share data with other clients. Although local SGD has been extensively studied, the convergence results do often not correctly capture the effectiveness Local SGD in practice, as they are often a bit too pessimistic [12]. This work aims to bridge this gap by introducing higher-order smoothness and Hessian similarity to allow for a more fine-grained analysis.

Related Work. The analysis of Local SGD have first been carried out under the assumption of homogeneous (IID) data distribution [14]. State-of-the-art analyses also consider the heterogeneous data case. Different measures are used to quantify the non-iid ness level of the data. One such measure is *gradient dissimilarity*, which comes in two flavors: Woodworth et al. [15] study Local SGD under $\bar{\zeta}$ assumption (uniform bound on the gradient dissimilarity) and show a benefit of local steps in the first term of their rate (optimization term). However, the works [4–6] proposed convergence rates for Local SGD based on the $\zeta_\star$ assumption (bounded gradient dissimilarity only locally at the solution) which is a weaker assumption. These analyses do not show a benefit of local steps in the optimization term. It is still unclear under which conditions we can have the benefit of local steps and $\zeta_\star$ assumption at the same time. A recent work [10] provides a lower bound for heterogeneous Local SGD in the convex setting, which shows that for this class of functions, it is not possible to have the benefit of local steps and $\zeta_\star$. There is also a line of research that argues that the gradient similarity assumption is too pessimistic based on insights from experiments. Wang et al. [12] proposed a new measure ρ termed *average drift at optimum* and showed that this measure is usually close to zero in practice. They provided a rate for strongly convex objectives with access to full gradients based on this parameter which can have a linear convergence rate if ρ is approximately zero. Another paper [13] introduced a new parameter for measuring heterogeneity called *heterogeneity-driven Lipschitz condition on averaged gradients*. They proposed a rate for non-convex functions based on this new parameter which also has the benefit of local steps in the optimization term. Another less-discovered line of research is higher-order smoothness. Yuan and Ma [16] studied Local SGD for convex and strongly convex objectives under third-order smoothness. However, this work considers only the homogeneous setting. Glasgow et al. [1] used third-order smoothness for non-convex functions in the homogeneous setting. In addition to higher-order smoothness, there is a hope that we can make use of Hessian similarity. This measure has been used in some works like [2–4] but it has not been utilized for Local SGD yet.

Our contribution. In this work, we derive a novel convergence rate for Local SGD that captures the influence of higher order smoothness and Hessian similarity at the same time for the case that the global function f is strongly convex. We show that Local SGD can benefit from local steps when the client's Hessians are similar (or identical, for the special case of quadratics). Although some previous results for the quadratic case are known [1, 4, 15] we are not aware of any previous results for quadratic functions that show this speedup even in the heterogeneous setting (i.e. with positive gradient dissimilarity). We further show that this speedup does not require convexity and does also hold on non-convex quadratics.

2. Setting

The goal in FL is to learn a shared model among M clients that has a good performance on each client's data for certain ML tasks. Each client $m \in [M]$ has only access to its own dataset (modeled by a local objective function $f_m\colon \mathbb{R}^d \to \mathbb{R}$) and samples one data point at each time step. At each round, every client performs exactly K steps of SGD on its data and then communicates its parameters to a central server for averaging. We have a total of R rounds which implies that each client performs a total of $T = KR$ steps of SGD. The problem can be formulated as below:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[f(\mathbf{x}) := \frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{\xi_m\sim\mathcal{D}_m} f_m(\mathbf{x},\xi_m) \right], \tag{1}$$

where $\mathcal{D}_m$ denotes the data distribution on client m and ξ_m denotes a sample of data drawn from $\mathcal{D}_m$. The Local SGD [11] update formula can be written as:

$$\mathbf{x}_{t+1}^m = \begin{cases} \mathbf{x}_t^m - \eta g_t^m, & \text{if } t+1 \notin \mathcal{I}_{\text{syn}} \\ \frac{1}{M}\sum_{m\in[M]} \left(\mathbf{x}_t^m - \eta g_t^m\right), & \text{if } t+1 \in \mathcal{I}_{\text{syn}} \end{cases}, \tag{2}$$

where $g_t^m = \nabla f_m(\mathbf{x}_t^m;\xi_m)$ and $\mathcal{I}_{syn}$ is a set of synchronization indices (we will always use $\mathcal{I}_{\text{syn}} = \{Kn \mid n \in \mathbb{N}\}$ in this work). Now we introduce a set of assumptions that will be used in this work. For each theorem, we will explicitly mention which assumptions are used.

Assumption 1 (Smoothness) *A function f is called to be L-smooth if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have*

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\left\|\mathbf{x} - \mathbf{y}\right\|^2 . \tag{3}$$

Assumption 2 (Strong Convexity) *A function f is called to be μ-strongly convex if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have*

$$\frac{\mu}{2}\left\|\mathbf{x} - \mathbf{y}\right\|^2 + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + f(\mathbf{x}) \le f(\mathbf{y}) . \tag{4}$$

For the case of $\mu = 0$, the function f is just convex.

