# On the Convergence of LocalSGD on Non-Convex Non-I.I.D. Functions

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

In federated learning (FL), by taking local steps, LocalSGD is a strong baseline method with significant communication saving in non-convex optimization McMahan et al. [2017]. However, the theoretical analysis of LocalSGD shows quite the opposite result, in which the rate of LocalSGD can match the rate of MbSGD (without local steps) only under very strict conditions.

In this work, we showed (for the first time) that:

1. LocalSGD can benefit from Hessian similarity, which yields an *improved* conditioning.
2. LocalSGD can converge provably faster than MbSGD for a class of *non-convex* functions.
3. LocalSGD can converge provably faster than MbSGD *without* dependency on uniform gradient similarity.

## 1 Introduction

We are interested in the problem class of [1]

| | | |
|---|---|---|
| **Model:** | 1. $\min_{\mathbf{x}\in\mathbb{R}^d} \left[ f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right]$ <br> 2. $f_i(\mathbf{x}) \in C_L^{1,1}(\mathbb{R}^d)$, $i \in [n]$ <br> 3. $f(\mathbf{x})$ is bounded below | (1) |
| **Oracle:** | $\mathcal{SO}$ | |
| **$\varepsilon$-solution:** | $\mathbb{E}\left\|\nabla f(\hat{\mathbf{x}})\right\|_2^2 \le \varepsilon$ | |

We assume that problem (1) is to be solved by iterative algorithms via subsequent calls to the stochastic oracle $\mathcal{SO}$. Specifically, at iteration $t \in [T]$ of the algorithm, $\left(\mathbf{x}_t^1, \cdots, \mathbf{x}_t^n\right) \in \mathbb{R}^{d\times n}$ being the input, the $\mathcal{SO}$ outputs vectors

$$\left(\mathbf{g}_t^1, \cdots, \mathbf{g}_t^n\right) := \left(G_1(\mathbf{x}_t^1, \xi_t^1), \cdots, G_n(\mathbf{x}_t^n, \xi_t^n)\right) \in \mathbb{R}^{d\times n}, \tag{2}$$

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

$$\mathbb{E}_{\xi_t^i}\left[G_i(\mathbf{x}, \xi_t^i)\right] = \nabla f_i(\mathbf{x}),\ \ \mathbb{E}_{\xi_t^i}\left\|G_i(\mathbf{x}, \xi_t^i) - \nabla f_i(\mathbf{x})\right\|_2^2 \le \sigma^2. \tag{3}$$

[1] Let $C_\ell^{m,p}(\mathbb{R}^d)$ denote the class of (possibly non-convex) functions on $\mathbb{R}^d$ which are $m$th-continuously differentiable and have $\ell$-Lipschitz continuous $p$th-order derivatives under $\|\cdot\|_2$.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

For simplicity, let $\mathbf{x}_0^1 = \cdots = \mathbf{x}_0^n$ and $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^i$. Assume that $f(\bar{\mathbf{x}}_0) - f^* \leq \varDelta$, where $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$.

We're interested in the analysis of LocalSGD [Stich, 2018, Lin et al., 2018, Woodworth et al., 2020a,b]:

$$\mathbf{x}_{t+1}^i = \begin{cases} \bar{\mathbf{x}}_{t-\tau+1} - \frac{1}{n}\sum_{j\in[n]}\sum_{k=0}^{\tau-1}\eta_{t-k}\mathbf{g}_{t-k}^j & \text{if } t+1 = r\tau, \\ \mathbf{x}_t^i - \eta_t \mathbf{g}_t^i & \text{otherwise,} \end{cases} \tag{4}$$

where $\tau$ is often referred to as the *communication interval*. Another baseline algorithm used for comparison is MinibatchSGD (*a.k.a.* MbSGD):

$$\mathbf{x}_{t+1}^i = \begin{cases} \bar{\mathbf{x}}_{t-\tau+1} - \frac{\widetilde{\eta}_t}{\tau n}\sum_{j\in[n]}\sum_{k=0}^{\tau-1}\mathbf{g}_{t-k}^j & \text{if } t+1 = r\tau, \\ \mathbf{x}_t^i & \text{otherwise.} \end{cases} \tag{5}$$

For both algorithms, they return $\hat{\mathbf{x}} \sim Uniform\left\{\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \cdots, \bar{\mathbf{x}}_{T-1}\right\}$. *W.l.o.g.*, let $T = \tau R$. The main concern of this paper is the (asymptotic) *communication complexity* (*i.e.*, the scale of $R$) for the algorithms.

