Distributed Random Reshuffling Over Networks

Kun Huang, Xiao Li, Member, IEEE, Andre Milzarek, Shi Pu, Member, IEEE, and Junwen Qiu

Abstract—In this paper, we consider distributed optimization problems where \(n\) agents, each possessing a local cost function, collaboratively minimize the average of the local cost functions over a connected network. To solve the problem, we propose a distributed random reshuffling (D-RR) algorithm that invokes the random reshuffling (RR) update in each agent. We show that D-RR inherits favorable characteristics of RR for both smooth strongly convex and smooth nonconvex objective functions. In particular, for smooth strongly convex objective functions, D-RR achieves \(O(1/T^2)\) rate of convergence (where \(T\) counts the epoch number) in terms of the squared distance between the iterate and the global minimizer. When the objective function is assumed to be smooth nonconvex, we show that D-RR drives the squared norm of the gradient to 0 at a rate of \(O(1/T^2)\). These convergence results match those of centralized RR (up to constant factors) and outperform the distributed stochastic gradient descent (DSGD) algorithm if we run a relatively large number of epochs. Finally, we conduct a set of numerical experiments to illustrate the efficiency of the proposed D-RR method on both strongly convex and nonconvex distributed optimization problems.

Index Terms—Distributed optimization, random reshuffling, stochastic gradient methods.

I. INTRODUCTION

In this paper, we consider solving the following optimization problems by a group of agents \(\{1, 2, \ldots, n\}\), connected over a network:

\[
\min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x) \quad \text{with} \quad f_i(x) = \frac{1}{m} \sum_{\ell=1}^{m} f_i,\ell(x),
\]

where each \(f_i : \mathbb{R}^p \rightarrow \mathbb{R}\) is a local cost function associated with the local private dataset of agent \(i\), and \(m\) denotes the number of data points or mini-batches in each local dataset. The finite sum structure of \(f_i\) naturally appears in many machine learning and signal processing problems that often involve a large amount of data, i.e., \(nm\) can be prohibitively large. Designing efficient distributed algorithms to solve Problem (1) has attracted great interest in recent years. In particular and initiated by the work [1], distributed algorithms implemented over networked agents with no central controller have become popular choices. In this setting, the agents only exchange information with their immediate neighbors in the network, which can help avoid the communication bottleneck of centralized protocols and increase algorithmic flexibility as well as the robustness to link and node failures [2], [3].

Due to the large size of data, distributed stochastic gradient (SG) methods implemented over networks\(^1\) have been studied extensively to solve Problem (1); see, e.g., [4], [5], [6], [7], [8], [9]. These methods have been shown to be efficient, among which some enjoy the comparable performance to the centralized stochastic gradient descent (SGD) algorithm under certain conditions [6], [9], [10], [11], [12], [13]. Moreover, targeting the finite sum structure of Problem (1), various distributed variance reduction (VR)-based methods have been developed to improve the algorithmic performance [14], [15]. Nevertheless, despite the existing SG- and VR-based (distributed) optimization schemes, random reshuffling (RR) has been a popular and successful method for solving the finite sum optimization problems in practice [16], [17], [18], [19], [20], [21], [22]. Compared to SGD that employs uniform random sampling with replacement at each iteration, RR proceeds in a cyclic sampling fashion. Namely, at each cycle (epoch), the data points or mini-batches are permuted uniformly at random and are then selected sequentially according to the permuted order for gradient computation. Under a centralized computation, RR is provably more efficient than SGD in certain situations (see the literature review subsection for further details) and does not require additional storage costs compared to VR-based methods; see, e.g., [22], [23]. Intuitively, RR allows to utilize all data points in every epoch which can lead to better theoretical and empirical performance. However, the development and study of distributed RR methods over networks seem to be fairly limited and less advanced. This

\(^1\)We also refer as decentralized stochastic gradient methods.
observation motivates the following question: *Can we design an efficient distributed RR algorithm over networks with similar convergence guarantees as centralized RR?*

In this paper, we give an affirmative answer to the above question. To solve Problem (1), we design a novel algorithm termed distributed random reshuffling (D-RR) that invokes the RR update in each agent. We will show that D-RR has comparable convergence properties to RR for both smooth strongly convex and smooth nonconvex objective functions. Here, the term ‘smooth’ refers to objective functions with Lipschitz continuous gradient.