Smoothness, convexity and strong convexity are very common assumptions that are used in many previous works [4, 5, 15]. Note that some works only need the global function to be convex while some require every f_m to be convex. The same argument holds for strong convexity as well.

Assumption 3 (Lipschitz Hessian) *Function f has a H Lipschitz Hessian if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have*

$$\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\|_F^2 \le H^2 \left\|\mathbf{x} - \mathbf{y}\right\|^2 . \tag{5}$$

The third-order smoothness is less common in the literature and is used by a few works like [1, 16]. Note that it can be the case that $H \lll L$.

Assumption 4 (Bounded Noise Variance) *For every client $m \in [M]$ the variance of noise imposed by SGD is uniformly upper bounded*

$$\mathbb{E}_{\xi_m \sim \mathcal{D}_m}\left[\left\|\nabla f_m(\mathbf{x}_t^m; \xi_m) - \nabla f_m(\mathbf{x}_t^m)\right\|^2\right] \le \sigma^2 . \tag{6}$$

Assumption 5 (Gradient Similarity) *We assume that the difference between the local gradients and global gradient is bounded by $\bar{\zeta}$ for any $\mathbf{x} \in \mathbb{R}^d$.*

$$\sup_{m \in [M]} \left\|\nabla f_m(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \le \bar{\zeta}^2 . \tag{7}$$

Assumption 6 (Gradient Similarity at Optimum) *We assume that the difference between the local gradients and global gradient is bounded by $\zeta_\star$ for $\mathbf{x} = \mathbf{x}^\star \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$.*

$$\sup_{m \in [M]} \left\|\nabla f_m(\mathbf{x}^\star) - \nabla f(\mathbf{x}^\star)\right\|^2 = \sup_{m \in [M]} \left\|\nabla f_m(\mathbf{x}^\star)\right\|^2 \le \zeta_\star^2 . \tag{8}$$

Both Assumptions 5 and 6 have been widely used in different works for measuring heterogeneity [4–6, 15]. Assumption 6 is a special case of Assumption 5 and the latter is stronger assumption. In addition, for using Assumption 5 we should be careful as for example it does not hold for the class of quadratic functions unless we can show that all the parameters generated by SGD in the optimization trajectory remain in a ball with a certain fixed radius.

Assumption 7 (Hessian Similarity) *We assume that the difference between the local Hessians and global Hessian at any point $\mathbf{x} \in \mathbb{R}^d$ is bounded by δ_h.*

$$\sup_{m \in [M]} \left\|\nabla^2 f_m(\mathbf{x}) - \nabla^2 f(\mathbf{x})\right\|_F^2 \le \delta_h^2 . \tag{9}$$

This similarity has been used in a few works such as [3, 4]. Note that if f_m is L-smooth, then we have that $\delta_h \le 2L$. In general, δ_h can be much less than L and even can be zero while L is very large.

3. Convergence Rates

In this section, we provide our convergence rates for Local SGD: (1) for the case that f is a general strongly convex function. This new rate is based on third-order smoothness and Hessian similarity and (2) for the case that all clients are quadratic non-convex functions with identical Hessians. We use intuition from simple quadratic functions to show that even for this simple case, existing convergence rates cannot capture the effectiveness of Local SGD while our new rate can recover this specific case. One of the main questions in FL is whether can we benefit from more local steps and less communications. And under which conditions can we beat mini-batch SGD? We start by showing two extreme cases where we assume that each client is a quadratic function. In these two scenarios, we are able to see the benefit of local steps.

Convex quadratics with identical Hessians: In this case, we have that $\delta_h = 0$ so the global optimum $\mathbf{x}^\star$ is simply the average of clients' optima so $\mathbf{x}^\star = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_\star^m$ where $\mathbf{x}_\star^m$ is the optimum of client m. It is clear that we only need $K \to \infty$ and $R = 1$ to achieve the global optimum. This is a case in which we should see the benefit of local steps. Having this benefit is important in many real-world applications as local steps are usually cheap to compute while communication comes at a high cost. So we would rather perform more local steps and less communication. However, none of the rates proposed by [5, 6, 15] can capture this setting. The reason is that these papers lack parameters for controlling the similarity between Hessians and measuring higher-order smoothness. We will recover this result as a special case of our Theorem 1 by introducing new parameters δ_h and H in our convergence rate.

Convex quadratics with identical optima: In this case, we have that $\zeta_\star = 0$. It implies that $\forall m \in [M], \mathbf{x}_\star^m = \mathbf{x}^\star$. So, it is clear that we need $K \to \infty$ and $R = 1$ to reach the global optimum. This case also cannot be recovered from the current rates [4–6] by just setting $\zeta_\star = 0$. We conjecture that for the case that each f_m is quadratic and $\zeta_\star = 0$, we should have a rate of $\mathcal{O}(\frac{1}{KR})$ which has the benefit of local steps. However, working with the weaker assumption of local gradient dissimilarity ($\zeta_\star$) is much more challenging than assuming uniformly bounded gradient diversity ($\bar{\zeta}$). That is why we focus on the $\bar{\zeta}$ assumption in this work and leave the extension to $\zeta_\star$ for future works.

