Decentralized Non-convex Stochastic Optimization with Heterogeneous Variance

Hongxu Chen[1], **Ke Wei**[1], **Luo Luo**[1*]

[1]School of Data Science, Fudan University
{hxchen20, kewei, luoluo}@fudan.edu.cn

Abstract

Decentralized optimization is critical for solving large-scale machine learning problems over distributed networks, where multiple nodes collaborate through local communication. In practice, the variances of stochastic gradient estimators often differ across nodes, yet their impact on algorithm design and complexity remains unclear. To address this issue, we propose D-NSS, a decentralized algorithm with node-specific sampling, and establish its sample complexity depending on the arithmetic mean of local standard deviations, achieving tighter bounds than existing methods that rely on the worst-case or quadratic mean. We further derive a matching sample complexity lower bound under heterogeneous variance, thereby proving the optimality of this dependence. Moreover, we extend the framework with a variance reduction technique and develop D-NSS-VR, which under the mean-squared smoothness assumption attains an improved sample complexity bound while preserving the arithmetic-mean dependence. Finally, numerical experiments validate the theoretical results and demonstrate the effectiveness of the proposed algorithms.

Introduction

With the increasing demand for data processing and computation, distributed optimization has become an essential tool for large-scale machine learning problems. In particular, decentralized optimization allows nodes to communicate with their neighbors, improving robustness and avoiding communication bottlenecks. In this paper, we consider the following decentralized non-convex stochastic optimization problem:

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

where $f_i(x) = \mathbb{E}_{\xi_i\sim\mathcal{D}_i}\left[F(x;\xi_i)\right]$ is a possibly non-convex objective defined by the local data distribution $\mathcal{D}_i$ at node i and ξ_i is the corresponding random index.

Heterogeneity in the local data distributions $\mathcal{D}_i$ leads to differences in the variances of stochastic gradient estimators across nodes. Specifically, for first-order methods applied to problem (1), node i can only access an unbiased stochastic gradient estimator $g_i(x;\xi_i)$ satisfying

$$\mathbb{E}_{\xi_i}\left[\|g_i(x;\xi_i) - \nabla f_i(x)\|^2\right] \le \sigma_i^2, \tag{2}$$

where $\sigma_i > 0$ may vary significantly across nodes due to the non-IID data distribution. However, most existing decentralized optimization methods (Lian et al. 2017; Tang et al. 2018; Assran et al. 2019; Yuan et al. 2022; Lu and De Sa 2023) typically assume uniform variance across nodes, resulting in complexity bounds that depend on the worst-case standard deviation $\sigma_{\max}$. Xin et al. (Xin, Khan, and Kar 2021b,a; Xin et al. 2021) analyze node-wise noise and derive bounds based on the quadratic mean of standard deviations, $\bar{\sigma}_{\rm QM} := \sqrt{(\sigma_1^2+\cdots+\sigma_m^2)/m}$, but their algorithms still assume identical sample sizes across nodes. Consequently, how variance heterogeneity affects the convergence and complexity of decentralized algorithms remains unclear, raising two fundamental questions:

(i) *How to design efficient decentralized algorithms under heterogeneous variance?*

(ii) *What is the optimal sample complexity in this setting?*

To answer these questions, we investigate the problem from a sampling perspective and propose a more efficient decentralized algorithm with node-specific sampling, called D-NSS (Decentralized Optimization with Node-Specific Sampling), which achieves the sample complexity (i.e., the total number of stochastic gradient evaluations) of

$$O\left(\frac{\Delta L\,\bar{\sigma}_{\rm AM}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right)$$

for finding the ϵ-stationary point, where $\bar{\sigma}_{\rm AM} = \frac{1}{m}\sum_{i=1}^m \sigma_i$ denotes the arithmetic mean of standard deviations. This dependence is tighter than the worst-case or quadratic-mean bounds in previous works. In highly heterogeneous regimes such $\bar{\sigma}_{\rm QM} = \Theta(\sqrt{m}\bar{\sigma}_{\rm AM})$), it yields an $O(m)$ improvement in sample complexity. In addition, we establish a lower bound for decentralized optimization with heterogeneous variance, thereby proving the optimality of D-NSS. Furthermore, under the mean-squared smoothness assumption, we incorporate variance reduction into the proposed scheme and develop D-NSS-VR, which achieves a sample complexity of

$$O\left(\frac{\Delta\bar{L}\bar{\sigma}_{\rm AM}}{\epsilon^3} + \frac{\bar{\sigma}_{\rm AM}^2}{\epsilon^2} + \frac{\sqrt{m}\Delta\bar{L}}{\epsilon^2} + m\right),$$

where Δ denotes the initial optimality gap, L and $\bar{L}$ are smoothness constants.

*Corresponding author.

Related Work

Decentralized optimization has developed rapidly over the past decade. A classical method is decentralized gradient descent (DGD), which performs gradient descent on local objectives followed by one round of communication. Under data heterogeneity, DGD requires diminishing step sizes to guarantee convergence, which results in slow convergence rates (Nedic and Ozdaglar 2009; Yuan, Ling, and Yin 2016). Gradient tracking techniques (Di Lorenzo and Scutari 2016; Pu and Nedić 2021; Qu and Li 2017; Scutari and Sun 2019) improve convergence by enabling each node to track the global gradient through additional communication of gradient information. Shi et al. (2015) propose EXTRA, a first-order method that introduces a correction term to achieve exact convergence with constant step sizes. Nedic, Olshevsky, and Shi (2017) combine gradient tracking with inexact gradient methods and establish the linear convergence of the DIGing algorithm under strong convexity and smoothness over time-varying graphs. To further improve communication efficiency, Chebyshev-accelerated multi-step communication methods have been developed to achieve optimal communication complexity in decentralized optimization (Scaman et al. 2017; Kovalev, Salim, and Richtárik 2020; Ye et al. 2023).