### A. Related Work

There is a vast literature on solving Problem (1) with distributed gradient or stochastic gradient methods; see, e.g., [1], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35]. Among the existing methods, the distributed gradient descent (DGD) algorithm considered in [1] has drawn remarkable attention due to its simplicity and robust performance. When the exact full gradient is not available or hard to evaluate, stochastic gradient methods provide an alternative to reduce the per-iteration sampling cost for solving large-scale machine learning problems. The distributed implementations of stochastic gradient methods over networks, including vanilla distributed stochastic gradient descent (DSDG) and more advanced methods, have been shown to achieve comparable performance to the centralized counterparts [5], [6], [7], [9], [10], [11], [12], [36], [37], [38], [39], [40]. Particularly, recent efforts have been focusing on reducing the *transient times* required by distributed algorithms to obtain the same convergence rate as centralized SGD. E.g., for strongly convex and smooth objective functions, the works [12], [38] have so far achieved the shortest transient time to match the $O(1/n^2)$ convergence rate of SGD, which behaves as $O(n/(1-\lambda))$ with $1-\lambda$ denoting the spectral gap related to the mixing matrix among the agents.

It is worth noting that algorithms based on stochastic gradients also work with online streaming data, which is different from the finite-sum (offline) setting we consider in this work.

RR is widely utilized in practice for tackling large-scale machine learning problems, such as the training of deep neural networks [16], [17], [18], [20], [21], [22]. Experimental evidence [41], [42] indicates that RR often has better empirical performance than SGD. Under the assumptions that the objective function is strongly convex and has Lipschitz Hessian, and the iterates are uniformly bounded, the work [20] establishes $O(1/T^2)$ asymptotic rate of convergence of RR with high probability in terms of the squared distance between the iterate and the unique optimal solution. Based on these motivating observations, a series of works have started to study the convergence behavior of RR; see [21], [22], [23], [24], [43], [44]. For instance, the work [23] establishes $O(1/mT^2)$ convergence rate of RR under the assumptions that each component function in the finite-sum is smooth and strongly convex, where $m$ represents the number of training samples. The authors claimed that this rate outperforms the rate of SGD under a similar setting; see Section 3.1 in [23]. When each component function in the finite-sum is smooth nonconvex and a certain bounded variance-type assumption holds, the works [22], [23] derive a $O(1/m^{1/3}T^{2/3})$ rate of convergence of RR in expectation in terms of the squared norm of gradient. This rate is superior to that of SGD under a similar setting (i.e., $O(1/m^{1/2}T^{1/2})$) after a number of epochs related to the sample size $m$; see the [22, Remark 2]. Very recently, the work [45] establishes strong limit-point convergence results of RR for smooth nonconvex minimization under the Kurdyka-Łojasiewicz inequality.

There are also recent works considering implementing RR over networked agents [46], [47]. In [46], a distributed variance-reduced RR method was introduced and it was shown to enjoy linear convergence for smooth and strongly convex objective functions. The authors in [47] considered a convex, structured problem and showed that the proposed algorithm converges to a neighborhood of the optimal solution in expectation at a sublinear rate. Though both the two works consider distributed RR-type methods, the superiority of RR over distributed SGD-type methods in certain settings was not demonstrated. Table I compares the theoretical results of the related works with those of D-RR.

#### B. Main Contributions

In this work, we propose an efficient algorithm termed distributed random reshuffling (D-RR) for solving distributed optimization problems over networks (see Algorithm 1), which invokes the RR update in each agent. For smooth strongly convex objective function, we conduct a non-asymptotic analysis for D-RR with both constant and decreasing stepsizes. We show that with a decreasing stepsize, D-RR achieves $O(1/mT^2)$ rate of convergence after $T$ epochs in terms of the squared distance between the iterate and the global minimizer (see Theorem 1), where $m$ is defined in (1). Note that this result is comparable to results known for centralized RR algorithms (up to constant factors depending on the network) [22]. In addition, under a constant stepsize $\alpha$, the expected error of D-RR decreases exponentially fast to a neighborhood of 0 with size being of order $O(m^2)$ (see Theorem 2). If the constant is appropriately chosen, D-RR has the same $O(1/mT^2)$ rate of convergence up to logarithmic factors as that of decreasing stepsize (see Corollary 1). The obtained results using constant