We now state our main result:

Theorem 1 *Let Assumptions 3, 4, 5 and 7 hold. Also consider each f_m to be L-smooth and let the global function f to be μ-strongly convex. Also let $\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\|^2 \leq B^2$. With a learning rate of $\eta \leq \frac{1}{6L}$, we have the following convergence rate for Local SGD on heterogeneous data.*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] \leq c_2\left(\frac{LB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \psi(\delta_h) + \varphi(H)\right), \tag{10}$$

$$\psi(\delta_h) := \frac{(B^6\delta_h^2\sigma^2)^{1/4}}{K^{2/4}R^{3/4}} + \frac{(B^6\delta_h^2\bar{\zeta}^2)^{1/4}}{K^{1/4}R^{2/3}} + \frac{(B^4\delta_h^2\sigma^2)^{1/3}}{\mu^{1/3}K^{1/3}R^{2/3}} + \frac{(B^4\delta_h^2\bar{\zeta}^2)^{1/3}}{\mu^{1/3}R^{2/3}}, \tag{11}$$

$$\varphi(H) := \frac{(B^{10}MH^2\sigma^4)^{1/6}}{K^{3/6}R^{5/6}} + \frac{(B^{10}MH^2\bar{\zeta}^4)^{1/6}}{K^{1/6}R^{5/6}} + \frac{(B^8MH^2\sigma^4)^{1/5}}{\mu^{1/5}K^{2/5}R^{4/5}} + \frac{(B^8MH^2\bar{\zeta}^4)^{1/5}}{\mu^{1/5}R^{4/5}}, \tag{12}$$

where c_2 is an absolute constant.

We define two extra functions to simplify the notations for the extra terms we have in our rate. The functions $\psi(\delta_h)$ and $\varphi(H)$ are defined to capture the terms that are affected by δ_h and H. To the best of our knowledge, this is the first work that analyses Local SGD with heterogeneous data, using Hessian similarity and third-order smoothness simultaneously. The work [16] proposed a rate for Local SGD for convex and strongly convex cases based on H but with homogeneous data. Note that for our Theorem 1, we only need the global function f to be strongly convex in contrast to some other papers that require each f_m to be strongly convex [4, 6, 15] which is a stronger assumption. One problem with the current analysis based on $\bar{\zeta}$ is that this parameter can get very large. However, in our rates, wherever we have $\bar{\zeta}$, it is multiplied by either δ_h or H, which can be very small or even zero. In contrast, in other works like [15], the term $L\bar{\zeta}$ appears in their rates, which can be significantly larger than $\delta_h\bar{\zeta}$ or $H\bar{\zeta}$ in our rates.

Corollary 2 (Convex Quadratics with identical Hessians) *Assume that the global function f is quadratic so $H = 0$ which yields $\varphi(H) = 0$. Also consider the clients to have identical Hessians so $\delta_h = 0$ and $\psi(\delta_h) = 0$. Then we have the following convergence rate for Local SGD on heterogeneous data (we do not make an assumption on $\bar{\zeta}$):*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] \leq \frac{LB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}}\,. \tag{13}$$

It is clear that for converging to the global optimum $\mathbf{x}^\star$, we need $K \to \infty$ and $R = 1$. Now we can see the benefit of local steps in this rate. Note that in our rate, only the global function f needs to be a quadratic function while clients f_m can be any arbitrary functions with identical Hessians. This is much weaker than assuming every client f_m to be a quadratic function.

Theorem 3 *Let Assumption 4 hold and suppose all clients to be arbitrary (possibly non-convex) quadratic functions with identical Hessians, which yields $\delta_h = 0$. Also let $f(\mathbf{x}_0) - f(\mathbf{x}^\star) \leq \Delta$. With a learning rate of $\eta \leq \frac{1}{L}$, we have the following convergence rate for Local SGD on heterogeneous data:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla f(\bar{\mathbf{x}}_t)\big\|^2\Big] \leq \frac{L\Delta}{KR} + \frac{\sigma\sqrt{L\Delta}}{\sqrt{MKR}}\,. \tag{14}$$

Previously, the work [1] proposed a rate for the non-convex case which is based on third-order smoothness but with homogeneous data. For the special case that functions are quadratics which implies $H = 0$ they get a rate of $\mathcal{O}(\frac{L\Delta}{KR} + \frac{\sigma\sqrt{L\Delta}}{\sqrt{MKR}})$. However, from the above theorem we know that we should be able to get the same rate even in the heterogeneous setting. We leave the rate for general non-convex setting based on δ_h and H to future works.

4. Conclusion

In this work, we introduced Hessian similarity and third-order smoothness as new controllable parameters for analyzing Local SGD with heterogeneous data. Our rate only needs the global function to be strongly convex and the clients to be just L-smooth. We also showed that non-convex quadratics with identical Hessians can benefit from local steps in the heterogeneous regime. Our new result can bridge the gap between theory and practice as for example, the widely used MSE loss function for linear regression tasks is third-order smooth.

References

[1] Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

[2] Samuel Horváth, Maziar Sanjabi, Lin Xiao, Peter Richtárik, and Michael Rabbat. Fedshuffle: Recipes for better use of local work in federated learning. *arXiv preprint arXiv:2204.13169*, 2022.

