DoCoM: Compressed Decentralized Optimization with Near-Optimal Sample Complexity

Chung-Yiu Yau, Hoi-To Wai

Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong

31st Jul, 2023

Introduction

- In this work, we consider a network-connected multi-agent system that solves the cooperative decentralized optimization problem.
- We proposed a communication-efficient and sample-efficient stochastic gradient algorithm by combining
  - communication compression and
  - hybrid gradient variance reduction.
- The proposed algorithm is proved to show fast convergence rate on smooth non-convex objective function.

Decentralized Optimization Problem

We investigate the distributed optimization problem over a network of n agents:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right] \tag{1}$$

for differentiable non-convex function $f_i : \mathbb{R}^d \to \mathbb{R}$ such that

- $f_i(x) > -\infty \ \forall x \in \mathbb{R}^d$,
- each agent only access its local objective function $f_i(x)$ (or $f_i(x;\xi)$), e.g., $f_i(x)$ follows a local data distribution D_i such that $f_i(x) = \mathbb{E}_{\xi \sim D_i}[f_i(x;\xi)]$ and $D_i \neq D_j$ if $i \neq j$.

Classical algorithm such as DSGD [Lian et al., 2017] send model parameters $x \in \mathbb{R}^d$ to neighbours in the graph defined on a mixing matrix $\mathbf{W} \in \mathbb{R}_+^{n \times n}$:

$$x_i^{t+1} = \sum_{j=1}^{n} W_{ij} x_j^t - \eta \nabla f_i(x_i^t; \xi_i^t) \tag{2}$$

Compressed Communication

- **Motivation.** High-dimensional model (e.g. deep neural network) poses a communication burden for computing $\sum_{j=1}^{n} W_{ij} x_j$.
- Compression on the exchanged message reduces network usage.

Assumption 1. *[Koloskova et al., 2019] A random* **contractive compression** *operator $\mathcal{Q}_\xi(x) : \mathbb{R}^d \to \mathbb{R}^d$ with $\xi \sim \pi_x$ should satisfy for any $x \in \mathbb{R}^d, 0 < \delta \leq 1$,*

$$\mathbb{E}_{\xi \sim \pi}[\|\mathcal{Q}_\xi(x) - x\|^2] \leq (1 - \delta)\|x\|^2. \tag{3}$$

Example 1. Top-k/random-k sparsifier satisfy Assumption 1 with $\delta = k/d$, for choosing an index set $\mathcal{I}_x \subseteq \{1, \ldots, d\}$, $|\mathcal{I}_x| = k$ and

$$\mathcal{Q}_{\mathcal{I}_x}(x)_i = \begin{cases} x_i & \text{if } i \in \mathcal{I}_x, \\ 0 & \text{if } i \notin \mathcal{I}_x. \end{cases} \tag{4}$$

- Top-k choose $\mathcal{I}_x$ as the first k coordinates of largest magnitude $|x_i|$.

Example 2. Randomized gossip satisfies Assumption 1 and allows agents to randomly skip updating its model parameters with neighbours by

$$\mathcal{Q}(x) = \begin{cases} x & \text{w.p. } \delta, \\ \mathbf{0} & \text{w.p. } 1 - \delta. \end{cases} \tag{5}$$

Communication Compression Algorithm - CHOCO-SGD [Koloskova et al., 2019][1]

- With step sizes $\eta, \gamma > 0$,

$$\text{(Initialization.) } \widehat{x}_i^0 = \mathbf{0} \tag{6a}$$

$$\widehat{x}_i^{t+1} = \widehat{x}_i^t + \mathcal{Q}(\overbrace{x_i^t - \eta \nabla f_i(x_i^t; \xi_i^t) - \widehat{x}_i^t}^{\text{forward difference}}) \tag{6b}$$

$$x_i^{t+1} = \underbrace{(x_i^t - \eta \nabla f_i(x_i^t; \xi_i^t))}_{\text{gradient step}} \underbrace{- \gamma \widehat{x}_i^{t+1} + \gamma \sum_{j=1}^{n} W_{ij} \widehat{x}_j^{t+1}}_{\text{gossip averaging}} \tag{6c}$$

- Agent j uses $\widehat{x}_i^{t+1}$ approximate $x_i^t - \eta \nabla f_i(x_i^t; \xi_i^t)$ by receiving compressed forward difference $\mathcal{Q}(x_i^t - \eta \nabla f_i(x_i^t; \xi_i^t) - \widehat{x}_i^t)$ from agent i.