---

**TABLE I**

|                  | Strongly Convex | Nonconvex |
|------------------|-----------------|-----------|
| **Stepsizes**    | **Constant**    | **Decreasing** | **Constant** |
| SGD/DGDG        | $O\left(\frac{1}{T}\right)$ [37] | $O\left(\frac{1}{mT}\right)$ [6] | $O\left(\frac{1}{\sqrt{mT}}\right)$ [9] |
| CRR             | $O\left(m\alpha^2\right)$ [23] | $O\left(\frac{1}{mT^2}\right)$ [22] | $O\left(\frac{1}{mT}\right)$ [22], [23] |
| D-RR            | $O\left(m\alpha^2\right)$ | $O\left(\frac{1}{mT^2}\right)$ | $O\left(\frac{1}{T}\right)$ |

For strongly convex objective functions with constant stepsize $\alpha$, we show the size of the final error bounds. The others are complexity results. The Notion $O(\cdot)$ additionally hides the logarithmic factors compared to $O(\cdot)$.
the popularity of centralized RR in centralized settings. We propose the distributed random reshuffling (D-RR) algorithm in Algorithm 1 to solve Problem (1). D-RR can be viewed as a combination of DGD and RR, where the local full gradient descent steps in DGD are replaced by local gradient descent steps that utilize only one of the local (permuted) component functions \( f_{i,t} \) at a time. Specifically, in each epoch \( t \), agent \( i \) first generates a random permutation \( \{ \pi_0^i, \pi_1^i, \ldots, \pi_{m-1}^i \} \) of \( [m] \) and then performs \( m \) stochastic gradient steps accessing the local component functions \( f_{i,\pi_j^i} \) \( \ell \in [m] \), consecutively in a shuffled order. Hence, in contrast to SGD, each agent has guaranteed access to its full local data in every epoch. Notice that such a sampling scheme leads to a biased stochastic gradient estimator, see, e.g., [19].

After agent \( i \) has performed a local stochastic gradient descent step, it sends the intermediate result to its direct neighbors in Line
Algorithm 1: Distributed Random Reshuffling (D-RR).

**Initialization:** Initialize $x_{i,0}$ for each agent $i \in [n]$.

1. for Epoch $t = 0, 1, 2, \ldots, T - 1$ do 
2. for Agent $i$ in parallel do
3. Independently sample a permutation $\{\pi_0, \pi_1, \ldots, \pi_m\}$ of $[m]$
4. Set $x_{0,i} = x_{i,t}$
5. for $\ell = 0, 1, \ldots, m - 1$ do
6. Agent $i$ updates $x_{\ell+1,i} = x_{\ell,i} - \alpha_t \nabla f_{i,\pi_i}(x_{\ell,i})$
7. Agent $i$ receives $x_{\ell+1,j}$ from its neighbors $j \in N_i$
8. end for
9. Set $x_{i,t+1} = x_{m,i}$
10. end for
11. end for
12. Output $x_{i,T}$.

6 of Algorithm 1. The received information is then combined in Line 7. Lines 6–7 are similar to the routine of DGD. From an optimization perspective, Line 7 plays the role of a “projection” for the consensus constraint. Compared to the work [47], where Line 7 is only performed after each epoch, D-RR has better control over the consensus errors.