[3] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020.

[4] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[5] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[6] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

[7] Jakub Konečnỳ, Brendan McMahan, and Daniel Ramage. Federated optimization: Distributed optimization beyond the datacenter. *arXiv preprint arXiv:1511.03575*, 2015.

[8] Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtárik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*, 2019.

[9] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[10] Kumar Kshitij Patel, Margalit Glasgow, Lingxiao Wang, Nirmit Joshi, and Nathan Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

[11] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[12] Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

[13] Jiayi Wang, Shiqiang Wang, Rong-Rong Chen, and Mingyue Ji. A new theoretical perspective on data heterogeneity in federated optimization. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

[14] Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020.

[15] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33: 6281–6292, 2020.

[16] Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020.

Appendix A. Proofs

Below, you can see the summary of the notations used in this paper.

Table 1: Summary of symbols used in this paper.

Symbol	Usage
$\mathbf{x}_t^m$	Parameters of client m at time step t.
g_t^m	Stochastic gradient on client m at time step t.
$\bar{\mathbf{x}}_t$	Average of iterates at time step t. $\frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_t^m$
$\mathbf{x}_\star^m$	Optimum of client m.
$\mathbf{x}^\star$	Global optimum.
$\bar{\zeta}$	Gradient similarity (Assumption 5).
$\zeta_\star$	Gradient similarity at optimum (Assumption 6).
σ	Uniform upper bound on the noise of stochastic gradient.
K	Number of local steps.
R	Number of rounds.
H	Lipschitz hessian parameter (Assumption 3).
δ_h	Hessian similarity (Assumption 7).
B	$\left\|\mathbf{x}_0 - \mathbf{x}^\star\right\| \leq B$.
Δ	$f(\mathbf{x}_0) - f(\mathbf{x}^\star) \leq \Delta$.
T	Total number of iterations. $T = KR$
c_1, c_2	Absolute constants for dropping numbers in proofs.
$\mathbb{E}_t$	Expectation conditioned on $\mathbf{x}_t^1, ..., \mathbf{x}_t^M$.
$\mathbb{E}$	Unconditional expectation.

We first introduce some useful lemmas which will be used throughout the proofs.

Lemma 4 *For a convex function f we have:*

$$f\Big(\frac{1}{M}\sum_{m=1}^{M}\mathbf{x}_m\Big) \leq \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{x}_m) \tag{15}$$

Lemma 5 *For a set of M vectors $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_M \in \mathbb{R}^d$ we have:*

$$\Big\|\sum_{m=1}^{M} \mathbf{a}_m\Big\|^2 \leq M \sum_{m=1}^{M} \big\|\mathbf{a}_m\big\|^2 \tag{16}$$

Lemma 6 *For a set of M vectors $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_M \in \mathbb{R}^d$ we have:*

$$\frac{1}{M}\sum_{m=1}^{M} \mathbf{a}_m^4 \leq M\Big(\frac{1}{M}\sum_{m=1}^{M} \mathbf{a}_m^2\Big)^2 \tag{17}$$

Lemma 7 *For two arbitrary vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $\gamma > 0$ we have:*

$$-\mathbf{a}^\top \mathbf{b} \leq \frac{\gamma}{2}\big\|\mathbf{a}\big\|^2 + \frac{1}{2\gamma}\big\|\mathbf{b}\big\|^2 \tag{18}$$

Lemma 8 *For two arbitrary vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $\gamma > 0$ we have:*

$$\big\|\mathbf{a} + \mathbf{b}\big\|^2 \leq \Big(1 + \frac{1}{\gamma}\Big)\big\|\mathbf{a}\big\|^2 + 2\gamma\big\|\mathbf{b}\big\|^2 \tag{19}$$

Lemma 9 *For any random variable X we have:*

$$\mathbb{E}\Big[\big\|X - \mathbb{E}[X]\big\|^2\Big] \leq \mathbb{E}\Big[\big\|X\big\|^2\Big] \tag{20}$$

Lemma 10 *Let Assumption 4 holds. Then we have:*

$$\mathbb{E}_t\Bigg[\Big\|\frac{1}{M}\sum_{m=1}^{M} g_t^m - \frac{1}{M}\sum_{m=1}^{M} \nabla f_m(\mathbf{x}_t^m)\Big\|^2\Bigg] \leq \frac{\sigma^2}{M} \tag{21}$$

Lemma 11 *Let f be a convex and L-smooth function. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have:*

$$\frac{1}{2L}\big\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\big\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle \tag{22}$$

Lemma 12 ([3, Lemma 3]) *Let Assumption 7 holds and $f(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} f_m(\mathbf{x})$. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have the following inequality:*

$$\big\|\nabla f_m(\mathbf{x}) - \nabla f(\mathbf{x}) + \nabla f(\mathbf{y}) - \nabla f_m(\mathbf{y})\big\|^2 \leq \delta_h^2\big\|\mathbf{x} - \mathbf{y}\big\|^2 \tag{23}$$