- With diminishing step size η, $x_i^t - \eta \nabla f_i(x_i^t; \xi_i^t)$ evolve slowly $\Longrightarrow \|x_i^t - \eta \nabla f_i(x_i^t; \xi_i^t) - \widehat{x}_i^t\|_2^2$ shrinks and compression error vanishes by contraction property (3).

[1] Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. ICML, 2019.

Variance Reduced Gradient Tracking - GT-HSGD[2]

- When only stochastic gradient is available, [Tran-Dinh et al., 2021] combine variance-reduced estimator (e.g., SARAH [Nguyen et al., 2017]) with SGD.

Assumption 2. *For every $i \in [n]$, stochastic function $f_i(x;\xi)$ satisfies* **mean-squared smoothness** *if there exists $L \geq 0$ such that*

$$\mathbb{E}_\xi[\|\nabla f_i(x;\xi) - \nabla f_i(y;\xi)\|^2] \leq L^2\|x-y\|^2 \quad \forall x, y \in \mathbb{R}^d \tag{7}$$

- In the context of gradient tracking, this is achieved by **GT-HSGD** [Xin et al., 2021] with initial batch size b_0, and step sizes $\eta, \beta > 0$.
- v_i^t approximates local exact gradient $\nabla f_i(x_i^t)$.

$$\text{(Initialization.) } v_i^0 = \frac{1}{b_0}\sum_{r=1}^{b_0} \nabla f_i(x_i^0; \xi_i^{0,r}), \quad g_i^0 = \sum_{j=1}^n W_{ij} v_j^0 \tag{8a}$$

$$x_i^{t+1} = \sum_{j=1}^n W_{ij}(x_j^t - \eta g_j^t) \tag{8b}$$

$$g_i^{t+1} = \sum_{j=1}^n W_{ij}(g_j^t + v_j^{t+1} - v_j^t) \tag{8c}$$

$$v_i^{t+1} = \beta \underbrace{\nabla f_i(x_i^{t+1}; \xi_i^{t+1})}_{\text{SGD}} + (1-\beta)\underbrace{(v_i^t + \nabla f_i(x_i^{t+1}; \xi_i^{t+1}) - \nabla f_i(x_i^t; \xi_i^{t+1}))}_{\text{SARAH}} \tag{8d}$$

- Variance-reduced algorithms achieve $\mathcal{O}(1/T^{2/3})$ convergence rate, whereas non-accelerated SGD algorithms only achieve $\mathcal{O}(1/\sqrt{T})$ convergence.

[2] Ran Xin, Usman A Khan, and Soummya Kar. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. ICML, 2021.

Stochastic Gradient Tracking with Compressed Communication - DoCoM

- We propose an algorithm that combines
  1. $\boxed{\textit{communication compression}}$ (9c) & (9d), (9f) & (9g),
  2. $\boxed{\textit{gradient tracking}}$ (9f) & (9g), and
  3. $\boxed{\textit{variance reduction}}$ (9e).

- Choose initial batch size b_0, initial model parameter $\bar{x}^0 \in \mathbb{R}^d$,

$$\text{(Initialization.) } \widehat{x}_i^0 = x_i^0 = \bar{x}^0 \tag{9a}$$

$$\text{(Initialization.) } g_i^0 = v_i^0 = b_0^{-1} \sum_{r=1}^{b_0} \nabla f_i(x_i^0; \xi_i^{0,r}) \tag{9b}$$

$$x_i^{t+1} = x_i^t - \eta g_i^t + \boxed{\gamma \sum_{j=1}^n W_{ij} (\widehat{x}_j^{t+1} - \widehat{x}_i^{t+1})} \tag{9c}$$

$$\widehat{x}_i^{t+1} = \boxed{\widehat{x}_i^t + \mathcal{Q}\left(x_i^t - \eta g_i^t - \widehat{x}_i^t\right)} \tag{9d}$$