For decentralized non-convex stochastic optimization, Lian et al. (2017) show that decentralized stochastic gradient descent can find an ϵ-stationary point (i.e., $\mathbb{E}[\|\nabla f(x)\|^2] \leq \epsilon^2$) with a sample complexity of $O(\epsilon^{-4})$. The same complexity guarantees are obtained for primal-dual algorithms by Yi et al. (2022). Further improvements are obtained by incorporating multi-step communication techniques, as shown in the works of Lu and De Sa (2021, 2023) and Yuan et al. (2022), resulting in nearly optimal sample and communication complexities. When the stochastic gradients additionally satisfy the mean-squared smoothness condition, variance reduction methods yield improved sample complexity. Sun, Lu, and Hong (2020) propose D-GET by integrating variance reduction with gradient tracking, while Pan, Liu, and Wang (2020) extend the SPIDER-SFO method (Fang et al. 2018) to the decentralized setting. Xin, Khan, and Kar (2021a) develop GT-HSGD, a single-loop method based on STORM (Cutkosky and Orabona 2019), and establish a network-independent $O(\epsilon^{-3})$ sample complexity.

However, most of these results rely on the assumption of uniform variance across nodes, and their complexity bounds depend on the worst-case variance $\sigma_{\max}$. Xin et al. (Xin, Khan, and Kar 2021b,a; Xin et al. 2021) consider heterogeneous variance in decentralized stochastic optimization and derive an upper bound that depends on the quadratic mean of standard deviations $\bar{\sigma}_{\mathrm{QM}}$, while Deng and Hu (2025) obtain similar results in the context of manifold optimization. It remains an open problem whether more efficient algorithms can be designed under heterogeneous variance.

Moreover, theoretical lower bounds are essential for characterizing the performance limits of algorithms. For distributed optimization, Arjevani and Shamir (2015) derive iteration lower bounds under deterministic convex setting. Scaman et al. (2017) analyze the lower complexity bounds for both centralized and decentralized algorithms in the smooth and strongly convex setting. In the non-convex case, Arjevani et al. (2023) establish the complexity lower bound for single-machine non-convex stochastic optimization. Lu and De Sa (2021, 2023) extend this result to the decentralized setting, though their analysis relies on a specific linear communication topology. Yuan et al. (2022) construct more general network structures and prove lower bounds under broader conditions, including the special case where the objective satisfies the Polyak–Łojasiewicz (PL) condition. Huang and Yuan (2022) extend the analysis to the time-varying networks. In addition, Huang et al. (2022) and He et al. (2023) investigate lower bounds under communication compression.

D-NSS Algorithm

In this section, we design D-NSS (Decentralized Optimization with Node-Specific Sampling), a sample-efficient algorithm under variance heterogeneity, and analyze its sample and communication complexities.

Algorithm Design

The design of the algorithm is motivated by formulating each iteration as a sample minimization problem with a target accuracy. If each node draws B_i stochastic gradients per iteration, the mean squared error of the global averaged gradient estimator is given by

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^{m}\frac{1}{B_i}\sum_{j=1}^{B_i} g_i(x^{(t)};\xi_{ij}) - \nabla f(x^{(t)})\right\|^2\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\frac{1}{B_i^2}\sum_{j=1}^{B_i}\mathbb{E}\left[\left\|g_i(x^{(t)};\xi_{ij}) - \nabla f_i(x^{(t)})\right\|^2\right] \\
&\leq \frac{1}{m^2}\sum_{i=1}^{m}\frac{\sigma_i^2}{B_i}, \qquad (3)
\end{aligned}$$

where the last step holds by equation (2). We desire to achieve the estimation accuracy of ϵ^2 based on equation (3). A simple choice for batch size setting is $B_i = \sigma_i^2/(m\epsilon^2)$, yielding a total sample size of $\bar{\sigma}_{\mathrm{QM}}^2/\epsilon^2$. Although Xin et al. (2021) use a different choice with $B_i = \bar{\sigma}_{\mathrm{QM}}^2/(m\epsilon^2)$, the total sample size per iteration remains the same. A natural question is how to minimize the total number of samples under the given accuracy. We formulate it as the following optimization problem:

$$\min_{B_i} \quad \sum_{i=1}^{m} B_i \quad \text{s.t.} \quad \frac{1}{m^2}\sum_{i=1}^{m}\frac{\sigma_i^2}{B_i} \leq \epsilon^2, \quad B_i > 0. \qquad (4)$$

It is a convex problem and its optimal solution can be characterized by the KKT conditions. To handle the inequality constraint, we introduce the Lagrange multiplier $\lambda \geq 0$ and define the Lagrangian as

$$\mathcal{L}(B,\lambda) = \sum_{i=1}^{m} B_i + \lambda\left(\sum_{i=1}^{m}\frac{\sigma_i^2}{B_i} - m^2\epsilon^2\right).$$

By the dual feasibility and first-order optimality, we have

$$B_i^\star = \frac{\sigma_i\sum_{j=1}^{m}\sigma_j}{m^2\epsilon^2}, \qquad \text{where } i = 1,\ldots,m. \qquad (5)$$

Algorithm 1: D-NSS

Input: Initial point $x^0 \in \mathbb{R}^d$, step size $\eta > 0$, communication matrix W, communication rounds R_t at iteration t, batch size B_i at node i

Initialize: $s_i^{-1} = y_i^{-1} = 0$ for all i

for $t = 0, 1, \ldots, T-1$ **do**
  for $i = 1, \ldots, m$ **in parallel do**
    Sample i.i.d. mini-batch $\mathcal{S}_i^t = \{\xi_{i,1}, \ldots, \xi_{i,B_i}\}$
    $y_i^t = \frac{1}{B_i} \sum_{j=1}^{B_i} g_i(x_i^t; \xi_{i,j})$
    $s_i^t = \texttt{FastMix}(\{s_i^{t-1} + y_i^t - y_i^{t-1}\}_{i=1}^m, W, R_t)$
    $x_i^{t+1} = \texttt{FastMix}(\{x_i^t - \eta s_i^t\}_{i=1}^m, W, R_t)$
  end for
end for