**Remark 1:** We present an intuitive idea why D-RR works as well as C-RR. Based on Assumption 1, we have the following relation for the averaged iterates over the network agents:

$$x_{i,t+1} = x_{i,t} - \alpha_t \sum_{i=1}^n \nabla f_{i,\pi_i}(x_{i,t}).$$

Notice that Problem (1) can also be written as

$$\min_{x \in \mathbb{R}^p} \frac{1}{m} \sum_{\ell=0}^{m-1} g_{\ell}(x) \quad \text{where} \quad g_{\ell}(x) = \frac{1}{n} \sum_{i=1}^n f_{i,\pi_i}(x).$$

Therefore, (2) can be viewed as approximately implementing the centralized RR method for solving Problem (3), since $\frac{1}{n} \sum_{i=1}^n \nabla f_{i,\pi_i}(x_{\ell,i})$ is close to $\frac{1}{n} \sum_{i=1}^n \nabla f_{i,\pi_i}(x_{\ell,i}')$ when all $x_{\ell,i}'$ are close to $x_{\ell,i}$. To achieve this objective, the consensus error $\sum_{\ell=0}^{m-1} \sum_{i=1}^n \|x_{\ell,i}' - x_{\ell,i}\|^2$ needs to be handled carefully, and thus we implement Line 7 (consensus step) in the inner loop to better control the aforementioned error term. Such an observation is critical for our analysis for D-RR and also explains why D-RR can work.

As a benchmark for the performance of D-RR, we consider a centralized counterpart of Algorithm 1 in Algorithm 2.

Using the notations from Section I-C, Algorithm 1 can be written in a compact form (4):

$$x_{\ell+1,i} = W \left( x_{\ell,i} - \alpha_t \nabla F_{\pi_i}(x_{\ell,i}) \right).$$

Algorithm 2: Centralized Random Reshuffling (C-RR).

**Initialization:** Initialize $x_0$ and stepsize $\alpha_t$.

1. for Epoch $t = 0, 1, 2, \ldots, T - 1$ do
2. Sample $\{\pi_0, \pi_1, \ldots, \pi_m\}$ of $[m]$
3. Set $x_{0,i} = x_t$
4. for $\ell = 0, 1, \ldots, m - 1$ do
5. Update $x_{\ell+1} = x_\ell - \alpha_t \sum_{k=0}^{m-1} \nabla f_{i,\pi_i}(x_{\ell,i})$
6. end for
7. Set $x_{t+1} = x_{m,i}$
8. end for
9. Output $x_{t,T}$.

In the rest of this section, we introduce the roadmaps for studying the convergence properties for D-RR under both strongly-convex objectives and nonconvex objectives. Parts of the analysis follow those in [23].

**B. Roadmap: Strongly-Convex Case**

We first consider $f_{i,\ell}$ satisfying the following assumption.

**Assumption 2:** Each $f_{i,\ell} : \mathbb{R}^p \to \mathbb{R}$ is $\mu-$strongly convex and $L-$smooth, i.e., for all $x, x' \in \mathbb{R}^p$ and $i, \ell$, we have

$$\langle \nabla f_{i,\ell}(x) - \nabla f_{i,\ell}(x'), x - x' \rangle \geq \mu \|x - x'\|^2,$$

$$\|\nabla f_{i,\ell}(x) - \nabla f_{i,\ell}(x')\| \leq L \|x - x'\|.$$

Under Assumption 2, there exists a unique solution $x^* \in \mathbb{R}^p$ to the Problem (1). Moreover, for any $f$ satisfying Assumption 2, we have the following lemma.

**Lemma 2:** Let $f : \mathbb{R}^p \to \mathbb{R}$ satisfy Assumption 2 and let us set $D_f(y, x) := f(y) - f(x) - \langle \nabla f(x), y - x \rangle$. Then

$$\frac{\mu}{2} \|x - y\|^2 \leq D_f(y, x) \leq \frac{L}{2} \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^p,$$

$$\frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \leq D_f(y, x), \quad \forall x, y \in \mathbb{R}^p.$$ (6)

**Proof:** The right-hand side of (5) and (6) are shown in [48, Theorem 2.1.5]. The left-hand side of (5) follows from the definition of $\mu-$strong convexity. \qed

We outline the procedures of the analysis under Assumption 2. According to Remark 1 and the discussions of Section 3.1 in [23], given a permutation $\pi$, for Problem (3), the real limit points for $\ell = 1, \ldots, m$ are defined as

$$x_{\ell} := x^* - \alpha_t \sum_{k=0}^{\ell-1} \nabla g_k(x^*) - \alpha_t \sum_{k=0}^{\ell-1} \nabla f_{i,\pi_i}(x^*)$$

$$\left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,\pi_i}(x^*) \right).$$

Since $x^*$ is the solution to Problem (1), we obtain

$$x_{m} = x^* - \frac{\alpha_t}{n} \sum_{k=0}^{m-1} \nabla f_{i,\pi_i}(x^*)$$

$$= x^* - \frac{\alpha_t}{n} \sum_{k=0}^{m-1} \nabla f_{i,\pi_i}(x^*) = x^*.$$
which is consistent with the observations in [23]. Our next steps are now based on the following principal ideas.