$$v_i^{t+1} = \beta \nabla f_i(x_i^{t+1}; \xi_i^{t+1}) + \boxed{(1-\beta)\left[v_i^t + \nabla f_i(x_i^{t+1}; \xi_i^{t+1}) - \nabla f_i(x_i^t; \xi_i^{t+1})\right]} \tag{9e}$$

$$g_i^{t+1} = \boxed{g_i^t + v_i^{t+1} - v_i^t} + \boxed{\gamma \sum_{j=1}^n W_{ij} (\widehat{g}_j^{t+1} - \widehat{g}_i^{t+1})} \tag{9f}$$

$$\widehat{g}_i^{t+1} = \boxed{\widehat{g}_i^t + \mathcal{Q}\left(\boxed{g_i^t + v_i^{t+1} - v_i^t} - \widehat{g}_i^t\right)} \tag{9g}$$

Convergence Analysis of DoCoM

Convergence Rate. Under Assumptions 1, 2, by carefully choosing the step sizes as $\beta = \frac{n^{1/3}}{T^{2/3}}, \eta = \frac{n^{2/3}}{LT^{1/3}}, \gamma = \gamma_\infty, b_0 = \frac{T^{1/3}}{n^{2/3}}$, it can be shown that a random iteration T drawn i.i.d uniformly from $\{0, \ldots, T-1\}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\|\nabla f(x_i^\mathsf{T})\|^2\right] = \mathcal{O}\left(\frac{L(f(\bar{x}^0) - f^\star)}{(nT)^{2/3}} + \frac{\sigma^2}{(nT)^{2/3}} + \underbrace{\frac{n\overline{G}_0}{\delta^2 \rho^4 T} + \frac{\sigma^2 n^{5/3}}{\delta^3 \rho^6 T^{4/3}}}_{\text{transient effect}} \right) \tag{10}$$

- **Dominant Terms (L, σ^2).** The local iterate x_i^T is $\mathcal{O}(1/T^{2/3})$-stationary on $f = \frac{1}{n}\sum_{i=1}^n f_i$, ($\mathcal{O}(1/T^{2/3})$-stationary is SOTA even in the centralized setting).
- **Transient Time.** For $T \geq T_{\text{trans}} = \Omega(n^3 \overline{G}_0^3 / (\sigma^6 \delta^6 \rho^{12}))$, the impact of network topology ($\rho := 1 - \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\}$) and compressor ($\delta$) vanishes.
- **Linear Speedup when $T \geq T_{\mathrm{trans}}$.** No. of iterations T reduces linearly with n (2x nodes $\Rightarrow \frac{1}{2}$ time).

Comparison over Iteration Complexity

Table: Comparison for smooth *non-convex* objective. Iteration complexity is the smallest T such that $\bar{x}$ is ϵ-stationary: $T^{-1}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{x}^t)\|^2] \leq \epsilon^2$. Highlighted in red are dominant terms when $\epsilon \to 0$.

Algorithms	**Iteration Complexity**	**Compress**
DSGD[3]	$\mathcal{O}\left(\max\left\{\frac{\sigma^2}{n}\epsilon^{-4}, \frac{n(\sigma^2+\varsigma^2)}{\rho^2\epsilon^2}\right\}\right)$	✗
GNSD (GT)[4]	$\mathcal{O}\left(\frac{1}{C_0^2 C_1^2}\epsilon^{-4}\right)$	✗
GT-HSGD[5]	$\mathcal{O}\left(\max\left\{\frac{\sigma^3}{n}\epsilon^{-3}, \frac{\overline{G}_0}{\rho^3\epsilon^2}, \frac{n^{0.5}\sigma^{1.5}}{\rho^{2.25}\epsilon^{1.5}}\right\}\right)$	✗
CHOCO-SGD[6]	$\mathcal{O}\left(\max\left\{\frac{\sigma^2}{n}\epsilon^{-4}, \frac{G}{\delta\rho^2\epsilon^3}\right\}\right)$	✓
BEER[7]	$\mathcal{O}\left(\max\left\{\frac{\sigma^2}{\delta^2\rho^3}\epsilon^{-4}, \frac{1}{\delta\rho^3\epsilon^2}\right\}\right)$	✓
DoCoM	$\mathcal{O}\left(\max\left\{\frac{\sigma^3}{n}\epsilon^{-3}, \frac{n\overline{G}_0}{\delta^2\rho^4\epsilon^2}, \frac{n^{1.25}\sigma^{1.5}}{\delta^{2.25}\rho^{4.5}\epsilon^{1.5}}\right\}\right)$	✓