Output: $x_{i,\text{out}}$ uniformly sampled from $\{x_i^0, x_i^1, \ldots, x_i^T\}$

Algorithm 2: $\texttt{FastMix}(\{\phi_i\}_{i=1}^m, W, R)$

Input: Initial value $\{\phi_i\}_{i=1}^m$, communication matrix W, number of rounds R

Initialize: $\eta = (1 - \sqrt{1 - \lambda_2^2(W)})/(1 + \sqrt{1 - \lambda_2^2(W)})$, $z_i^0 = z_i^{-1} = \phi_i$.

for $r = 0, 1, \ldots, R-1$ **do**
  $z_i^{r+1} = (1+\eta) \sum_{j=1}^m W_{ij} z_j^r - \eta z_i^{r-1}$
end for

Output: z_i^R

Therefore, the minimal total number of samples to achieve an ϵ-accurate estimation is $\sum_{i=1}^m B_i^\star = \bar{\sigma}_{\text{AM}}^2/\epsilon^2$, which depends on the arithmetic mean of the standard deviations.

As discussed above, existing decentralized optimization algorithms (Lian et al. 2017; Yi et al. 2022; Lu and De Sa 2021, 2023; Yuan et al. 2022; Xin et al. 2021) typically adopt uniform sampling across nodes, which ignores substantial variance heterogeneity and leads to suboptimal dependence in sample complexity.

Based on the optimal sampling strategy given in equation (5), we propose D-NSS, a decentralized algorithm that allocates node-specific batch sizes, described in Algorithm 1. At iteration t, node i holds a local variable $x_i^t \in \mathbb{R}^d$ and computes a stochastic gradient estimator y_i^t using a mini-batch of size B_i, where B_i is determined by the local gradient noise standard deviation σ_i. The local variable x_i^t is then updated using a gradient tracking variable s_i^t, which combines y_i^t with information communicated from the neighboring nodes $\mathcal{N}_i$. Fast consensus is achieved through the multi-consensus step `FastMix` (Algorithm 2). With this design, D-NSS attains the theoretically optimal sample complexity and is suited for decentralized settings with heterogeneous variance.

Complexity Analysis of D-NSS

To analyze the complexity of the D-NSS algorithm, we first present several standard assumptions in decentralized stochastic optimization.

Assumption 1. *The objective function f is lower bounded, i.e.,* $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$.

In addition, we denote the initial optimal function value gap by $\Delta = f(x^0) - \inf_{x \in \mathbb{R}^d} f(x)$, where x^0 is the initial point of the algorithm. It holds $\Delta < +\infty$ under Assumption 1.

Assumption 2. *For each node i, the stochastic gradient $g_i(x, \xi)$ satisfies unbiasedness and bounded variance as*

$$\mathbb{E}_\xi[g_i(x;\xi)] = \nabla f_i(x) \quad \textit{and}$$
$$\mathbb{E}_\xi\left[||g_i(x;\xi) - \nabla f_i(x)||^2\right] \le \sigma_i^2,$$

where $\sigma_i > 0$.

Assumption 2 captures the variance heterogeneity across nodes, which is central to the problem studied in this work.

Assumption 3. *The global objective f is L-smooth, and each local function f_i is mL-smooth, i.e., for all $x, y \in \mathbb{R}^d$, we have*

$$\|\nabla f(x) - \nabla f(y)\|^2 \le L^2 \|x - y\|^2,$$

and for any node i and $x, y \in \mathbb{R}^d$, we have

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \le m^2 L^2 \|x - y\|^2.$$

Remark 1. *We emphasize our Assumption 3 is weaker than the smoothness condition in existing works that requires every f_i satisfying $\|\nabla f_i(x) - \nabla f_i(y)\|^2 \le L^2\|x-y\|^2$ (Lu and De Sa 2021, 2023; Yuan et al. 2022; Xin et al. 2021). See Appendix A for details.*

Finally, we introduce the assumption on the communication matrix W, which is standard in the decentralized optimization (Ye et al. 2023; Bai, Liu, and Luo 2024).

Assumption 4. *Let $W \in \mathbb{R}^{m \times m}$ be the communication matrix, satisfying:*

(a) W is symmetric and element-wise nonnegative, with $W_{ij} \neq 0$ if and only if nodes i and j are connected or $i = j$;

(b) $0 \preceq W \preceq I$ and $W^\top \mathbf{1}_m = W \mathbf{1}_m = \mathbf{1}_m$; moreover, the null space of $(I - W)$ is $\text{span}(\mathbf{1})$.

Assumption 4 indicates that $1 - \lambda_2(W) > 0$, where $\lambda_2(W)$ is the second largest eigenvalue of W. We define $\chi := 1 - \lambda_2(W)$.

Following theorem provides upper bounds on the sample and communication complexity required by Algorithm 1 to reach an ϵ-stationary point.

Theorem 1. *Under Assumptions 1–4, consider Algorithm 1 with the following parameter choices:*

$$\eta = \frac{1}{2L}, \quad B_i = \left\lceil \frac{16\sigma_i \sum_{j=1}^m \sigma_j}{m^2 \epsilon^2} \right\rceil, \quad T = \left\lceil \frac{32 \Delta L}{\epsilon^2} \right\rceil,$$

$$R_0 = O\left(\frac{1}{\sqrt{\chi}} \log\left(\frac{m}{\epsilon}\right)\right), \textit{ and } R_t = O\left(\frac{1}{\sqrt{\chi}} \log(m)\right),$$

then the output of the algorithm is an ϵ-stationary point satisfying $\mathbb{E}\left[\|\nabla f(x_{i,\text{out}})\|^2\right] \le \epsilon^2$. The sample complexity is upper bounded by

$$O\left(\frac{\Delta L \bar{\sigma}_{\text{AM}}^2}{\epsilon^4} + \frac{m \Delta L}{\epsilon^2}\right),$$

and the communication complexity is bounded by

$$\tilde{O}\left(\frac{\Delta L}{\sqrt{\chi}\epsilon^2}\right).$$

In Table 1, we compare the result with representative existing algorithms. Since all these methods achieve near-optimal (up to a logarithmic factor) communication complexity, we focus on their sample complexity. D-NSS achieves a dependence on the arithmetic mean of standard deviations $\bar{\sigma}_{\mathrm{AM}}$, which is tighter than the worst-case or quadratic-mean bounds in previous methods. Under significant variance heterogeneity, the parameters can satisfy the relation $\sigma_{\max} = \Theta(\sqrt{m}\,\bar{\sigma}_{\mathrm{QM}}) = \Theta(m\,\bar{\sigma}_{\mathrm{AM}})$. Therefore, in such scenarios, our D-NSS algorithm achieves an $O(m)$ improvement in sample efficiency over existing methods.