1) For any $\ell$, $t \geq 0$, decompose the errors for the inner loop (8) and the outer loop (9) respectively:

$$\frac{1}{n} \sum_{i=1}^{n} \left\| x_{i,t} - x_{i}^0 \right\|^2 = \left\| x_{t}^\ell - x_{t}^0 \right\|^2 + \frac{1}{n} \sum_{i=1}^{n} \left\| x_{i}^\ell - x_{t}^\ell \right\|^2,$$

(8)

$$\frac{1}{n} \sum_{i=1}^{n} \left\| x_{0,i,t} - x_{i}^0 \right\|^2 = \left\| x_{t}^0 - x_{t}^* \right\|^2 + \frac{1}{n} \sum_{i=1}^{n} \left\| x_{i}^0 - x_{t}^0 \right\|^2,$$

(9)

The first term in (9) is comparable to the error term when studying the performance of centralized RR. The second term is caused by decentralization, this is the consensus error. Such arguments also apply to (8). Dealing with the consensus error is critical and nontrivial.

2) Treat the inner loop and outer loop separately as it can be seen from Algorithm 1 that $x_{0,i,t} = x_{i,t}$ and $x_{0,i,t+1} = x_{i,t+1}$. Specifically, we perform the following four steps to derive the results for the strongly convex case:

i) We first construct two coupled recursions for the terms $E[\|x_{t}^\ell - x_{t}^0\|^2]$ and $E[\|x_{0,i}^t - x_{t}^*\|^2]$ in Lemmas 5 and 7 respectively and introduce a Lyapunov function $H_{t}^\ell$ to decouple the two error terms. Then, we relate the results of the inner loop and the outer loop in Lemma 8.

ii) Using a decreasing stepsize policy $\alpha_t = \frac{\theta}{\mu (t + K)}$, we obtain the upper bounds for $H_t$ in Lemma 9 and $H^\ell_t$ in Lemma 10 which are in the order of $O(1/(t + K)^2)$. In addition, the expected errors of $H_t$ and $H^\ell_t$ decrease exponentially fast to a neighborhood of 0 with size being of order $O(m^2)$ when using a constant stepsize. Such results are also stated in Lemmas 9 and 10, respectively.

iii) Noting that the bounds in Step (ii) can also be applied to $E[\|x_{0,i}^t - x_{i}^*\|^2]$ and $E[\|x_{0,i}^0 - x_{i}^0\|^2]$ for the two stepsize choices, we utilize them in Lemma 7 to obtain a decoupled bound for $E[\|x_{0,i}^t - x_{i}^*\|^2]$ in Lemma 11. Invoking Lemma 11 in Lemma 5, we obtain a decoupled and refined bound for $E[\|x_{0,i}^t - x_{i}^0\|^2]$ in Lemma 12.

iv) Finally, combining (9) and Lemmas 11 and 12, we prove the main results, i.e., Theorem 1 for decreasing stepsizes and Theorem 2 under a constant stepsize.

We highlight the main technical challenges of analyzing D-RR compared to the analysis of SGD-type decentralized methods [6, 12] and that of centralized RR [23]. Note that the decomposition in Step 1 is different from that of SGD-type decentralized methods due to the existence of the real limit point $x_{t}^\ell$ and the two-loop structure of D-RR. Although the analysis of the term $E[\|x_{t}^\ell - x_{t}^0\|^2]$ borrows ideas from [23], the extra consensus term due to decentralization imposes further challenges for constructing the Lyapunov function $H_{t}^\ell$. Utilizing the bound of $H_{t}^\ell$ in both the outer and the inner loops also differs from the analysis of previous SGD-type decentralized methods.