Lemma 13 ([8, Definition 2]) *Let function f satisfies Assumption 3, then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have the following inequality:*

$$\big\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}) - \nabla^2 f(\mathbf{y})(\mathbf{x} - \mathbf{y})\big\| \leq \frac{H}{2}\big\|\mathbf{x} - \mathbf{y}\big\|^2 \tag{24}$$

Lemma 14 *Let Assumptions 4 and 5 hold and the global function f be convex. We can upper bound the consensus error Ξ_t using the Lemma from [15]. For any fixed learning rate $\eta_t = \eta$ we have:*

$$\Xi_t = \frac{1}{M}\sum_{m=1}^{M} \big\|\mathbf{x}_t^m - \bar{\mathbf{x}}_t\big\|^2 \leq 3K\sigma^2\eta^2 + 6K^2\eta^2\bar{\zeta}^2 \tag{25}$$

Where $\mathbf{x}_t^m$ is the parameters of client m at time step t and $\bar{\mathbf{x}}_t = \frac{1}{M}\sum_{m=1}^{M} \mathbf{x}_t^m$.

A.1. Proof of Theorem 1

Proof We start with the distance from the optimal point and taking the conditional expectation on the previous iterate $\mathbf{x}_t^m, \forall m \in [M]$ we have:

$$\begin{aligned}
&\mathbb{E}_t \left\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}^\star\right\|^2 \\
&= \mathbb{E}_t \left\|\bar{\mathbf{x}}_t - \mathbf{x}^\star - \frac{\eta}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m) + \frac{\eta}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m) - \frac{\eta}{M}\sum_{m=1}^M g_t^m\right\|^2 \\
&\overset{\text{(Lemma 10)}}{\leq} \left\|\bar{\mathbf{x}}_t - \mathbf{x}^\star - \frac{\eta}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m)\right\|^2 + \frac{\eta^2\sigma^2}{M} \\
&= \left\|\bar{\mathbf{x}}_t - \mathbf{x}^\star\right\|^2 + \underbrace{\eta^2\left\|\frac{1}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m)\right\|^2}_{\dagger} \underbrace{- \frac{2\eta}{M}\sum_{m=1}^M \left\langle \bar{\mathbf{x}}_t - \mathbf{x}^\star, \nabla f_m(\mathbf{x}_t^m)\right\rangle}_{\dagger\dagger} + \frac{\eta^2\sigma^2}{M} \qquad (26)
\end{aligned}$$

For the term † we have:

$$\begin{aligned}
\dagger &= \eta^2\left\|\frac{1}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t) + \nabla f(\bar{\mathbf{x}}_t)\right\|^2 \\
&\overset{\text{(Lemma 5)}}{\leq} 2\eta^2\left\|\frac{1}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t)\right\|^2 + 2\eta^2\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \qquad (27)
\end{aligned}$$

For the first term in (27) we have:

$$\begin{aligned}
&2\eta^2\left\|\frac{1}{M}\sum_{m=1}^M \nabla f_m(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t)\right\|^2 \\
&= 2\eta^2\left\|\frac{1}{M}\sum_{m=1}^M \Big(\nabla f_m(\mathbf{x}_t^m) - \nabla f(\mathbf{x}_t^m) + \nabla f(\bar{\mathbf{x}}_t) - \nabla f_m(\bar{\mathbf{x}}_t)\Big) + \frac{1}{M}\sum_{m=1}^M \nabla f(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t)\right\|^2 \\
&\overset{\text{(Lemma 4,5)}}{\leq} \frac{4\eta^2}{M}\sum_{m=1}^M \left\|\nabla f_m(\mathbf{x}_t^m) - \nabla f(\mathbf{x}_t^m) + \nabla f(\bar{\mathbf{x}}_t) - \nabla f_m(\bar{\mathbf{x}}_t)\right\|^2 + 4\eta^2\left\|\frac{1}{M}\sum_{m=1}^M \nabla f(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t)\right\|^2 \\
&\leq 4\eta^2\delta_h^2\Xi_t + 4\eta^2\left\|\frac{1}{M}\sum_{m=1}^M \nabla f(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t)\right\|^2
\end{aligned}$$

For the second term in the above inequality we have:

$$
\begin{aligned}
&4\eta^2\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla f(\mathbf{x}_t^m)-\nabla f(\bar{\mathbf{x}}_t)\Big\|^2\\
&=4\eta^2\Big\|\frac{1}{M}\sum_{m=1}^{M}\Big(\nabla f(\mathbf{x}_t^m)-\nabla f(\bar{\mathbf{x}}_t)-\nabla^2 f(\bar{\mathbf{x}}_t)^\top(\mathbf{x}_t^m-\bar{\mathbf{x}}_t)\Big)+\underbrace{\frac{1}{M}\sum_{m=1}^{M}\nabla^2 f(\bar{\mathbf{x}}_t)^\top(\mathbf{x}_t^m-\bar{\mathbf{x}}_t)}_{=0}\Big\|^2\\
&\overset{(\text{Lemma 4})}{\leq}\frac{4\eta^2}{M}\sum_{m=1}^{M}\Big\|\nabla f(\mathbf{x}_t^m)-\nabla f(\bar{\mathbf{x}}_t)-\nabla^2 f(\bar{\mathbf{x}}_t)^\top(\mathbf{x}_t^m-\bar{\mathbf{x}}_t)\Big\|^2\\
&\overset{(\text{Lemma 13})}{\leq}\frac{H^2\eta^2}{M}\sum_{m=1}^{M}\big\|\mathbf{x}_t^m-\bar{\mathbf{x}}_t\big\|^4\\
&\overset{(\text{Lemma 6})}{\leq}MH^2\eta^2\Xi_t^2
\end{aligned}
$$