Lower Bounds Under Heterogeneous Variance

In the previous section, we established upper bounds for decentralized non-convex stochastic optimization under heterogeneous variance. Although the proposed algorithm achieves a sample complexity that depends on the arithmetic mean $\bar{\sigma}_{\mathrm{AM}}$, it remains unclear whether this dependence can be further reduced through more refined algorithmic designs. To address this, we investigate the lower bounds under heterogeneous variance. We begin by formally defining the class of algorithms.

Definition 1 (Decentralized first-order algorithm class)**.** *A decentralized first-order algorithm is defined over a network of nodes and satisfies the following constraints:*

- ***Local memory****: At time t, each node i maintains a local memory $\mathcal{M}_{i,t} \subset \mathbb{R}^d$ that stores previously accessed or generated information. The memory is updated through either local computation or communication, i.e.,*

$$\mathcal{M}_{i,t} \subseteq \mathcal{M}_{i,t}^{comp} \cup \mathcal{M}_{i,t}^{comm},$$

  where $\mathcal{M}_{i,t}^{comp}$ and $\mathcal{M}_{i,t}^{comm}$ represent the computational and communication memories, respectively.
- ***Local computation****: At time t, node i can query a local first-order stochastic oracle to access $g_i(x;\xi_i)$ for any $x \in \mathcal{M}_{i,t-1}$. The computational memory is given by*

$$\mathcal{M}_{i,t}^{comp} = \mathrm{Span}\left(\{x,\, g_i(x;\xi_i) : x \in \mathcal{M}_{i,t-1}\}\right).$$

- ***Local communication****: At time t, node i can receive information from its neighbours $\mathcal{N}(i)$. The communication memory is defined as*

$$\mathcal{M}_{i,t}^{comm} = \mathrm{Span}\left(\bigcup_{j\in\mathcal{N}(i)} \mathcal{M}_{j,t-\tau}\right),$$

  where $\tau < t$ denotes the communication delay parameter.
- ***Output value****: At time t, node i must output one vector from its local memory as its current output, i.e.,*

$$x_i^t \in \mathcal{M}_{i,t}.$$

Establishing lower bounds for decentralized stochastic optimization under heterogeneous variance presents two key challenges. First, existing results (Lu and De Sa 2021, 2023; Yuan et al. 2022) rely on constructions in which all local objectives are identical, i.e., $f_1 = f_2 = \cdots = f_m$. Such constructions are suitable for uniform-variance settings but fail to capture the difficulties introduced by heterogeneous variance.

Algorithm	Sample Complexity
DeTAG Lu and De Sa (2023)	$O\left(\frac{\Delta L\sigma_{\max}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right)$
MG-DSGD Yuan et al. (2022)	$O\left(\frac{\Delta L\sigma_{\max}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right)$
GT-SA Xin et al. (2021)	$O\left(\frac{\Delta L\bar{\sigma}_{\mathrm{QM}}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right)$
D-NSS Theorem 1	$O\left(\frac{\Delta L\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right)$
Lower Bound Theorem 2	$\Omega\left(\frac{\Delta L\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right)$

Table 1: Comparison of sample complexity for first-order decentralized methods in the non-convex stochastic setting with heterogeneous variance.

Motivated by lower bounds for finite-sum problems (Zhou and Gu 2019), we address this limitation by constructing orthogonal local functions that satisfy $\langle \nabla f_i(x), \nabla f_j(x)\rangle = 0$ for all $i \neq j$. Second, heterogeneous variance breaks the symmetry across nodes, as algorithms typically allocate different numbers of samples to different nodes. This asymmetry introduces additional challenges in the analysis that are not addressed by existing lower bound techniques. Our construction explicitly incorporates this sample allocation asymmetry. The detailed construction and proof are provided in the appendix.

We state the main result as follows.

Theorem 2. *For any algorithm $\mathcal{A}$ satisfying Definition 1, there exists a distributed objective function of the form $f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)$ with corresponding stochastic gradients satisfying Assumptions 1–3, such that, for sufficiently small ϵ, the number of samples required to find an ϵ-stationary point is at least*

$$\Omega\left(\frac{\Delta L\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^4} + \frac{m\Delta L}{\epsilon^2}\right).$$

The result in Theorem 2 shows that the D-NSS algorithm achieves the optimal sample complexity, and its dependence on the arithmetic mean $\bar{\sigma}_{\mathrm{AM}}$ is tight. Moreover, since the construction of the communication lower bound does not involve stochastic gradients, the lower bound matches that in Yuan et al. (2022) and is given by

$$\Omega\left(\frac{\Delta L}{\sqrt{\chi}\epsilon^2}\right).$$

Therefore, the D-NSS algorithm achieves nearly optimal communication complexity (up to a logarithmic factor).

D-NSS-VR Algorithm

In non-convex stochastic optimization, the optimal sample complexity is $O(\epsilon^{-4})$ under the standard bounded-variance

assumption. When the mean-squared smoothness condition holds, this complexity can be improved to $O(\epsilon^{-3})$ (Arjevani et al. 2023). A variety of variance-reduction methods, such as SARAH, SPIDER, PAGE, and their decentralized extensions, have been developed to attain this improved rate.