### 1.1 Assumptions

For theoretical analyses of the algorithms, the following assumptions on gradient similarity and Hessian similarity are conventionally made in the literature [Koloskova et al., 2020, Karimireddy et al., 2020, Patel et al., 2022]:

**Assumption 1** (($\zeta, \bar{\zeta}$)-GS)**.** $0 \leq \bar{\zeta} \leq \sqrt{2L\varDelta}$, $\bar{\zeta} \leq \zeta \leq \sqrt{n}\bar{\zeta}$. *For any* $i \in [n]$,

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\right\|_2 \leq \zeta, \tag{6}$$

*or for any* $\mathbf{x} \in \mathbb{R}^d$,

$$\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\right\|_2^2 \leq \bar{\zeta}^2. \tag{7}$$

**Assumption 2** ($\delta$-HS)**.** $0 \leq \delta \leq 2L$. *For any* $i \in [n]$, $f_i \in C_L^{2,1}(\mathbb{R}^d)$, *and*

$$\sup_{\mathbf{x}\in\mathbb{R}^d} \left\|\nabla^2 f_i(\mathbf{x}) - \nabla^2 f(\mathbf{x})\right\|_2 \leq \delta. \tag{8}$$

*Further,* $\mathcal{M} \geq 0$ *s.t. there exists a function in* $\mathbf{conv}\{f_1, \cdots, f_n\}$ *with* $\mathcal{M}$*-Lipschitz continuous Hessian.*

For technical reasons, in Assumption 2, a (relatively weak) assumption of higher-order smoothness is required in some of our results. Indeed, this higher-order smoothness assumption is significantly weaker than the uniform higher-order smoothness, *i.e.*, $f_i \in C_{\mathcal{M}}^{2,2}(\mathbb{R}^d)$, $i \in [n]$, and is also weaker than the higher-order smoothness of $f$, *i.e.*, $f \in C_{\mathcal{M}}^{2,2}(\mathbb{R}^d)$.

Also, in some of our results, we need the following (relatively weak) assumption of weak convexity:

**Assumption 3** ($\rho$-weak convexity)**.** $0 \leq \rho \leq L$. *There exists a* $\rho$*-weakly convex function*[2] *in* $\mathbf{conv}\{f_1, \cdots, f_n\}$.

### 1.2 Related Work

#### LocalSGD and its analyses

- Under different names, LocalSGD was widely proposed and explored in the literature [Bijral et al., 2017, Zhang et al., 2016, McMahan et al., 2017].
- Early works [Stich, 2018, Lin et al., 2018, Woodworth et al., 2020a] analyzed LocalSGD for i.i.d. functions.

[2]Function $g$ is weakly convex if $g(\mathbf{x}) + \frac{\rho}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}$ is convex.

- Under Assumption 1, Koloskova et al. [2020] analyzed LocalSGD for convex and non-convex functions. In their analysis, the statistics term is *optimal*, and the heterogeneity term is proven to be *tight*.[3] But, the communication complexity of LocalSGD at their bests can only match MbSGD under the (standard) conditioning of

$$\bar{\zeta}^2 = \mathcal{O}(1/R). \tag{9}$$

- Under Assumption 1, Woodworth et al. [2020b] showed that, for convex functions, the optimization term in [Karimireddy et al., 2020, Koloskova et al., 2020] can be improved, so that the communication complexity of LocalSGD can surpass MbSGD under the (stronger) conditioning of

$$\zeta^2 = \mathcal{O}(1/R). \tag{10}$$

- Wang et al. [2022] proposed a different heterogeneity measure $\hat{\rho}$ (*cf.* Equation (15) in their paper), namely "average drift at optimum". They claimed that their measure $\hat{\rho} \approx 0$, and when $\sigma = 0$, they obtained a superfast $\mathcal{O}(1)$ rate for strongly convex functions. However, it's not clear how close the claimed approximation is, and whether it can be generalized to stochastic and non-convex settings.
- Under Assumption 1, Karimireddy et al. [2020], Yang et al. [2021] generalized LocalSGD with different local and global stepsizes, and used the analysis of 'merged local updates'. But it's believed that, at least under their analysis, the algorithm is uninterestingly reduced to (an inferior version of) MbSGD (*cf.* the discussions in Appendix G of [Woodworth et al., 2020b] or the comments in Appendix B of this paper).

**"Variance Reduction" algorithms relieving Assumption 1**

- SCAFFOLD is proposed by [Karimireddy et al., 2020] for both convex and non-convex functions. They showed that, without any assumptions on similarity, the communication complexity of SCAFFOLD can match MbSGD. They further made Assumption 2, under which the communication complexity of SCAFFOLD can surpass MbSGD, though for quadratic functions only.
- ProxSkip [Mishchenko et al., 2022] is a non-deterministic algorithm for strongly-convex functions. They showed by a different technique that, without any assumptions on similarity, the communication complexity of ProxSkip can surpass MbSGD. Their optimization term is *optimal*. However, their statistics term is not optimal (no linear speedup), and moreover, it's not clear whether their techniques can be generalized to non-convex settings.
- Beyond the problem (1) of our interests, CE-LSGD [Patel et al., 2022] is an algorithm for non-convex functions using a stronger two-point stochastic oracle that requires the stronger $L$-mean Lipschitz continuity of $\nabla G(\cdot, \xi_t^i)$. Under Assumption 2, the communication complexity of CE-LSGD can surpass MbSGD. However, as long as $\sigma > 0$, the optimization term in their analysis is not optimal.

**Open problems for distributed non-convex optimization**

1. Is the optimization term tight in the analyses in [Karimireddy et al., 2020, Koloskova et al., 2020] of LocalSGD, under Assumption 1?
2. Can LocalSGD benefit from higher-order similarity?
3. Whether LocalSGD can match (or surpass) MbSGD under a weaker conditioning than Equation (9) (or Equation (10))?
4. Is there any first-order deterministic algorithm that gives *optimal* optimization and statistics terms, and can surpass MbSGD (not only for quadratic functions)?