[3] [Lian et al., 2017]
[4] [Lu et al., 2019]
[5] [Xin et al., 2021]
[6] [Koloskova et al., 2019]
[7] [Zhao et al., 2022]

Faster Convergence under PL Condition

Assumption 3. *For any* $x \in \mathbb{R}^d$, $f(x) \geq f^\star > -\infty$, *there exists* $\mu > 0$ *such that*

$$\|\nabla f(x)\|_2^2 \geq 2\mu\big[f(x) - f^\star\big] \tag{11}$$

- By properly choosing the step sizes as $\eta = \beta = \log T/T$, $\gamma = \gamma_\infty$, $b_0 = \Omega(1)$, the last iterates $\bar{x}^T = n^{-1}\sum_{j=1}^n x_j^T$ of DoCoM satisfy

$$\mathbb{E}[f(\bar{x}^T)] - f^\star = \mathcal{O}(\log T/T), \tag{12}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_i^T - \bar{x}^T\|^2] = \mathcal{O}(\log T/T). \tag{13}$$

- ($\sigma = 0$) With deterministic gradient, DoCoM achieves linear convergence

$$\mathbb{E}[f(\bar{x}^T)] - f^\star = \mathcal{O}((1 - \widetilde{\beta})^T) \tag{14}$$

where $\widetilde{\beta} = \min\{\eta\mu, \overline{\beta}/2\}$.

Proof Outline

- **Descent Lemma.**

$$f(\bar{x}^{t+1}) \leq f(\bar{x}^t) - \frac{\eta}{2}\|\nabla f(\bar{x}^t)\|^2 + \frac{L^2\eta}{n}\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{X}^t\|_F^2 + \eta\|\overline{v}^t - \overline{\nabla F}^t\|^2 - \frac{\eta}{4}\|\overline{g}^t\|^2$$

- We can show contraction of the following potential function $\mathtt{V}^t$ for some analytically deduced weightings $a, b, c > 0$:

$$\begin{aligned}\mathtt{V}^t &= \mathbb{E}\left[L^2\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{X}^t\|_F^2 + n\|\overline{v}^t - \overline{\nabla F}^t\|^2 + n^{-1}\|\mathbf{V}^t - \nabla F^t\|_F^2\right] \\ &+ \mathbb{E}\left[a\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{G}^t\|_F^2 + b\|\mathbf{G}^t - \widehat{\mathbf{G}}^t\|_F^2 + c\|\mathbf{X}^t - \eta\mathbf{G}^t - \widehat{\mathbf{X}}^t\|_F^2\right]\end{aligned} \tag{15}$$

- A vector inequality (in expectation) is developed with $\mathbf{M}_{\eta,\gamma,\beta} \in \mathbb{R}^{6\times 6}$,

$$\mathbb{E}\begin{bmatrix} L^2\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{X}^{t+1}\|_F^2 \\ n\|\overline{v}^{t+1} - \overline{\nabla F}^{t+1}\|^2 \\ n^{-1}\|\mathbf{V}^{t+1} - \nabla F^{t+1}\|_F^2 \\ a\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{G}^{t+1}\|_F^2 \\ b\|\mathbf{G}^{t+1} - \widehat{\mathbf{G}}^{t+1}\|_F^2 \\ c\|\mathbf{X}^{t+1} - \eta\mathbf{G}^{t+1} - \widehat{\mathbf{X}}^{t+1}\|_F^2 \end{bmatrix} \leq \mathbf{M}_{\eta,\gamma,\beta}\,\mathbb{E}\begin{bmatrix} L^2\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{X}^{t}\|_F^2 \\ n\|\overline{v}^{t} - \overline{\nabla F}^{t}\|^2 \\ n^{-1}\|\mathbf{V}^{t} - \nabla F^{t}\|_F^2 \\ a\|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^\top)\mathbf{G}^{t}\|_F^2 \\ b\|\mathbf{G}^{t} - \widehat{\mathbf{G}}^{t}\|_F^2 \\ c\|\mathbf{X}^{t} - \eta\mathbf{G}^{t} - \widehat{\mathbf{X}}^{t}\|_F^2 \end{bmatrix} + \mathcal{O}(\beta^2\sigma^2)\mathbf{1},$$