For the second term in (27) we also have:

$$
\begin{aligned}
2\eta^2\big\|\nabla f(\bar{\mathbf{x}}_t)\big\|^2&=2\eta^2\big\|\nabla f(\bar{\mathbf{x}}_t)-\nabla f(\mathbf{x}^\star)\big\|^2\\
&\overset{(\text{Lemma 11})}{\leq}4\eta^2L\Big[f(\bar{\mathbf{x}}_t)-f(\mathbf{x}^\star)\Big]
\end{aligned}
$$

Now by putting everything together we can bound the term † as follows:

$$
\begin{aligned}
&\dagger\leq 4\eta^2\delta_h^2\Xi_t+2MH^2\eta^2\Xi_t^2+4\eta^2L\Big[f(\bar{\mathbf{x}}_t)-f(\mathbf{x}^\star)\Big]\\
&\overset{(\text{Lemma 14})}{\leq}12\eta^4\delta_h^2\sigma^2K+24\eta^4\delta_h^2\bar{\zeta}^2K^2+18\eta^6MH^2K^2\sigma^4+72\eta^6MH^2K^4\bar{\zeta}^4+4\eta^2L\Big[f(\bar{\mathbf{x}}_t)-f(\mathbf{x}^\star)\Big]
\end{aligned}
$$

Now we bound the term ††.

$$
\begin{aligned}
&\dagger\,\dagger\\
&=-\frac{2\eta}{M}\sum_{m=1}^{M}\Big\langle\bar{\mathbf{x}}_t-\mathbf{x}^\star,\nabla f_m(\mathbf{x}_t^m)\Big\rangle\\
&=-\frac{2\eta}{M}\sum_{m=1}^{M}\Big\langle\bar{\mathbf{x}}_t-\mathbf{x}^\star,\nabla f_m(\mathbf{x}_t^m)-\nabla f(\bar{\mathbf{x}}_t)+\nabla f(\bar{\mathbf{x}}_t)\Big\rangle\\
&=-\frac{2\eta}{M}\sum_{m=1}^{M}\Big\langle\bar{\mathbf{x}}_t-\mathbf{x}^\star,\nabla f(\bar{\mathbf{x}}_t)\Big\rangle-2\eta\Big\langle\bar{\mathbf{x}}_t-\mathbf{x}^\star,\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\mathbf{x}_t^m)-\nabla f(\bar{\mathbf{x}}_t)\Big\rangle\\
&\overset{(4,\text{Lemma 7})}{\leq}-2\eta\Big[f(\bar{\mathbf{x}}_t)-f(\mathbf{x}^\star)\Big]-\eta\mu\big\|\bar{\mathbf{x}}_t-\mathbf{x}^\star\big\|^2+\eta\mu\big\|\bar{\mathbf{x}}_t-\mathbf{x}^\star\big\|^2+\frac{\eta}{\mu}\Big\|\frac{1}{M}\sum_{m=1}^{M}\nabla f_m(\mathbf{x}_t^m)-\nabla f(\bar{\mathbf{x}}_t)\Big\|^2
\end{aligned}
$$

$$
\begin{aligned}
&= -2\eta\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] + \frac{\eta}{\mu}\Big\|\frac{1}{M}\sum_{m=1}^{M}\Big(\nabla f_m(\mathbf{x}_t^m) - \nabla f(\mathbf{x}_t^m) + \nabla f(\bar{\mathbf{x}}_t) - \nabla f_m(\bar{\mathbf{x}}_t)\Big) + \\
&\qquad\qquad\qquad \frac{1}{M}\sum_{m=1}^{M}\nabla f(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t)\Big\|^2 \\
&\overset{\text{(Lemma 12)}}{\leq} -2\eta\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] + \frac{2\eta\delta_h^2}{\mu}\varXi_t + \frac{\eta}{\mu}\Big\|\frac{1}{M}\sum_{m=1}^{M}\Big(\nabla f(\mathbf{x}_t^m) - \nabla f(\bar{\mathbf{x}}_t) - \nabla^2 f(\bar{\mathbf{x}}_t)^\top(\mathbf{x}_t^m - \bar{\mathbf{x}}_t)\Big) + \\
&\qquad\qquad\qquad \underbrace{\frac{1}{M}\sum_{m=1}^{M}\nabla^2 f(\bar{\mathbf{x}}_t)^\top(\mathbf{x}_t^m - \bar{\mathbf{x}}_t)}_{=0}\Big\|^2 \\
&\overset{\text{(Lemma 4,13)}}{\leq} -2\eta\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] + \frac{2\eta\delta_h^2}{\mu}\varXi_t + \frac{\eta M H^2}{4\mu}\varXi_t^2 \\
&\overset{\text{(Lemma 14)}}{\leq} -2\eta\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] + \frac{6}{\mu}\eta^3\delta_h^2 K\sigma^2 + \frac{12}{\mu}\eta^3\delta_h^2 K^2\bar{\zeta}^2 + \frac{5}{\mu}\eta^5 M H^2 K^2\sigma^4 + \frac{18}{\mu}\eta^5 M H^2 K^4\bar{\zeta}^4
\end{aligned}
$$