### 1.3 Contributions

We showed that, for *non-convex functions* under Assumption 1,

[3] Throughout the paper, "tightness" means the term in the upper bound of the analysis of the specific algorithm cannot be improved, while "optimal" means the term cannot be improved for any algorithm under the same setting.

1. The optimization term is *tight* in the analyses in [Karimireddy et al., 2020, Koloskova et al., 2020] of LocalSGD.
2. With the benefits from Assumption 2, the heterogeneity term can be provably improved, so that LocalSGD can match MbSGD under a much *weaker* conditioning of

$$\delta^2\bar{\zeta}^2 = \mathcal{O}(1/R) \text{ and } \bar{\zeta}^4 = \mathcal{O}(1/R). \tag{11}$$

   This is the first work, to the best of our knowledge, to show that LocalSGD can also *benefit* from Hessian similarity.
3. Under Assumption 3, we show (for the first time) for a class of *non-convex* functions that LocalSGD can surpass MbSGD under the conditioning of Equation (10), as an extension of the results in [Woodworth et al., 2020b] for convex functions.
4. Under Assumption 2, 3, we show (for the first time) that LocalSGD can be provably faster than MbSGD *only* with dependency of $\bar{\zeta}$.

# 2 Theory

## 2.1 Convergence analysis

**Theorem 1** (non-convex functions)**.** *Let's make Assumption 1. For Equation* (4)*, in each of the below cases, there exists some value of $\eta$, s.t.*

*0. cf. [Koloskova et al., 2020, Karimireddy et al., 2020, Yang et al., 2021],*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \le \mathcal{O}\left(\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{L\Delta\bar{\zeta}}{R}\right)^{\frac{2}{3}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}\right); \tag{12}$$

*1. under Assumption 2,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \le \mathcal{O}\left(\frac{L\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + K(T)\right); \tag{13}$$

*2. under Assumption 3,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \le \mathcal{O}\left(\left(\frac{L}{\tau} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{L\Delta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(L\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}}\right); \tag{14}$$

*3. under Assumption 2, 3,*[4]

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \le \mathcal{O}\left(\left(\frac{L}{\tau} + \rho + \delta + (\mathcal{M}\bar{\zeta})^{\frac{1}{2}} + \frac{(\mathcal{M}\sigma)^{\frac{1}{2}}}{\tau^{\frac{1}{4}}}\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + K(T)\right). \tag{16}$$

*In above formulas,*

$$K(T) := \left(\frac{\delta\Delta\bar{\zeta}}{R}\right)^{\frac{2}{3}} + \frac{(\delta\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}} + \left(\frac{\mathcal{M}^2\Delta^4\bar{\zeta}^4}{R^4}\right)^{\frac{1}{5}} + \frac{(\mathcal{M}^2\Delta^4\sigma^4)^{\frac{1}{5}}}{\tau^{\frac{2}{5}}R^{\frac{4}{5}}}. \tag{17}$$

**Remark 1** (communication complexity)**.** *For simplicity, let $L = \Delta = \mathcal{M} = \sigma = 1$ and $\tau = \infty$. It's well known that, for general non-convex functions (and also for $\rho$-weakly convex functions),* MbSGD *gets a $\mathcal{O}\left(\frac{1}{R}\right)$-substationary point.*

[4] This is far from trivially merging Equation (13) and Equation (14), which would yield, instead,

$$\mathcal{O}\left(\left(\frac{L}{\tau} + \rho\right)\frac{\Delta}{R} + \sqrt{\frac{L\Delta\sigma^2}{n\tau R}} + \left(\frac{\delta\Delta\zeta}{R}\right)^{\frac{2}{3}} + \frac{(\delta\Delta\sigma)^{\frac{2}{3}}}{\tau^{\frac{1}{3}}R^{\frac{2}{3}}} + \left(\frac{\mathcal{M}^2\Delta^4\zeta^4}{R^4}\right)^{\frac{1}{5}} + \frac{(\mathcal{M}^2\Delta^4\sigma^4)^{\frac{1}{5}}}{\tau^{\frac{2}{5}}R^{\frac{4}{5}}}\right). \tag{15}$$

1. *For general non-convex functions with Hessian similarity, in Equation (13), we get a*

$$\mathcal{O}\left(\frac{1}{R}+\left(\frac{\delta\bar{\zeta}}{R}\right)^{2/3}+\left(\frac{\bar{\zeta}}{R}\right)^{4/5}\right) \tag{18}$$

*-substationary point. Therefore, with benefit from Hessian similarity,* LocalSGD *has matching rate to* MbSGD *under the conditioning of Equation (11), which is much* weaker *than the conditioning of Equation (9) required in [Koloskova et al., 2020, Karimireddy et al., 2020, Yang et al., 2021].*

2. *For $\rho$-weakly convex function, in Equation (14), we get a*

$$\mathcal{O}\left(\frac{\rho}{R}+\left(\frac{\zeta}{R}\right)^{2/3}\right) \tag{19}$$