We have established a complexity bound that depends on the arithmetic mean $\bar{\sigma}_{\mathrm{AM}}$ in the general setting. In contrast, decentralized variance-reduction methods involve more intricate structures, such as recursive gradient updates, nested inner–outer loops, and synchronization of local information, which require more delicate parameter tuning. Whether a similar dependence on $\bar{\sigma}_{\mathrm{AM}}$ can be preserved within such a framework remains an open question.

In this section, we address this question by extending the sampling strategy to the variance-reduced setting. We propose D-NSS-VR (Decentralized Node-Specific Sampling with Variance Reduction), a decentralized algorithm that incorporates node-specific sampling into a SARAH-type variance reduction framework (Nguyen et al. 2017). We show that D-NSS-VR also achieves sample complexity that depends on the arithmetic mean $\bar{\sigma}_{\mathrm{AM}}$.

Algorithm Design

The core idea of variance reduction is to construct gradient estimates recursively using historical information and small mini-batches in most iterations, thereby significantly reducing sample usage while maintaining estimation accuracy.

As presented in Algorithm 3, D-NSS-VR integrates node-specific sampling with recursive gradient updates. During the large-batch update phase, the algorithm follows the design of D-NSS by allocating to each node i a batch size B_i proportional to its local noise level σ_i. This estimate serves as the initialization for subsequent recursive updates. In the inner update phase, a fixed mini-batch size b is employed across all nodes to construct variance-reduced gradient estimates recursively, as specified in line 15 of Algorithm 3, which further improves the sample complexity.

Moreover, inspired by the PAGE method (Li et al. 2021), the algorithm employs a probabilistic update scheme in place of a fixed inner–outer loop, making the overall framework simpler and more unified from an analytical perspective. At the implementation level, D-NSS-VR also introduces a skip variable w_i^t, which allows each node to reuse the previous gradient estimate in certain iterations, thereby reducing computational overhead. In particular, as the noise level tends to zero, the algorithm reduces to a variance reduction method as in the finite-sum setting.

Algorithm 3: D-NSS-VR

Input: Initial point $x^0 \in \mathbb{R}^d$, step size $\eta > 0$, communication matrix W, communication rounds R_t at iteration t, batch sizes B_i and b, probability parameters $p, q \in (0, 1]$

1: For all i, sample large batch $\mathcal{S}_i^0 = \{\xi_{i,1}, \ldots, \xi_{i,B_i}\}$
2: $y_i^0 = \dfrac{1}{B_i} \sum_{j=1}^{B_i} g_i(x_i^0; \xi_{i,j})$;
3: $s_i^0 = \texttt{FastMix}(\{y_i^0\}_{i=1}^m, W, R_0)$;
4: $x_i^1 = \texttt{FastMix}(\{x_i^0 - \eta s_i^0\}_{i=1}^m, W, R_0)$
5: **for** $t = 1, 2, \ldots, T-1$ **do**
6: Sample $\zeta_t \sim$ Bernoulli(p)
7: **for** $i = 1, \ldots, m$ **in parallel do**
8: **if** $\zeta_t = 1$ **then**
9: Sample large batch $\mathcal{S}_i^t = \{\xi_{i,1}, \ldots, \xi_{i,B_i}\}$
10: $y_i^t = \dfrac{1}{B_i} \sum_{j=1}^{B_i} g_i(x_i^t; \xi_{i,j})$
11: **else**
12: Sample $\omega_i^t \sim$ Bernoulli(q)
13: **if** $\omega_i^t = 1$ **then**
14: Sample mini-batch $\mathcal{S}_i^t = \{\xi_{i,1}, \ldots, \xi_{i,b}\}$
15: $y_i^t = y_i^{t-1} + \dfrac{\omega_i^t}{bq} \sum_{j=1}^{b} \Big(g_i(x_i^t; \xi_{i,j}) - g_i(x_i^{t-1}; \xi_{i,j}) \Big)$
16: **else**
17: $y_i^t = y_i^{t-1}$
18: **end if**
19: **end if**
20: $s_i^t = \texttt{FastMix}(\{s_i^{t-1} + y_i^t - y_i^{t-1}\}_{i=1}^m, W, R_t)$
21: $x_i^{t+1} = \texttt{FastMix}(\{x_i^t - \eta s_i^t\}_{i=1}^m, W, R_t)$
22: **end for**
23: **end for**

Output: $x_{i,\mathrm{out}}$ uniformly sampled from $\{x_i^0, x_i^1, \ldots, x_i^T\}$

Complexity Analysis of D-NSS-VR

The convergence analysis of variance reduction methods relies on the following mean-squared smoothness assumption.

Assumption 5 (Mean-Squared Smoothness)**.** *There exists a constant $\bar{L} > 0$ such that for any $x, y \in \mathbb{R}^d$,*

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\xi_i} \left\| g_i(x, \xi_i) - g_i(y, \xi_i) \right\|^2 \leq \bar{L}^2 \|x - y\|^2.$$

Assumption 5 is weaker than that in previous works (Pan, Liu, and Wang 2020; Sun, Lu, and Hong 2020; Xin, Khan, and Kar 2021a; Xin et al. 2021), which require mean-squared smoothness of the stochastic gradient for each local function. Moreover, this assumption also implies that the global objective function f is $\bar{L}$-smooth.

The following theorem provides upper bounds on the sampling and communication complexity of Algorithm 3 for finding an ϵ-stationary point.