C. Roadmap: Nonconvex Case

In the following, we formalize the assumption for analyzing D-RR when it is utilized to solve smooth nonconvex optimization problems over networks. Basically, we only require smoothness and lower boundedness of the cost functions.

Assumption 3: Each $f_{i,t} : \mathbb{R}^p \to \mathbb{R}$ is $L$-smooth and bounded from below, i.e., for all $x, x' \in \mathbb{R}^p$ and $i, \ell$, we have

$$\left\| \nabla f_{i,t}(x) - \nabla f_{i,t}(x') \right\| \leq L \|x - x'\| \quad \text{and} \quad f_{i,t}(x) \geq \bar{f}_{i,t}.$$ 

From Assumption 3, we can obtain the following lemma. A similar result can be found in, e.g., [23, Proposition 2].

Lemma 3: Let Assumption 3 hold. Then, there exist nonnegative constants $A, B \geq 0$ such that for any $x \in \mathbb{R}^p$, we have

$$\frac{1}{mn} \sum_{i=1}^{n} \sum_{t=1}^{m} \left\| \nabla f_{i,t}(x) - \nabla f(x) \right\|^2 \leq 2A \left( f(x) - \bar{f} \right) + B^2,$$

(10)

where $\bar{f} := \inf_{x \in \mathbb{R}^p} f(x)$, $A = 2L$, and $B^2 = 2L \cdot \left( \bar{f} - \frac{1}{mn} \sum_{i=1}^{m} \bar{f}_{i,t} \right)$.

Proof: See Supplementary Material C-A.

On the one hand, as mentioned in [23, Section 3.3], Lemma 3 bounds the variance of the gradient $\nabla f(x)$. Such a result generalizes the bounded gradient assumption $\|\nabla f_{i,t}(x)\| \leq G$ and the uniformly bounded variance assumption, which is equivalent to (10) when $A = 0$. On the other hand, Lemma 3 also includes the so-called bounded gradient dissimilarity assumption in the distributed setting; see, e.g., in [49, A3]. From such a perspective, the result also characterizes the non-i.i.d. level among the local datasets.

Under Assumption 3, we are able to establish the convergence result for D-RR. The main result is given in Theorem 4. The central idea is to construct a novel Lyapunov function $Q_t$ as in (24) so that we can utilize the technique in Lemma 16. Compared with the analysis in [23] that directly deals with the term $f(x_{0,i}^t) - f$, we apply Lemma 16 to the Lyapunov function $Q_t$ which is nontrivial because of the extra terms related to the consensus error. The core steps for showing Theorem 4 are given as follows:

i) Based on standard analysis techniques for optimization algorithms, we first establish an approximate descent property for D-RR in Lemma 13. D-RR does not have exact descent at each epoch due to two types of errors: the consensus errors and the algorithmic errors.

ii) To derive convergence from the approximate descent property, we further provide upper bounds for these two types of errors in Lemma 14.

iii) By carefully checking the relationship between Lemmas 13 and 14, we construct a novel Lyapunov function $Q_t$ in Lemma 15.

iv) Finally, applying Lemma 16 to the recursion of $Q_t$ in Lemma 15 yields the complexity result of D-RR in Theorem 4.

III. Convergence Analysis: Strongly-Convex Case

In this section, we analyze D-RR for smooth and strongly convex objective functions and present the main convergence
result. We first derive Lemmas 5 and 7, which introduce coupled recursions for two decomposed expected error terms and serve as the cornerstones for the convergence analysis. A novel Lyapunov function $H_t^f$ is constructed in Lemma 8 to decouple these two recursions. In Section III-B, we bound the Lyapunov function $H_t^f$ and then obtain the recursion for the two decomposed errors. With all the preliminary results in hand, we are able to show the convergence results for the strongly convex case in Theorem 1 (decreasing stepsizes) and Theorem 2 (constant stepsize).

The following technical result is used repeatedly for unrolling the recursion when decreasing stepsizes are employed.