*-substationary point. Therefore, under the conditioning of Equation (10), our analysis shows that* LocalSGD *can* surpass MbSGD *for a class of* non-convex *functions.*

3. *For $\rho$-weakly convex function with Hessian similarity, in Equation (16), we get a*

$$\mathcal{O}\left(\frac{\rho+\delta+\sqrt{\bar{\zeta}}}{R}+\left(\frac{\delta\bar{\zeta}}{R}\right)^{2/3}+\left(\frac{\bar{\zeta}}{R}\right)^{4/5}\right) \tag{20}$$

*-substationary point. Therefore, under the conditioning of Equation (11), our analysis shows that* LocalSGD *can* surpass MbSGD without *dependency on $\zeta$.*

**Remark 2** (linear speedup)**.** *For simplicity, let $L = \Delta = \mathcal{M} = \sigma = 1$. For a fixed communication interval $\tau = \mathcal{O}(1)$, when*

$$T = \Omega\left(\delta^4 n^3 + n^{5/3}\right), \tag{21}$$

*for general non-convex functions (and also for $\rho$-weakly convex functions), according to Theorem 1,* LocalSGD *achieves a linear speedup with respect to the number of workers,* i.e., *the statistics term dominates and yields a*

$$\mathcal{O}\left(K_1(T)\right) = \mathcal{O}\left(\frac{1}{\sqrt{nT}}\right) \tag{22}$$

*-stationary point. The conditioning of Equation (21) for linear speedup is weaker than the previous s.o.t.a. conditioning of $T = \Omega\left(n^3\right)$ (cf. [Yu et al., 2019]).*

**Remark 3.** *Karimireddy et al. [2020] showed that for quadratic functions, Assumption 2, 3 can yield faster convergence of* SCAFFOLD*, and posed "a challenging open problem" of its generalization to non-quadratic functions. In Equation (16), we successfully generalized the result to non-quadratic functions for even the baseline algorithm of* LocalSGD*.*

## 2.2 Proof sketch

**Lemma 4** (descent lemma)**.** *For Equation (4), under Assumption 2, for $\eta_t \leq \frac{1}{L}$, we have*

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_{t+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \frac{\eta_t}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 + \eta_t^2\frac{L\sigma^2}{2n} + \frac{\eta_t}{2}\mathbb{E}\left[8\delta^2\Xi_t + \frac{\mathcal{M}^2}{2}\Xi_t^2\right], \tag{23}$$

*where $\Xi_t = \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t\right\|_2^2$.*

**Lemma 5** (distance lemma)**.** *Let $k \in [0, \tau-1]$ s.t. $t-k$ is a multiple of $\tau$, and let $\eta_{t-k} = \cdots = \eta_t := \eta$. For Equation (4), under Assumption 1, 2, 3, for*

$$\eta \leq \min\left\{\frac{1}{36\rho\tau}, \frac{1}{17\delta\tau}, \frac{1}{6\sqrt{\mathcal{M}\bar{\zeta}}\tau}, \frac{1}{5\sqrt{\mathcal{M}\sigma}\tau^{\frac{3}{4}}}\right\}, \tag{24}$$

*we have*

$$\Xi_t \leq 36\eta^2\tau^2\bar{\zeta}^2 + 6\eta^2\tau\sigma^2. \tag{25}$$

*Proof sketch of Equation* (16). Plugging Equation (25) into Equation (23), we have

$$\mathbb{E}\left[f(\bar{\mathbf{x}}_{t+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \frac{\eta}{2}\mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 + \eta^2\frac{L\sigma^2}{2n} + 144\eta^3\delta^2\tau^2\bar{\zeta}^2 + 24\eta^3\delta^2\tau\sigma^2 + 648\eta^5\mathcal{M}^2\tau^4\bar{\zeta}^4 + 18\eta^5\mathcal{M}^2\tau^2\sigma^4. \tag{26}$$

After summing Equation (26) over $t$ from 0 to $T-1$ and dividing by $T/2$, Equation (16) follows immediately from the following choice of $\eta$:

$$\eta_t \equiv \eta := \min\left\{\frac{1}{L}, \frac{1}{36\rho\tau}, \frac{1}{17\delta\tau}, \frac{1}{6\sqrt{\mathcal{M}\bar{\zeta}}\tau}, \frac{1}{5\sqrt{\mathcal{M}\sigma}\tau^{\frac{3}{4}}}, \sqrt{\frac{2n\varDelta}{L\tau R\sigma^2}}, \left(\frac{\varDelta}{144\delta^2\tau^3R\bar{\zeta}^2}\right)^{\frac{1}{3}}, \left(\frac{\varDelta}{24\delta^2\tau^2R\sigma^2}\right)^{\frac{1}{3}}, \left(\frac{\varDelta}{648\mathcal{M}^2\tau^5R\bar{\zeta}^4}\right)^{\frac{1}{5}}, \left(\frac{\varDelta}{18\mathcal{M}^2\tau^3R\sigma^4}\right)^{\frac{1}{5}}\right\}. \tag{27}$$

□

## 3 Conclusion

| Method | Benefit from HS | Conditioning | Improvement over MbSGD | Non-convex |
|---|---|---|---|---|
| LocalSGD [Koloskova et al., 2020] | ✗ | (9) **standard** | ✗ | ✓ |
| LocalSGD [Woodworth et al., 2020b] | ✗ | (10) **worse** | ✓ | ✗ |
| SCAFFOLD [Karimireddy et al., 2020] | ✓ | variance reduction | ✗<br>✓ | ✓<br>Quadratic |
| LocalSGD **(Ours)** | ✓ | (11) **better** | ✓ | ✓ |

Table 1: Comparisons over related work

# References


Avleen S. Bijral, Anand D. Sarwate, and Nathan Srebro. Data-dependent convergence for consensus stochastic optimization. *IEEE Transactions on Automatic Control*, 62(9):4483–4498, 2017. doi: 10.1109/TAC.2017.2671377.