Theorem 3. *Under Assumptions 1, 2, 4, and 5, consider Algorithm 3 with the following parameter choices:*

$$\eta = \frac{1}{48\bar{L}}, \quad B_i = \max\left\{ \left\lceil \frac{32\sigma_i \sum_{j=1}^m \sigma_j}{m^2 \epsilon^2} \right\rceil, 1 \right\},$$

$$b = \left\lceil \frac{\sqrt{\sum_{i=1}^m B_i}}{m} \right\rceil, \quad q = \frac{\sqrt{\sum_{i=1}^m B_i}}{bm},$$

$$p = \frac{bq}{bq + \sum_{i=1}^m B_i/m}, \quad T = \left\lceil \frac{384\Delta\bar{L}}{\epsilon^2} + \frac{2}{p} \right\rceil,$$

$$R_0 = O\left(\frac{1}{\sqrt{\chi}} \log\left(\frac{m}{\epsilon} \right) \right), \text{ and } R_t = O\left(\frac{1}{\sqrt{\chi}} \log m \right),$$

Algorithm	Sample Complexity	Communication Rounds
GT-HSGD Xin, Khan, and Kar (2021a)	$O\left(\frac{(\Delta\bar{L}+\bar{\sigma}_{\mathrm{QM}}^2)^{3/2}}{\epsilon^3}\right)$	$O\left(\frac{(\Delta\bar{L}+\bar{\sigma}_{\mathrm{QM}}^2)^{3/2}}{\epsilon^3}\right)$
GT-SR-O Xin et al. (2021)	$O\left(\frac{\Delta\bar{L}\bar{\sigma}_{\mathrm{QM}}}{\epsilon^3}+\frac{\bar{\sigma}_{\mathrm{QM}}^2}{\epsilon^2}+\frac{m\Delta\bar{L}}{\epsilon^2}+m\right)$	$\tilde{O}\left(\frac{\Delta\bar{L}}{\sqrt{\chi}\epsilon^2}+\frac{\bar{\sigma}_{\mathrm{QM}}}{\sqrt{\chi}\epsilon}\right)$
D-NSS-VR Theorem 3	$O\left(\frac{\Delta\bar{L}\bar{\sigma}_{\mathrm{AM}}}{\epsilon^3}+\frac{\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^2}+\frac{\sqrt{m}\Delta\bar{L}}{\epsilon^2}+m\right)$	$\tilde{O}\left(\frac{\Delta\bar{L}}{\sqrt{\chi}\epsilon^2}+\frac{\bar{\sigma}_{\mathrm{AM}}}{\sqrt{\chi}\epsilon}\right)$
Lower Bound Theorem 4	$\Omega\left(\frac{\Delta\bar{L}\bar{\sigma}_{2/3}}{\epsilon^3}+\frac{\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^2}+\frac{\sqrt{m}\Delta\bar{L}}{\epsilon^2}+m\right)$	$\Omega\left(\frac{\Delta\bar{L}}{\sqrt{\chi}\epsilon^2}\right)$

Table 2: Comparison of sample and communication complexity for decentralized variance reduction methods under mean-squared smoothness and heterogeneous variance.

then the output $x_{i,\mathrm{out}}$ of the algorithm is an ϵ-stationary point satisfying $\mathbb{E}\left[\|\nabla f(x_{i,\mathrm{out}})\|^2\right] \leq \epsilon^2$. The sample complexity is upper bounded by

$$O\left(\frac{\Delta\bar{L}\bar{\sigma}_{\mathrm{AM}}}{\epsilon^3}+\frac{\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^2}+\sqrt{m}\frac{\Delta\bar{L}}{\epsilon^2}+m\right),$$

and the communication complexity is bounded by

$$\tilde{O}\left(\frac{\Delta\bar{L}}{\sqrt{\chi}\epsilon^2}+\frac{\bar{\sigma}_{AM}}{\sqrt{\chi}\epsilon}\right).$$

Theorem 3 shows that the variance reduction method can also achieve a dependence on the arithmetic mean of standard deviations. As shown in Table 2, we compare the sample and communication complexity of D-NSS-VR with state-of-the-art decentralized variance reduction methods (Xin, Khan, and Kar 2021a; Xin et al. 2021). D-NSS-VR exhibits better dependence on the variance parameters, yielding an $O(\sqrt{m})$ improvement in sample complexity under strong heterogeneity. Moreover, when all variances $\sigma_i \to 0$, the sample complexity of D-NSS-VR matches that of variance reduction method for the finite-sum problem (Li et al. 2021), which is not achieved by the compared methods.

Lower Bounds with Mean-Squared Smoothness

We establish the following lower bound on the sample complexity for variance reduction algorithms.

Theorem 4. *For any algorithm $\mathcal{A}$ satisfying Definition 1, there exists a distributed objective function of the form $f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)$ with corresponding stochastic gradients satisfying Assumptions 1, 2, and 5, such that, for sufficiently small ϵ, the number of stochastic gradient samples required to find an ϵ-stationary point is at least*

$$\Omega\left(\frac{\Delta\bar{L}\bar{\sigma}_{2/3}}{\epsilon^3}+\frac{\bar{\sigma}_{\mathrm{AM}}^2}{\epsilon^2}+\frac{\sqrt{m}\Delta\bar{L}}{\epsilon^2}+m\right),$$

where $\bar{\sigma}_{2/3} = ((\sigma_1^{2/3}+\cdots+\sigma_m^{2/3})/m)^{3/2}$.

Under mean-squared smoothness assumption, the communication lower bound remains

$$\Omega\left(\frac{\Delta\bar{L}}{\sqrt{\chi}\epsilon^2}\right),$$

as its construction does not rely on the stochastic gradients (Yuan et al. 2022).

Remark 2. *Theorems 3 and 4 show that while several terms in the upper bounds match the corresponding lower bounds, a gap remains in the leading term. Due to the inherent difficulty of constructing lower bounds, particularly under the mean-squared smoothness assumption, closing this gap is left as an open problem for future work. Although the upper and lower bounds do not fully match, our result demonstrates that the complexity of variance reduction methods can depend on $\bar{\sigma}_{\mathrm{AM}}$. Finally, the communication complexity is nearly optimal in the small-ϵ regime.*

