New Nonconvex Analysis of LocalSGD and SCAFFOLD

Ruichen Luo[1]

Joint work with
Sebastian Stich[2] Samuel Horvath[3] Martin Takac[3]

[1]Institute of Science and Technology Austria (ISTA)

[2]CISPA Helmholtz Center for Information Security

[3]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

EUROPT'2025, Southampton, England

Distributed non-convex optimization problem

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

- $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable

Distributed non-convex optimization problem

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

- $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable
- f is the global objective function, bounded below by f^*

Distributed non-convex optimization problem

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

- $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable
- f is the global objective function, bounded below by f^*
- n is the number of workers

Distributed non-convex optimization problem

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

- $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable
- f is the global objective function, bounded below by f^*
- n is the number of workers
- f_i is the local objective function distributed to the ith worker, and f_i has L-Lipschitz continuous gradient, for each $i \in \{1, \cdots, n\}$

Distributed non-convex optimization problem

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

- $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable
- f is the global objective function, bounded below by f^*
- n is the number of workers
- f_i is the local objective function distributed to the ith worker, and f_i has L-Lipschitz continuous gradient, for each $i \in \{1, \cdots, n\}$
- The local objective functions are *heterogeneous* in general, *i.e.*, $f_i \neq f$

Stochastic oracle and intermittent communication

T subsequent queries to a *fully stochastic oracle* $\mathcal{SO}$:

- The workers input $\left(\mathbf{x}_t^1, \cdots, \mathbf{x}_t^n\right) \in \mathbb{R}^{d \times n}$
- The $\mathcal{SO}$ outputs $\left(G_1(\mathbf{x}_t^1, \xi_t^1), \cdots, G_n(\mathbf{x}_t^n, \xi_t^n)\right) \in \mathbb{R}^{d \times n}$
- $\left\{\xi_t^i : 0 \leq t \leq T-1\right\}$ are i.i.d. random variables
- Assume $\mathbb{E}_{\xi_t^i}\left[G_i(\mathbf{x}, \xi_t^i)\right] = \nabla f_i(\mathbf{x}),\ \mathbb{E}_{\xi_t^i}\left\|G_i(\mathbf{x}, \xi_t^i) - \nabla f_i(\mathbf{x})\right\|_2^2 \leq \sigma^2$

Stochastic oracle and intermittent communication

T subsequent queries to a *fully stochastic oracle* $\mathcal{SO}$:

- The workers input $\left(\mathbf{x}_t^1, \cdots, \mathbf{x}_t^n\right) \in \mathbb{R}^{d \times n}$
- The $\mathcal{SO}$ outputs $\left(G_1(\mathbf{x}_t^1, \xi_t^1), \cdots, G_n(\mathbf{x}_t^n, \xi_t^n)\right) \in \mathbb{R}^{d \times n}$
- $\left\{\xi_t^i : 0 \leq t \leq T-1\right\}$ are i.i.d. random variables
- Assume $\mathbb{E}_{\xi_t^i}\left[G_i(\mathbf{x}, \xi_t^i)\right] = \nabla f_i(\mathbf{x}),\ \mathbb{E}_{\xi_t^i}\left\|G_i(\mathbf{x}, \xi_t^i) - \nabla f_i(\mathbf{x})\right\|_2^2 \leq \sigma^2$

The worker communicates after every τ iterations:

- Assume T is a multiple of τ

Stochastic oracle and intermittent communication

T subsequent queries to a *fully stochastic oracle* $\mathcal{SO}$:

- The workers input $\left(\mathbf{x}_t^1, \cdots, \mathbf{x}_t^n\right) \in \mathbb{R}^{d\times n}$
- The $\mathcal{SO}$ outputs $\left(G_1(\mathbf{x}_t^1, \xi_t^1), \cdots, G_n(\mathbf{x}_t^n, \xi_t^n)\right) \in \mathbb{R}^{d\times n}$
- $\left\{\xi_t^i : 0 \le t \le T-1\right\}$ are i.i.d. random variables
- Assume $\mathbb{E}_{\xi_t^i}\left[G_i(\mathbf{x}, \xi_t^i)\right] = \nabla f_i(\mathbf{x}),\ \mathbb{E}_{\xi_t^i}\left\|G_i(\mathbf{x}, \xi_t^i) - \nabla f_i(\mathbf{x})\right\|_2^2 \le \sigma^2$