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.

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.

Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

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.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! *arXiv preprint arXiv:2202.09357*, 2022.

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

Kumar Kshitij Patel, Lingxiao Wang, Blake Woodworth, Brian Bullins, and Nathan Srebro. Towards optimal communication complexity in distributed non-convex optimization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=SNElc7QmMDe.

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

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.

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, 2020a.

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, 2020b.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Hao Yu, Sen Yang, and Shenghuo Zhu. 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, 2019.

Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel sgd: When does averaging help?, 2016.

# A Techniques

## A.1 Basic techniques

**Lemma 6** (Young's inequality)**.** *For* $\gamma > 0$*,*

$$\|\mathbf{x}+\mathbf{y}\|_2^2 \le (1+\gamma)\|\mathbf{x}\|_2^2 + (1+\gamma^{-1})\|\mathbf{y}\|_2^2 . \tag{28}$$

**Lemma 7.** *Under Assumption 2, we have, for any* $i \in [n]$*,*

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y}) - \nabla f(\mathbf{x}) + \nabla f(\mathbf{y})\|_2 \le \delta \|\mathbf{x}-\mathbf{y}\|_2 . \tag{29}$$

*Proof.* It follows from the $\delta$-Lipschitz continuous gradient of $f_i - f$. ☐

## A.2 Proof of descent lemmas

**Lemma 8.** *Under Assumption 2, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}^i) - \nabla f(\bar{\mathbf{x}})\right\|_2^2 \le 8\delta^2 \varXi + \frac{\mathcal{M}^2}{2}\varXi^2, \tag{30}$$

*where* $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}^i$ *and* $\varXi = \frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}^i - \bar{\mathbf{x}}\right\|_2^2$*.*

*Proof.* Let $\hat{f} \in \mathbf{conv}\{f_1, \cdots, f_n\} \cap C^{2,2}_{\mathcal{M}}(\mathbb{R}^d)$. Then

$$\hat{f} = \frac{1}{n}\sum_{i=1}^n \alpha_i f_i, \quad \frac{1}{n}\sum_{i=1}^n \alpha_i = 1.\ \alpha_i \ge 0. \tag{31}$$

We have

$$
\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}^i) - \nabla f(\bar{\mathbf{x}}) \right\|_2^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \nabla^2 f_i(\bar{\mathbf{x}}_t + u(\mathbf{x}_t^i - \bar{\mathbf{x}}_t))(\mathbf{x}_t^i - \bar{\mathbf{x}}_t) du \right\|_2^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left[ \nabla^2 f_i(\bar{\mathbf{x}}_t + u(\mathbf{x}_t^i - \bar{\mathbf{x}}_t)) - \nabla^2 \hat{f}(\bar{\mathbf{x}}_t) \right] (\mathbf{x}_t^i - \bar{\mathbf{x}}_t) du \right\|_2^2 \\
&\leq \left[ \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left\| \nabla^2 f_i(\bar{\mathbf{x}}_t + u(\mathbf{x}_t^i - \bar{\mathbf{x}}_t)) - \nabla^2 \hat{f}(\bar{\mathbf{x}}_t) \right\|_2 \cdot \left\| \mathbf{x}_t^i - \bar{\mathbf{x}}_t \right\|_2 du \right]^2 \\
&\overset{(32)\text{-}❶}{\leq} \left[ \frac{1}{n} \sum_{i=1}^{n} \int_0^1 \left( 2\delta + \mathcal{M} u \left\| \mathbf{x}_t^i - \bar{\mathbf{x}}_t \right\|_2 \right) \cdot \left\| \mathbf{x}_t^i - \bar{\mathbf{x}}_t \right\|_2 du \right]^2 \\
&= \left[ 2\delta \left( \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{x}_t^i - \bar{\mathbf{x}}_t \right\|_2 \right) + \frac{\mathcal{M}}{2} \varXi_t \right]^2 \\
&\leq 8\delta^2 \left( \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{x}_t^i - \bar{\mathbf{x}}_t \right\|_2 \right)^2 + \frac{\mathcal{M}^2}{2} \varXi_t^2 \\
&\leq 8\delta^2 \varXi_t + \frac{\mathcal{M}^2}{2} \varXi_t^2,
\end{aligned} \tag{32}
$$

