# Revisiting LocalSGD and SCAFFOLD: Improved Rates and Missing Analysis

Ruichen Luo, Sebastian Stich, Samuel Horvath and Martin Takac
rluo@ista.ac.at stich@cispa.de {samuel.horvath, martin.takac}@mbzuai.ac.ae
ISTA
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

## Optimization Problem

**Distributed stochastic (non-convex) optimization:**

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

where $f_i$'s are $L$-smooth and $f$ is bounded below.
**Stochastic gradients:** for $t \in [0, T-1]$, $i \in [n]$,

$$\mathbb{E}\left[\mathbf{g}_t^i\right] = \nabla f_i(\mathbf{x}_t^i), \quad \mathbb{E}\left\|\mathbf{g}_t^i - \nabla f_i(\mathbf{x}_t^i)\right\|_2^2 \le \sigma^2 .$$

**Communication interval:** $\tau$ ($T$ is a multiple of $\tau$).
**Notations:** $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^i$, $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$, $\Delta = f(\bar{\mathbf{x}}_0) - f^*$.

## MbSGD vs. LocalSGD/SCAFFOLD

**MbSGD:** $T = R\tau$. For $t \in [0, T-1]$, $i \in [n]$,

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

**LocalSGD:** $T = R\tau$. For $t \in [0, T-1]$, $i \in [n]$,

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

**SCAFFOLD [Kar+20]:** $T = 2R\tau$.

**Algorithm 1** SCAFFOLD

1: **for** $r = 0, 1, \cdots, R-1$ **do**
2: **for** $i \in [n]$ **do in parallel**
3: **for** $k = 0, 1, \cdots, \tau-1$ **do**
4: $\mathbf{x}_{2r\tau+k+1}^i = \mathbf{x}_{2r\tau+k}^i$
5: **end for**
6: $\widehat{\mathbf{g}}_{(r\tau)}^i = \frac{1}{\tau}\sum_{k=0}^{\tau-1} \mathbf{g}_{2r\tau+k}^i$
7: **end for**
8: Compute and broadcast: $\widehat{\mathbf{g}}_{(r\tau)} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{g}}_{(r\tau)}^i$
9: **for** $i \in [n]$ **do in parallel**
10: **for** $k = \tau, \tau+1, \cdots, 2\tau-2$ **do**
11: $\mathbf{x}_{2r\tau+k+1}^i = \mathbf{x}_{2r\tau+k}^i - \eta\left(\mathbf{g}_{2r\tau+k}^i - \widehat{\mathbf{g}}_{(r\tau)}^i + \widehat{\mathbf{g}}_{(r\tau)}\right)$
12: **end for**
13: **end for**
14: Compute:
$$\bar{\mathbf{x}}_{2(r+1)\tau} = \bar{\mathbf{x}}_{2r\tau} - \frac{\eta}{n}\sum_{j=1}^n \sum_{l=\tau}^{2\tau-1} \mathbf{g}_{2r\tau+l}^j$$
15: Broadcast: $\mathbf{x}_{2(r+1)\tau}^i = \bar{\mathbf{x}}_{2(r+1)\tau}$, for $i \in [n]$
16: **end for**

REFERENCES:

[Kar+20] Sai Praneeth Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: *ICML*. 2020.

[Kol+20] Anastasia Koloskova et al. "A unified theory of decentralized sgd with changing topology and local updates". In: *ICML*. 2020.

[WPS20] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. "Minibatch vs local sgd for heterogeneous distributed learning". In: *NIPS*. 2020.

## Assumptions

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

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

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

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

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

$$\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x}) - \nabla f_i(\mathbf{y}) + \nabla f(\mathbf{y})\|_2^2 \le \delta^2 \|\mathbf{x}-\mathbf{y}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d .$$

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

$$\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x}) - \nabla f_i(\mathbf{y}) + \nabla f(\mathbf{y})\|_2 \le \bar{\delta}\|\mathbf{x}-\mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad \forall i \in [n] .$$

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

$$f_i(\mathbf{x}) + \frac{\rho}{2}\mathbf{x}^\top\mathbf{x} \text{ is convex}, \quad \forall i \in [n] .$$

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

$$\left\|\nabla^2 \widehat{f}(\mathbf{x}) - \nabla^2 \widehat{f}(\mathbf{y})\right\|_2 \le \mathcal{M}\|\mathbf{x}-\mathbf{y}\|_2, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d .$$

## Existing Analysis

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

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

**Lemma 2** ([Kol+20]). *Under Assumption 1, there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_t)\|_2^2$:*

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

**Lemma 3** ([WPS20]). *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 $\|\bar{\mathbf{x}}_0 - \mathbf{x}^*\|_2 \le D$, then there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(\bar{\mathbf{x}}_t)] - f^*$:*

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

**Lemma 4** ([Kar+20]). *Suppose in Line 14 of Algorithm 1, a different global stepsize $\eta_g$ can be used when aggregating the updates. There exists $\eta_g \ge \eta > 0$ such that* SCAFFOLD *ensures the following upper bound on $\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_{2r\tau})\|_2^2$:*

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

**Lemma 5** ([Kar+20]). *Suppose $\widehat{\mathbf{g}}_{(r\tau)}^i = \nabla f_i(\bar{\mathbf{x}}_{2r\tau})$ in Line 6 of Algorithm 1. Under Assumptions 2+ and 3, if all $f_i$ are quadratic, then there exists $\eta > 0$ such that* SCAFFOLD *ensures the following upper bound on $\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_{2r\tau+\tau+k})\|_2^2$:*

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

**Remark.** *There is no theoretical speedup without more restrictive assumptions!*

## Our Analysis

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

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

**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 $\|\bar{\mathbf{x}}_0 - \mathbf{x}^*\|_2 \le D$, then there exists $\eta > 0$ such that* LocalSGD *ensures the following upper bound on $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[f(\bar{\mathbf{x}}_t)] - f^*$:*

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

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

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

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

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

**Theorem 5.** *Under Assumptions 2 to 4 with $\mathcal{M} = 0$, there exists $\eta > 0$ s.t.* SCAFFOLD *ensures the following upper bound on $\frac{2}{T}\sum_{r=0}^{R-1}\sum_{k=0}^{\tau-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_{2r\tau+\tau+k})\|_2^2$:*

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

**Remark.** *Our analyses are based on existing or weaker assumptions!*