Now we plug † and †† into (26) and setting $\eta \leq \frac{1}{4L}$ and taking the unconditional expectation we have:

$$
\begin{aligned}
&\eta\,\mathbb{E}\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] \\
&\leq \mathbb{E}\left[\big\|\bar{\mathbf{x}}_t - \mathbf{x}^\star\big\|^2 - \big\|\bar{\mathbf{x}}_{t+1} - \mathbf{x}^\star\big\|^2\right] + \frac{\eta^2\sigma^2}{M} + 12\eta^4\delta_h^2\sigma^2 K + 24\eta^4\delta_h^2 K^2 + 18\eta^6 M H^2 K^2\sigma^4 + 72\eta^6 M H^2 K^4\bar{\zeta}^4 + \\
&\quad \frac{6}{\mu}\eta^3\delta_h^2 K\sigma^2 + \frac{12}{\mu}\eta^3\delta_h^2 K^2\bar{\zeta}^2 + \frac{5}{\mu}\eta^5 M H^2 K^2\sigma^4 + \frac{18}{\mu}\eta^5 M H^2 K^4\bar{\zeta}^4
\end{aligned}
$$

Now we divide by ηT and sum over t and we have:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[f(\bar{\mathbf{x}}_t) - f(\mathbf{x}^\star)\Big] \leq \frac{B^2}{\eta T} + \frac{\eta\sigma^2}{M} &+ 12\eta^3\delta_h^2\sigma^2 K + 24\eta^3\delta_h^2 K^2\bar{\zeta}^2 + 18\eta^5 M H^2 K^2\sigma^4 + 72\eta^5 M H^2 K^4\bar{\zeta}^4 + \\
&\frac{6}{\mu}\eta^2\delta_h^2 K\sigma^2 + \frac{12}{\mu}\eta^2\delta_h^2 K^2\bar{\zeta}^2 + \frac{5}{\mu}\eta^4 M H^2 K^2\sigma^4 + \frac{18}{\mu}\eta^4 M H^2 K^4\bar{\zeta}^4
\end{aligned}
$$

With this choice of learning rate

$$
\begin{aligned}
\eta = c_1.\min\Bigg\{&\frac{1}{L}, \frac{B\sqrt{M}}{\sigma\sqrt{T}}, \Big(\frac{B^2}{\delta_h^2\sigma^2 KT}\Big)^{1/4}, \Big(\frac{B^2}{\delta_h^2\bar{\zeta}^2 K^2 T}\Big)^{1/4}, \Big(\frac{B^2}{M H^2\sigma^4 K^2 T}\Big)^{1/6}, \Big(\frac{B^2}{M H^2\bar{\zeta}^4 K^4 T}\Big)^{1/6}, \\
&\Big(\frac{\mu B^2}{\delta_h^2\sigma^2 KT}\Big)^{1/3}, \Big(\frac{\mu B^2}{\delta_h^2\bar{\zeta}^2 K^2 T}\Big)^{1/3}, \Big(\frac{\mu B^2}{M H^2\sigma^4 K^2 T}\Big)^{1/5}, \Big(\frac{\mu B^2}{M H^2\bar{\zeta}^4 K^4 T}\Big)^{1/5}\Bigg\}
\end{aligned}
$$

We get a rate of:

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[f(\bar{\mathbf{x}}_t)-f(\mathbf{x}^\star)\Big] \\ &\leq c_2\Bigg(\frac{LB^2}{KR}+\frac{\sigma B}{\sqrt{MKR}}+\frac{(B^6\delta_h^2\sigma^2)^{1/4}}{K^{2/4}R^{3/4}}+\frac{(B^6\delta_h^2\bar{\zeta}^2)^{1/4}}{K^{1/4}R^{2/3}}+\frac{(B^{10}MH^2\sigma^4)^{1/6}}{K^{3/6}R^{5/6}}+ \\ &\qquad\frac{(B^{10}MH^2\bar{\zeta}^4)^{1/6}}{K^{1/6}R^{5/6}}+\frac{(B^4\delta_h^2\sigma^2)^{1/3}}{\mu^{1/3}K^{1/3}R^{2/3}}+\frac{(B^4\delta_h^2\bar{\zeta}^2)^{1/3}}{\mu^{1/3}R^{2/3}}+\frac{(B^8MH^2\sigma^4)^{1/5}}{\mu^{1/5}K^{2/5}R^{4/5}}+\frac{(B^8MH^2\bar{\zeta}^4)^{1/5}}{\mu^{1/5}R^{4/5}}\Bigg)\end{aligned} \tag{28}$$