in which relation (32)-❶ follows from

$$
\begin{aligned}
&\left\| \nabla^2 f_i(\mathbf{x} + \mathbf{u}) - \nabla^2 \hat{f}(\mathbf{x}) \right\|_2 \\
&\leq \left\| \nabla^2 f_i(\mathbf{x} + \mathbf{u}) - \nabla^2 \hat{f}(\mathbf{x} + \mathbf{u}) \right\|_2 + \left\| \nabla^2 \hat{f}(\mathbf{x} + \mathbf{u}) - \nabla^2 \hat{f}(\mathbf{x}) \right\|_2 \\
&= \left\| \sum_{i=1}^{M} \alpha_i \left( \nabla^2 f_i(\mathbf{x} + \mathbf{u}) - \nabla^2 F_i(\mathbf{x} + \mathbf{u}) \right) \right\|_2 + \left\| \nabla^2 \hat{f}(\mathbf{x} + \mathbf{u}) - \nabla^2 \hat{f}(\mathbf{x}) \right\|_2 \\
&\leq \sum_{i=1}^{M} \alpha_i \cdot 2\delta + \mathcal{M} \left\| \mathbf{u} \right\|_2 \\
&= 2\delta + \mathcal{M} \left\| \mathbf{u} \right\|_2 .
\end{aligned} \tag{33}
$$

□

**Lemma 9** (descent lemmas)**.** *For Equation (4), for $\eta_t \leq \frac{1}{L}$, we have*

$$
\mathbb{E}\left[f(\bar{\mathbf{x}}_{t+1})\right] \leq \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \frac{\eta_t}{2} \mathbb{E} \left\| \nabla f(\bar{\mathbf{x}}_t) \right\|_2^2 + \eta_t^2 \frac{L\sigma^2}{2n} + \frac{\eta_t}{2} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t) \right\|_2^2, \tag{34}
$$

*where*

-
$$
\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t) \right\|_2^2 \leq L^2 \varXi_t; \tag{35}
$$

- *under Assumption 2,*
$$
\left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t) \right\|_2^2 \leq 8\delta^2 \varXi_t + \frac{\mathcal{M}^2}{2} \varXi_t^2. \tag{36}
$$

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

$$\begin{aligned}
&\mathbb{E}\left[f(\bar{\mathbf{x}}_{t+1})\right] \\
&\overset{(3)}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \eta_t \, \mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}_t), \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) \right\rangle + \eta_t^2 \frac{L}{2} \, \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) \right\|_2^2 + \eta_t^2 \frac{L}{2}\frac{\sigma^2}{n} \\
&\overset{(37)\text{-}❶}{\leq} \mathbb{E}\left[f(\bar{\mathbf{x}}_t)\right] - \frac{\eta_t}{2} \, \mathbb{E}\left\|\nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 + \frac{\eta_t}{2} \, \mathbb{E}\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t) \right\|_2^2 + \eta_t^2 \frac{L}{2}\frac{\sigma^2}{n},
\end{aligned} \tag{37}$$

where relation (37)-❶ follows from $\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{2}\left[\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2\right]$ and $\eta_t \leq \frac{1}{L}$.

Then, we have

$$\left\| \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t) \right\|_2^2 \leq \frac{1}{n}\sum_{i=1}^{n} \left\|\nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t)\right\|_2^2 \leq L^2 \Xi_t, \tag{38}$$

or Equation (30) under Assumption 2. □

### A.3 Proof of distance lemmas

**Lemma 10.** *If $f \in C_L^{1,1}(\mathbb{R}^d)$ is $\rho$-weakly convex, then*

- $\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \geq \frac{1}{2L}\left\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\right\|_2^2 - (\frac{\rho^2}{L} + \rho)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2$.
- *For $\eta \in (0, \frac{1}{L}]$,*

$$\left\|\mathbf{x}-\mathbf{y} - \eta(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))\right\|_2^2 \leq (1 + 6\rho\eta)\left\|\mathbf{x}-\mathbf{y}\right\|_2^2. \tag{39}$$

*Proof.* It follows from applying Theorem 2.1.5 in [Nesterov, 2003] to $f(\mathbf{x}) + \frac{\rho}{2}\mathbf{x}^{\mathrm{T}}\mathbf{x}$. □

**Lemma 11** (distance lemmas)**.** *Let's make Assumption 1, 3. Let $k \in [0, \tau-1]$ s.t. $t-k$ is a multiple of $\tau$, and let $\eta_{t-k} = \cdots = \eta_t := \eta$. For Equation (4), we have*

- *for $\eta \leq \frac{1}{12\rho\tau}$,*

$$\mathbb{E}\left[\Xi_t\right] \leq 24\eta^2\tau^2\zeta^2 + 6\eta^2\tau\sigma^2; \tag{40}$$

- *under Assumption 2, for $\eta \leq \min\left\{\frac{1}{36\rho\tau}, \frac{1}{17\delta\tau}, \frac{1}{6\sqrt{\mathcal{M}}\bar{\zeta}\tau}, \frac{1}{5\sqrt{\mathcal{M}}\sigma\tau^{\frac{3}{4}}}\right\}$,*

$$\mathbb{E}\left[\Xi_t\right] \leq 36\eta^2\tau^2\bar{\zeta}^2 + 6\eta^2\tau\sigma^2. \tag{41}$$