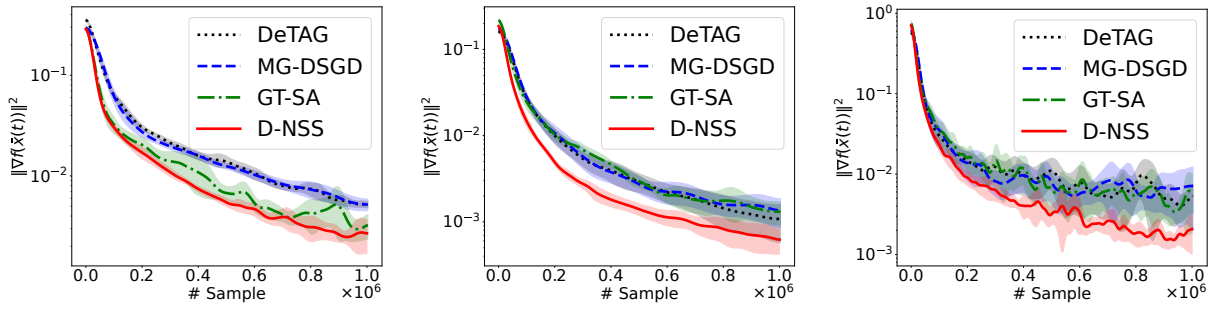
Numerical Experiments

In this section, we validate the convergence of the proposed algorithms through numerical experiments on multiple datasets and compare them with existing methods. Specifically, we consider a widely used regularized logistic regression problem for binary classification in the decentralized setting (Xin, Khan, and Kar 2021a; Luo et al. 2022; Gao et al. 2023):

$$f_i(x) = \frac{1}{N_i}\sum_{j=1}^{N_i}\log\left(1+\exp\left(-b_{ij}a_{ij}^\top x\right)\right)+r\sum_{k=1}^{d}\frac{[x]_k^2}{1+[x]_k^2},$$

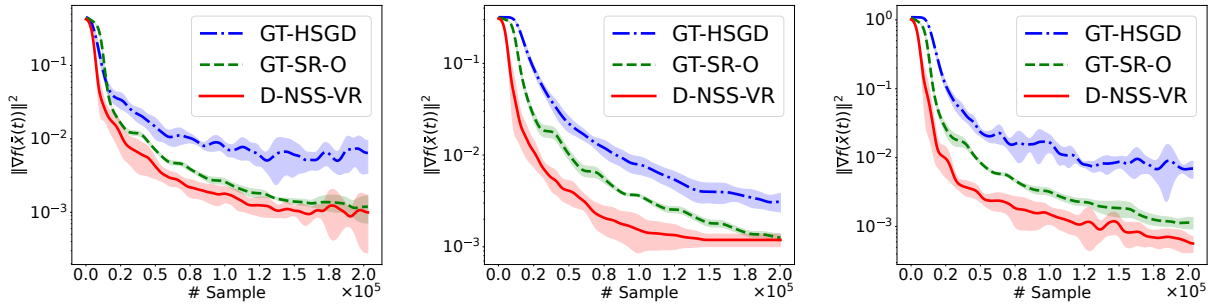
where $a_{ij} \in \mathbb{R}^d$ is the feature vector of the j-th sample at node i, $b_{ij} \in \{\pm 1\}$ is the corresponding binary label, and $r > 0$ is the regularization parameter, which is set to 10^{-4} in our experiments. Each node i holds N_i local samples, which may vary across nodes. To explicitly model heterogeneous variance across nodes, Gaussian noise with node-specific variances is added to the stochastic gradients.

We conduct numerical experiments on three real-world datasets: `a9a`, `w8a`, and `MNIST`. The datasets `a9a` ($N = 32,561$, $d = 123$) and `w8a` ($N = 49,749$, $d = 300$) can



(a) a9a (b) w8a (c) MNIST

Figure 1: Performance comparison of decentralized algorithms in terms of the number of samples on datasets a9a, w8a, and mnist. The lines represent averages over 5 runs, and the shaded regions denote the standard deviations.



(a) a9a (b) w8a (c) MNIST

Figure 2: Performance comparison of decentralized variance reduction algorithms in terms of the number of samples on datasets a9a, w8a, and mnist. The lines represent averages over 5 runs, and the shaded regions denote the standard deviations.

be downloaded from the LIBSVM repository (Chang and Lin 2011). For the MNIST dataset (LeCun et al. 2002), the digits 4 and 5 are utilized for the classification task, with $N = 11,263$ and $d = 784$. We set the number of nodes $m = 20$. Regarding the communication network, a random graph with $\chi = 0.41$ is used and Metropolis-Hastings weights (Xiao, Boyd, and Lall 2005) are applied to construct the communication matrix W.

We first compare D-NSS (Algorithm 1) with three baseline algorithms: DeTAG (Lu and De Sa 2023), MG-DSGD (Yuan et al. 2022), and GT-SA (Xin et al. 2021). For each baseline, the batch sizes and step sizes are tuned to achieve their best empirical performance. In D-NSS, each σ_i is estimated by computing the variance on a small batch (50 on each node) of stochastic gradients and broadcast at initialization. The additional cost of this variance estimation and communication is negligible compared with the overall cost. The experimental results are presented in Figure 1. It can be observed that the proposed D-NSS consistently outperforms the baseline methods across all three datasets, which demonstrates the effectiveness of the node-specific sampling strategy.

For the variance-reduced method D-NSS-VR (Algorithm 3), we compare it with the state-of-the-art variance-reduced algorithms GT-HSGD (Xin, Khan, and Kar 2021a) and GT-SR-O (Xin et al. 2021). The results are presented in Figure 2. It can be observed that D-NSS-VR also consistently achieves the best performance across all three datasets. Although the incorporation of randomness (lines 6 and 12 in Algorithm 3) increases the variance, the combination with the node-specific sampling strategy leads to a significant advantage in practical performance.

Conclusion

This work studies decentralized non-convex stochastic optimization under heterogeneous variance. We propose D-NSS, a node-specific sampling algorithm that allocates samples according to local noise levels, and establish a sample complexity bound depending on the arithmetic mean of local standard deviations. A matching lower bound is derived, demonstrating the optimality of the proposed approach. Furthermore, by incorporating variance reduction, we develop D-NSS-VR, which achieves improved sample complexity under mean-squared smoothness while maintaining the arithmetic-mean dependence. Numerical experiments on multiple datasets validate our theoretical results and demonstrate the practical advantages of the proposed algorithms. Future work includes closing the remaining gap in the lower bound for variance reduction methods and investigating the performance of higher-order algorithms under heterogeneous variance.