The worker communicates after every τ iterations:

- Assume T is a multiple of τ

Notations:

- $\left(\mathbf{g}_t^1, \cdots, \mathbf{g}_t^n\right) = \left(G_1(\mathbf{x}_t^1, \xi_t^1), \cdots, G_n(\mathbf{x}_t^n, \xi_t^n)\right)$
- $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^i$
- $f^* = \inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$
- $\Delta = f(\bar{\mathbf{x}}_0) - f^*$

Minibatch SGD vs. Local SGD/SCAFFOLD

Initialization: $\mathbf{x}_0^1 = \cdots = \mathbf{x}_0^n$

MbSGD.

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

Minibatch SGD vs. Local SGD/SCAFFOLD

Initialization: $\mathbf{x}_0^1 = \cdots = \mathbf{x}_0^n$

MbSGD.

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

LocalSGD.

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

Minibatch SGD vs. Local SGD/SCAFFOLD

Initialization: $\mathbf{x}_0^1 = \cdots = \mathbf{x}_0^n$

MbSGD.

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

LocalSGD.

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

SCAFFOLD.

(next page)

Minibatch SGD vs. Local SGD/SCAFFOLD

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

Assumptions (gradient smilarity)

Assumption 1 (Standard gradient similarity – SGS)

For some $\zeta \geq 0$, we have

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

Assumptions (gradient smilarity)

Assumption 1 (Standard gradient similarity – SGS)

For some $\zeta \geq 0$, we have

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

Assumption 1+ (Uniform gradient similarity – UGS)

For some $\bar{\zeta} \geq 0$, we have

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

Assumptions (Hessian similarity)

Assumption 2 (Standard Hessian similarity – SHS)

For some $\delta \in [0, L]$, we have

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

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

Assumptions (Hessian similarity)

Assumption 2 (Standard Hessian similarity – SHS)

For some $\delta \in [0, L]$, we have

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

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

Assumption 2+ (Uniform Hessian similarity – UHS)

For some $\bar{\delta} \in [0, 2L]$, we have

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

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

Assumptions (weak convexity, Lipschitz continuous Hessian)

Assumption 3 (Weak convexity – WC)

For some $\rho \in [0, L]$, we have

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

for all $i \in [n]$.

Assumptions (weak convexity, Lipschitz continuous Hessian)

Assumption 3 (Weak convexity – WC)

For some $\rho \in [0, L]$, we have

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

for all $i \in [n]$.

Assumption 4 (Lipschitz continuous Hessian – LCH)

For some $\mathcal{M} \geq 0$, there exists (at least) one function $\hat{f}$ such that: $\hat{f} \in \mathbf{conv}\{f_1, \cdots, f_n\}$, and

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

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

MbSGD

Lemma 1

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

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

LocalSGD: non-convex speedup from WC

Lemma 2 ([Kol+20])

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

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

LocalSGD: non-convex speedup from WC

Lemma 2 ([Kol+20])

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

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

Theorem 1 (Ours)

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

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

LocalSGD: convex speedup without UGS

Lemma 3 ([WPS20])

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

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

LocalSGD: convex speedup without UGS

Lemma 3 ([WPS20])

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

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

Theorem 2 (Ours)

Under Assumption 1, $\cdots$

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

LocalSGD: improved conditioning from UHS & LCH

Theorem 3 (Ours)

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

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

SCAFFOLD: existing analyses

Lemma 4 ([Kar+20])

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

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

SCAFFOLD: existing analyses

Lemma 4 ([Kar+20])

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

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

Lemma 5 ([Kar+20])

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

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

SCAFFOLD: speedup without quadratic/UHS

Theorem 4 (Ours)

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

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

SCAFFOLD: speedup without quadratic/UHS

Theorem 4 (Ours)

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

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

Theorem 5 (Ours)

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

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

References

[Kar+20] Sai Praneeth Karimireddy et al. "Scaffold: Stochastic controlled averaging for federated learning". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5132–5143.

[Kol+20] Anastasia Koloskova et al. "A unified theory of decentralized sgd with changing topology and local updates". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5381–5393.

[WPS20] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. "Minibatch vs local sgd for heterogeneous distributed learning". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6281–6292.