*Proof of Equation* (40). We claim no novelty in this proof of Equation (40) (*cf.* Lemma 8 in [Woodworth et al., 2020b]). We will use Equation (28) frequently in the proof. We have

$$\Xi_t \leq \frac{1}{n}\sum_{i=2}^{n} \mathbb{E}\left\|\mathbf{x}_t^i - \mathbf{x}_t^1\right\|_2^2 := \mathcal{E}_t. \tag{42}$$

Let $\widetilde{f} \in \mathbf{conv}\{f_1, \cdots, f_n\}$ be $\rho$-weakly convex. For $\eta \leq \frac{1}{12\rho\tau}$,

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{E}_{t+1}\right] \\
&\overset{(3)}{\leq} \frac{1}{n}\sum_{i=2}^{n} \mathbb{E}\left\|\mathbf{x}_t^i - \mathbf{x}_t^1 - \eta\left(\nabla f_i(\mathbf{x}_t^i) - \nabla f_1(\mathbf{x}_t^1)\right)\right\|_2^2 + 2\eta^2\sigma^2 \\
&\overset{(6)}{\leq} \left(1 + \frac{1}{2(\tau-1)}\right)\frac{1}{n}\sum_{i=2}^{n} \mathbb{E}\left\|\mathbf{x}_t^i - \mathbf{x}_t^1 - \eta\left(\nabla \widetilde{f}(\mathbf{x}_t^i) - \nabla \widetilde{f}(\mathbf{x}_t^1)\right)\right\|_2^2 + 8\eta^2\tau\zeta^2 + 2\eta^2\sigma^2 \\
&\overset{(39)}{\leq} \left(1 + \frac{1}{2(\tau-1)}\right)(1 + 6\rho\eta)\,\mathbb{E}\left[\mathcal{E}_t\right] + 8\eta^2\tau\zeta^2 + 2\eta^2\sigma^2 \\
&\leq \left(1 + \frac{1}{2(\tau-1)}\right)^2 \mathbb{E}\left[\mathcal{E}_t\right] + 8\eta^2\tau\zeta^2 + 2\eta^2\sigma^2 \\
&\leq 24\eta^2(\tau-1)\tau\zeta^2 + 6\eta^2(\tau-1)\sigma^2.
\end{aligned} \tag{43}$$

□

*Proof of Equation* (41). Let $k(t) \in [0, \tau-1]$ *s.t.* $t - k(t)$ is a multiple of $\tau$. We first strengthen the statement

$$\mathbb{E}\left[\Xi_t\right] \leq 36\eta^2 k(t)\tau\bar{\zeta}^2 + 6\eta^2 k(t)\sigma^2, \tag{44}$$

for which we will prove by induction.

If $k(t) = 0$, Equation (44) trivially follows.

Assume that for $0 \leq s \leq k-1$,

$$\mathbb{E}\left[\Xi_{t-s}\right] \leq 36\eta^2(k-s-1)\tau\bar{\zeta}^2 + 6\eta^2(k-s-1)\sigma^2, \tag{45}$$

where $k := k(t+1) \geq 1$.

We will use Equation (28) frequently in the following proof. In view of

$$\begin{aligned}
&\mathbb{E}\left[\Xi_{t+1}\right] \\
&\overset{(3)}{\leq} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t - \eta\left(\nabla f_i(\mathbf{x}_t^i) - \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_t^j)\right)\right\|_2^2 + \frac{n-1}{n}\eta^2\sigma^2 \\
&\overset{(30)}{\leq} \left(1 + \frac{1}{6(\tau-1)}\right)\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t - \eta\left(\nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t)\right)\right\|_2^2 + 6\eta^2\tau\,\mathbb{E}\left[8\delta^2\Xi_t + \frac{\mathcal{M}^2}{2}\Xi_t^2\right] + \eta^2\sigma^2,
\end{aligned} \tag{46}$$

and (let $\widetilde{f} \in \mathbf{conv}\{f_1, \cdots, f_n\}$ be $\rho$-weakly convex)

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t - \eta\left(\nabla f_i(\mathbf{x}_t^i) - \nabla f(\bar{\mathbf{x}}_t)\right)\right\|_2^2 \\
&\overset{(7)}{\leq} \left(1 + \frac{1}{6(\tau-1)}\right)\frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t - \eta\left(\nabla f_i(\mathbf{x}_t^i) - \nabla f_i(\bar{\mathbf{x}}_t)\right)\right\|_2^2 + 6\eta^2\tau\bar{\zeta}^2 \\
&\overset{(29)}{\leq} \left(1 + \frac{1}{6(\tau-1)}\right)^2\frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{x}_t^i - \bar{\mathbf{x}}_t - \eta\left(\nabla \widetilde{f}(\mathbf{x}_t^i) - \nabla \widetilde{f}(\bar{\mathbf{x}}_t)\right)\right\|_2^2 + 6\eta^2\tau\delta^2\Xi_t + 6\eta^2\tau\bar{\zeta}^2 \\
&\overset{(39)}{\leq} \left(1 + \frac{1}{6(\tau-1)}\right)^2 (1 + 6\rho\eta + 6\tau\delta^2\eta^2)\Xi_t + 6\eta^2\tau\bar{\zeta}^2,
\end{aligned} \tag{47}$$