Acknowledgments

Luo is supported by the Major Key Project of Pengcheng Laboratory (No. PCL2024A06), National Natural Science Foundation of China (No. 12571557), National Natural Science Foundation of China (No. 62206058), and Shanghai Basic Research Program (23JC1401000). Chen and Wei were partially supported by the National Key R&D Program of China (No. 2023YFA1009300).

References

Arjevani, Y.; Carmon, Y.; Duchi, J. C.; Foster, D. J.; Srebro, N.; and Woodworth, B. 2023. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214.

Arjevani, Y.; and Shamir, O. 2015. Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28.

Assran, M.; Loizou, N.; Ballas, N.; and Rabbat, M. 2019. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, 344–353. PMLR.

Bai, Y.; Liu, Y.; and Luo, L. 2024. On the Complexity of Finite-Sum Smooth Optimization under the Polyak-{\L} ojasiewicz Condition. *arXiv preprint arXiv:2402.02569*.

Chang, C.-C.; and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3): 1–27.

Cutkosky, A.; and Orabona, F. 2019. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32.

Deng, K.; and Hu, J. 2025. Decentralized Projected Riemannian Stochastic Recursive Momentum Method for Nonconvex Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11203–11211.

Di Lorenzo, P.; and Scutari, G. 2016. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2): 120–136.

Fang, C.; Li, C. J.; Lin, Z.; and Zhang, T. 2018. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31.

Gao, J.; Liu, X.-W.; Dai, Y.-H.; Huang, Y.; and Gu, J. 2023. Distributed stochastic gradient tracking methods with momentum acceleration for non-convex optimization. *Computational Optimization and Applications*, 84(2): 531–572.

He, Y.; Huang, X.; Chen, Y.; Yin, W.; and Yuan, K. 2023. Lower bounds and accelerated algorithms in distributed stochastic optimization with communication compression. *arXiv preprint arXiv:2305.07612*.

Huang, X.; Chen, Y.; Yin, W.; and Yuan, K. 2022. Lower bounds and nearly optimal algorithms in distributed learning with communication compression. *Advances in Neural Information Processing Systems*, 35: 18955–18969.

Huang, X.; and Yuan, K. 2022. Optimal complexity in non-convex decentralized learning over time-varying networks. *arXiv preprint arXiv:2211.00533*.

Kovalev, D.; Salim, A.; and Richtárik, P. 2020. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33: 18342–18352.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, Z.; Bao, H.; Zhang, X.; and Richtárik, P. 2021. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International conference on machine learning*, 6286–6295. PMLR.

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. *Advances in neural information processing systems*, 30.

Lu, Y.; and De Sa, C. 2021. Optimal complexity in decentralized training. In *International conference on machine learning*, 7111–7123. PMLR.

Lu, Y.; and De Sa, C. 2023. Decentralized learning: Theoretical optimality and practical improvements. *Journal of Machine Learning Research*, 24(93): 1–62.

Luo, L.; Bai, Y.; Chen, L.; Liu, Y.; and Ye, H. 2022. On the Complexity of Decentralized Smooth Nonconvex Finite-Sum Optimization. *arXiv preprint arXiv:2210.13931*.

Nedic, A.; Olshevsky, A.; and Shi, W. 2017. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4): 2597–2633.

Nedic, A.; and Ozdaglar, A. 2009. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1): 48–61.

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*, 2613–2621. PMLR.

Pan, T.; Liu, J.; and Wang, J. 2020. D-SPIDER-SFO: A decentralized optimization algorithm with faster convergence rate for nonconvex problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1619–1626.

Pu, S.; and Nedić, A. 2021. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1): 409–457.

Qu, G.; and Li, N. 2017. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3): 1245–1260.

Scaman, K.; Bach, F.; Bubeck, S.; Lee, Y. T.; and Massoulié, L. 2017. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, 3027–3036. PMLR.

Scutari, G.; and Sun, Y. 2019. Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176: 497–544.

Shi, W.; Ling, Q.; Wu, G.; and Yin, W. 2015. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2): 944–966.

Sun, H.; Lu, S.; and Hong, M. 2020. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International conference on machine learning*, 9217–9228. PMLR.

Tang, H.; Lian, X.; Yan, M.; Zhang, C.; and Liu, J. 2018. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, 4848–4856. PMLR.

Xiao, L.; Boyd, S.; and Lall, S. 2005. A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, 63–70. IEEE.

Xin, R.; Das, S.; Khan, U. A.; and Kar, S. 2021. A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency. *arXiv preprint arXiv:2110.01594*.

Xin, R.; Khan, U.; and Kar, S. 2021a. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *International Conference on Machine Learning*, 11459–11469. PMLR.

Xin, R.; Khan, U. A.; and Kar, S. 2021b. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69: 1842–1858.

Ye, H.; Luo, L.; Zhou, Z.; and Zhang, T. 2023. Multi-consensus decentralized accelerated gradient descent. *Journal of machine learning research*, 24(306): 1–50.

Yi, X.; Zhang, S.; Yang, T.; Chai, T.; and Johansson, K. H. 2022. A primal-dual SGD algorithm for distributed nonconvex optimization. *IEEE/CAA Journal of Automatica Sinica*, 9(5): 812–833.

Yuan, K.; Huang, X.; Chen, Y.; Zhang, X.; Zhang, Y.; and Pan, P. 2022. Revisiting optimal convergence rate for smooth and non-convex stochastic decentralized optimization. *Advances in Neural Information Processing Systems*, 35: 36382–36395.

Yuan, K.; Ling, Q.; and Yin, W. 2016. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3): 1835–1854.

Zhou, D.; and Gu, Q. 2019. Lower bounds for smooth non-convex finite-sum optimization. In *International Conference on Machine Learning*, 7574–7583. PMLR.