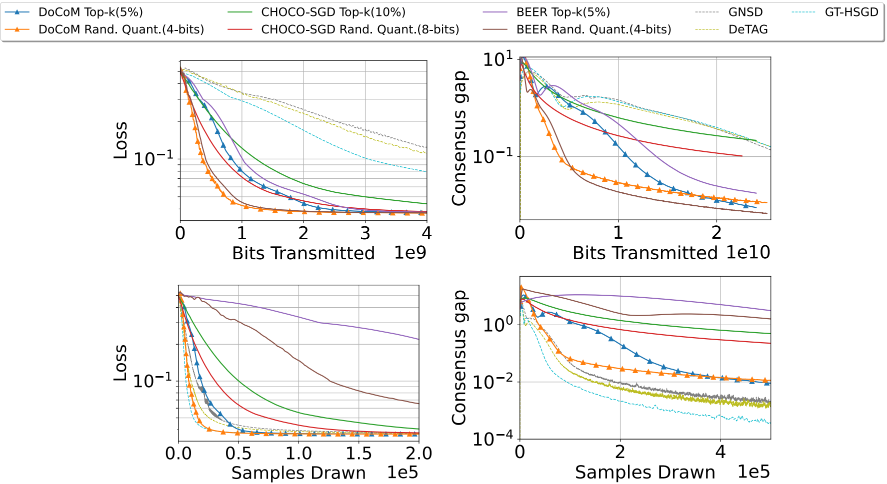
$$M_{11} = 1 - \rho\gamma/2,\ M_{22} = M_{33} = (1-\beta)^2,\ M_{44} = 1 - \rho\gamma/4,\ M_{55} = M_{66} = 1 - \delta/8.$$

- Off-diagonals $\mathbf{M}$ are controlled by η, γ, β.

- $\exists \eta, \gamma, \beta$ s.t. $\max\{|\lambda_1(\mathbf{M})|, \ldots, |\lambda_6(\mathbf{M})|\} < 1$.

Numerical Experiments

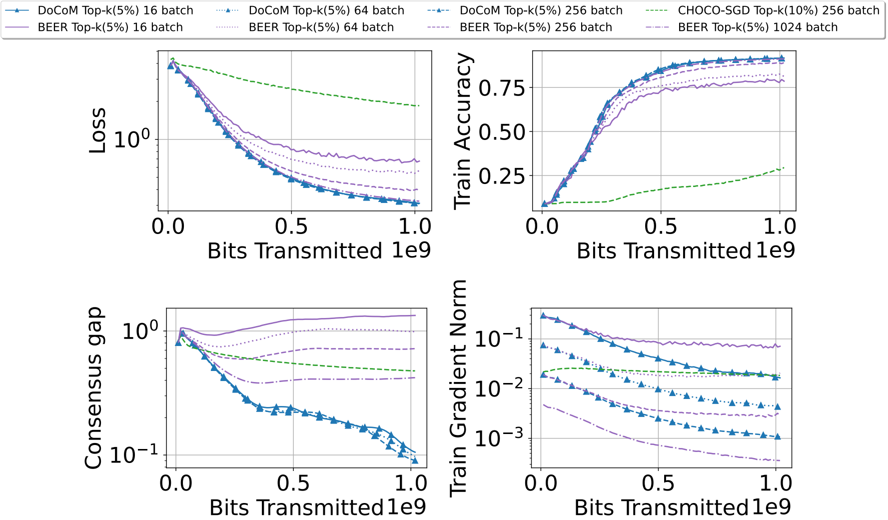
- **Exp 1.** Sigmoid loss linear model on generated dataset.