for $\eta \leq \min\left\{\frac{1}{36\rho\tau}, \frac{1}{17\delta\tau}\right\}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\Xi_{t+1}\right] \\
&\overset{(46)(47)}{\leq} \left[\left(1+\frac{1}{6(\tau-1)}\right)^3 (1+6\rho\eta+6\tau\delta^2\eta^2)+48\eta^2\tau\delta^2\right]\mathbb{E}\left[\Xi_t\right]+3\eta^2\tau\mathcal{M}^2\,\mathbb{E}\left[\Xi_t^2\right]+6\eta^2\tau\bar{\zeta}^2+\eta^2\sigma^2 \\
&\leq \left(1+\frac{1}{6(\tau-1)}\right)^6 \mathbb{E}\left[\Xi_t\right]+3\eta^2\tau\mathcal{M}^2\,\mathbb{E}\left[\Xi_t^2\right]+6\eta^2\tau\bar{\zeta}^2+\eta^2\sigma^2 \\
&\leq 9\eta^2\tau\mathcal{M}^2\sum_{s=0}^{k-1}\mathbb{E}\left[\Xi_{t-s}^2\right]+18\eta^2 k\tau\bar{\zeta}^2+3\eta^2 k\sigma^2.
\end{aligned}
\tag{48}
$$

For $\eta \leq \min\left\{\frac{1}{6\sqrt{\mathcal{M}\bar{\zeta}\tau}}, \frac{1}{5\sqrt{\mathcal{M}\sigma}\tau^{\frac{3}{4}}}\right\}$, Equation (45) induces

$$
\begin{aligned}
\mathbb{E}\left[\Xi_{t+1}\right] &\overset{(48)}{\leq} 18\eta^2\tau\mathcal{M}^2 k\left(36^2\eta^4\tau^4\bar{\zeta}^4+36\eta^4\tau^2\sigma^4\right)+18\eta^2 k\tau\bar{\zeta}^2+3\eta^2 k\sigma^2 \\
&\leq 36\eta^2 k\tau\bar{\zeta}^2+6\eta^2 k\sigma^2.
\end{aligned}
\tag{49}
$$

Therefore, Equation (44) is proven by induction, and yields

$$
\mathbb{E}\left[\Xi_t\right] \leq 36\eta^2(\tau-1)\tau\bar{\zeta}^2+6\eta^2(\tau-1)\sigma^2. \tag{50}
$$

□

# B Comments on the generalized variant of LocalSGD and the 'merged local updates' analysis

In [Karimireddy et al., 2020, Yang et al., 2021], they proposed a generalized variant of LocalSGD, which can be formulated as:

$$
\mathbf{x}_{t+1}^i = \begin{cases} \bar{\mathbf{x}}_{t-\tau+1} - \frac{\bar{\eta}_t}{n}\sum_{j\in[n]}\sum_{k=0}^{\tau-1}\eta_{t-k}\mathbf{g}_{t-k}^j & \text{if } t+1=r\tau, \\ \mathbf{x}_t^i-\eta_t\mathbf{g}_t^i & \text{otherwise,} \end{cases} \tag{51}
$$

where they can set $\bar{\eta}_t > 1$.

In their analyses, they merged all local updates between two communication rounds (*i.e.*, $\frac{1}{|\mathcal{S}_{r-1}|}\sum_{j\in\mathcal{S}_{r-1}}\sum_{k=0}^{\tau-1}\eta_{t-k}\mathbf{g}_{t-k}^j$) to derive the descent lemma. As a result, in their descent lemmas (*cf.* Lemma 7 in the appendix of [Karimireddy et al., 2020], Relation (a8) in the appendix of [Yang et al., 2021]), they can only show the progress between two communication rounds.

Technically, for the progress at communication round $r=(t+1)/\tau$, they worked very hard to upper bound the variance of $\frac{1}{n}\sum_{i=1}^n\left\|\mathbf{x}_k^i-\bar{\mathbf{x}}_{t-\tau+1}\right\|_2^2$, in order to show that (informally) $\mathbf{g}_k^i \approx \nabla f_i(\bar{\mathbf{x}}_{t-\tau+1})$, $t-\tau+1\leq k\leq t$. But, following their analyses, the rate of LocalSGD can never surpass the rate of MbSGD, in which $\frac{1}{n}\sum_{i=1}^n\left\|\mathbf{x}_k^i-\bar{\mathbf{x}}_{t-\tau+1}\right\|_2^2 \equiv 0$.

Moreover, as long as keeping the equivalent stepsize $\bar{\eta}_t\cdot\sum_{k=0}^{\tau-1}\eta_{t-k}\leq\mathcal{O}(\frac{1}{L})$, people are always encouraged to set $\eta_{t-k}\to 0$ (*cf.* Theorem V in the appendix of [Karimireddy et al., 2020], Theorem 1 in [Yang et al., 2021]), and consequently, their generalized algorithm is uninterestingly reduced to MbSGD.

Therefore, in terms of the theoretical rate for problem 1, at least under their analyses, it's not clear whether the generalized variant is truly interesting, because it seems always inferior to MbSGD.