Where c_1 and c_2 are absolute constant numbers that are used for simplifying the rates and dropping some constant numbers in the proof. They don't have any effect on the convergence rate. ■

A.2. Proof of Theorem 3

Proof For the proof of this section, we assume that our quadratic functions are in the form of:

$$\begin{aligned}f_m(\mathbf{x})&=\frac{1}{2}\mathbf{x}^\top A\mathbf{x}+\mathbf{b}_m^\top\mathbf{x}+\mathbf{d}_m \\ f(\mathbf{x})&=\frac{1}{2}\mathbf{x}^\top A\mathbf{x}+\bar{\mathbf{b}}^\top\mathbf{x}+\bar{\mathbf{d}}\end{aligned} \tag{29}$$

Where $\bar{\mathbf{b}}=\frac{1}{M}\sum_{m=1}^M\mathbf{b}_m$, $\bar{\mathbf{d}}=\frac{1}{M}\sum_{m=1}^M\mathbf{d}_m$ and $A\in\mathbb{R}^{d\times d}$ is an arbitrary symmetric matrix and $\mathbf{x},\mathbf{b},\mathbf{d}\in\mathbb{R}^d$.

We start by the L-smoothness property of global function and by using the fact that $\bar{\mathbf{x}}_{t+1}=\bar{\mathbf{x}}_t-\frac{\eta}{M}\sum_{m=1}^M g_t^m$ we have:

$$\begin{aligned}f(\bar{\mathbf{x}}_{t+1})&\leq f(\bar{\mathbf{x}}_t)+\nabla f(\bar{\mathbf{x}}_t)^\top(\bar{\mathbf{x}}_{t+1}-\bar{\mathbf{x}}_t)+\frac{L}{2}\big\|\bar{\mathbf{x}}_{t+1}-\bar{\mathbf{x}}_t\big\|^2 \\ &=f(\bar{\mathbf{x}}_t)-\nabla f(\bar{\mathbf{x}}_t)^\top\frac{\eta}{M}\sum_{m=1}^M g_t^m+\frac{\eta^2L}{2}\Big\|\frac{1}{M}\sum_{m=1}^M g_t^m\Big\|^2\end{aligned}$$

By taking the conditional expectation on the previous iterate $\mathbf{x}_t^m,\forall m\in[M]$ we have:

$$\begin{aligned}\mathbb{E}_t[f(\bar{\mathbf{x}}_{t+1})]&\leq f(\bar{\mathbf{x}}_t)-\nabla f(\bar{\mathbf{x}}_t)^\top\frac{\eta}{M}\sum_{m=1}^M\nabla f_m(\mathbf{x}_t^m)+\frac{\eta^2L}{2}\,\mathbb{E}_t\Big\|\frac{1}{M}\sum_{m=1}^M g_t^m\Big\|^2 \\ &\overset{\text{(Lemma 10)}}{\leq} f(\bar{\mathbf{x}}_t)-\nabla f(\bar{\mathbf{x}}_t)^\top\frac{\eta}{M}\sum_{m=1}^M\nabla f_m(\mathbf{x}_t^m)+\frac{\eta^2L}{2}\Big\|\frac{1}{M}\sum_{m=1}^M\nabla f_m(\mathbf{x}_t^m)\Big\|^2+\frac{\eta^2\sigma^2L}{2M} \\ &=f(\bar{\mathbf{x}}_t)-\eta\nabla f(\bar{\mathbf{x}}_t)^\top\underbrace{(A\bar{\mathbf{x}}_t+\bar{b})}_{=\nabla f(\bar{\mathbf{x}}_t)}+\frac{\eta^2L}{2}\|\underbrace{A\bar{\mathbf{x}}_t+\bar{b}}_{=\nabla f(\bar{\mathbf{x}}_t)}\|^2+\frac{\eta^2\sigma^2L}{2M} \\ &=f(\bar{\mathbf{x}}_t)-\eta\big\|\nabla f(\bar{\mathbf{x}}_t)\big\|^2+\frac{\eta^2L}{2}\big\|\nabla f(\bar{\mathbf{x}}_t)\big\|^2+\frac{\eta^2\sigma^2L}{2M}\end{aligned}$$

By choosing $\eta \leq \frac{1}{L}$, rearranging the terms and taking the unconditional expectation we have:

$$\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \leq \frac{2}{\eta}\,\mathbb{E}\Big[f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})\Big] + \frac{\eta L \sigma^2}{M}$$

Then we sum over t and divide by T and we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \leq \frac{2\mathit{\Delta}}{\eta T} + \frac{\eta L \sigma^2}{M}$$

By the choice of learning rate in the following way we get:

$$\eta = \min\left\{\frac{1}{L}, \sqrt{\frac{M\mathit{\Delta}}{LT\sigma^2}}\right\}$$

Which results the following convergence rate:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|^2 \leq \frac{L\mathit{\Delta}}{KR} + \frac{\sigma\sqrt{L\mathit{\Delta}}}{\sqrt{MKR}}$$

■