- DoCoM converges fastest with least network usage.
- DoCoM is the only compressed algorithm achieving the same sample/iteration efficiency as uncompressed algorithms.

Numerical Experiments

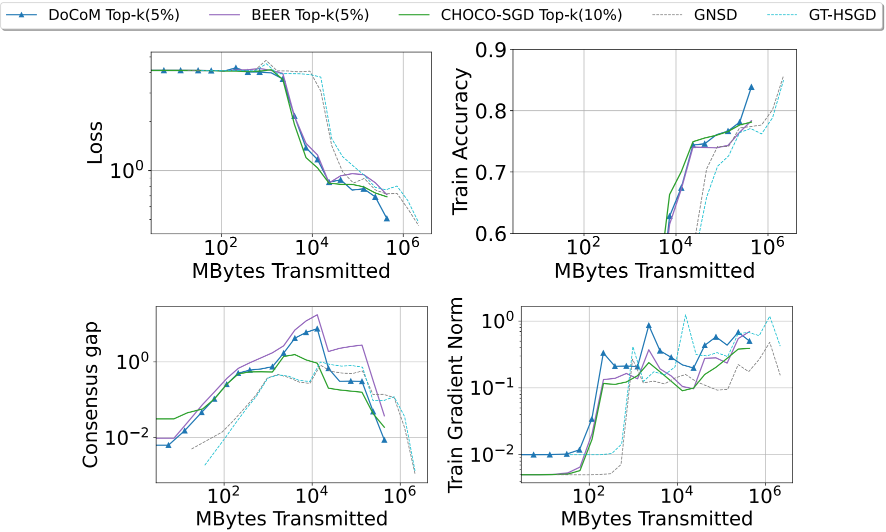
- **Exp 2.** Feed-forward Neural Network on unshuffled MNIST.



- DoCoM converge fastest under <u>small batch size</u>.

Numerical Experiments

- **Exp 3.** LeNet-5 (convolutional neural network) on FEMNIST[8].



[8] Extended MNIST for Federated setting. [Caldas et al., 2019]

Conclusion

- We proposed **DoCoM**, achieving fast convergence rate of $\mathcal{O}(1/T^{2/3})$ for smooth non-convex stochastic optimization under communication compression.
- Open research problems such as analyzing communication-efficient algorithms on time-varying communication graph, and/or asynchronous communication.

Reference I

[Caldas et al., 2019] Caldas, S., Duddu, S. M. K., Wu, P., Li, T., KoneÄnÃ½, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2019).
Leaf: A benchmark for federated settings.
In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*.

[Koloskova et al., 2019] Koloskova, A., Stich, S., and Jaggi, M. (2019).
Decentralized stochastic optimization and gossip algorithms with compressed communication.
In *International Conference on Machine Learning*, pages 3478–3487. PMLR.

[Lian et al., 2017] Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. (2017).
Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent.
In *Neural Information Processing Systems*, pages 5336–5346.

[Lu et al., 2019] Lu, S., Zhang, X., Sun, H., and Hong, M. (2019).
Gnsd: a gradient-tracking based nonconvex stochastic algorithm for decentralized optimization.
In *2019 IEEE Data Science Workshop (DSW)*, pages 315–321.

Reference II

[Nguyen et al., 2017] Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017).
Sarah: A novel method for machine learning problems using stochastic recursive gradient.
In *International Conference on Machine Learning*, pages 2613–2621. PMLR.

[Tran-Dinh et al., 2021] Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. (2021).
A hybrid stochastic optimization framework for composite nonconvex optimization.
Mathematical Programming, pages 1–67.

[Xin et al., 2021] Xin, R., Khan, U. A., and Kar, S. (2021).
A hybrid variance-reduced method for decentralized stochastic non-convex optimization.
In *ICML*.

[Zhao et al., 2022] Zhao, H., Li, B., Li, Z., Richtárik, P., and Chi, Y. (2022).
Beer: Fast $o(1/t)$ rate for decentralized nonconvex optimization with communication compression.
arXiv preprint arXiv:2201.13320.