**Lemma 4:** For all $1 < a < k$, $a, k \in \mathbb{N}$, and $1 < \gamma \leq a/2$, we have

$$
\alpha^2 \gamma \leq \frac{k-1}{k} \left( 1 - \frac{\gamma}{t} \right) \leq \frac{\alpha^2 \gamma}{k}. 
$$

**Proof:** See Lemma 11 in [6].

A. Supporting Lemmas

The contents of this subsection correspond to Step (i) of Section II-B. In Lemma 5, we follow the intuition in Remark 1 and the arguments in [23] to construct a bound for $\mathbb{E}[\|x_t^2 - x_t^1\|^2]$. First, we define the shuffling variance $\sigma_{\text{shuffle}}^2$ in (11) as a distributed counterpart of corresponding variance for C-RR defined in [23].

**Definition 1 (Shuffling Variance):** Given a permutation $\pi$ of $[m]$, let $x_t^i$ be defined as in (7). The shuffling variance of agent $t$ is defined via

$$
\sigma_{\text{shuffle}}^2 := \max_{\ell = 0, \ldots, m-1} \mathbb{E}\left[ \frac{1}{n} \sum_{i=1}^{n} f_i(x_t^\pi_{\ell}(\bar{x}_t^\ell)) - \frac{1}{n} \sum_{i=1}^{n} f_i(x_t^\pi_{\ell}(\bar{x}_t^\ell)) \right].
$$

**Lemma 5:** Under Assumptions 1 and 2, let $\alpha_t \leq \frac{1}{2L}$. We have

$$
\mathbb{E}[\|x_t^{i+1} - x_t^i\|^2] \leq \left( 1 - \frac{\alpha_t \mu}{2} \right) \mathbb{E}[\|x_t^i - x_t^i\|^2] + 2\alpha_t \sigma_{\text{shuffle}}^2 + \frac{2\alpha_t \bar{L}^2}{n} \left( \frac{1}{\mu} + \alpha_t \right) \mathbb{E}[\|x_t^i - 1 (x_t^i)\|^2].
$$

**Proof:** See Appendix A-A.

Compared to [23, Theorem 1], one more term $\mathbb{E}[\|x_t^1 - 1 (x_t^1)\|^2]$ appears in Lemma 5 which is related to the expected consensus error of the decision variables among different agents. Therefore, if the consensus error decreases fast enough, we can expect that Algorithm 1 achieves a similar convergence rate compared to the centralized RR Algorithm 2. In fact, we will show in Lemma 7 that $\mathbb{E}[\|x_t^1 - 1 (x_t^1)\|^2]$ decreases in the order of $O(\alpha_t^2)$. Before we proceed to derive the recursion for $\mathbb{E}[\|x_t^1 - 1 (x_t^1)\|^2]$, we establish a relation between the shuffling variance $\sigma_{\text{shuffle}}^2$ and $\sigma_\varepsilon^2 := \frac{1}{mn} \sum_{i=1}^{m} \sum_{i=1}^{m} \|\nabla f_i(x^\ast)\|^2$ in Lemma 6, which shows that $\sigma_{\text{shuffle}}^2 \sim O(m\alpha_t^2 \sigma_\varepsilon^2)$. The term $\sigma_\varepsilon^2$ is similar to the variance of the gradient noises in SGD for solving finite sum problems. Hence, it makes sense that we use it as a baseline in our analysis.

**Lemma 6:** Under Assumption 2, we have

$$
\alpha_t \mu m \cdot \sigma_\varepsilon^2 \leq \sigma_{\text{shuffle}}^2 \leq \frac{\alpha_t^2 \bar{L} m}{4} \cdot \sigma_\varepsilon^2,
$$

where $\sigma_\varepsilon^2 := \frac{1}{mn} \sum_{i=1}^{m} \sum_{i=1}^{m} \|\nabla f_i(x^\ast)\|^2$.

**Proof:** See Appendix B-B in the Supplementary Material.

**Lemma 7:** Let Assumption 1 hold and assume

$$
\alpha_t \leq \sqrt{\frac{2 - \rho_w^2}{24\rho_w^2 (5 - \rho_w^2)}} \frac{1 - \rho_w^2}{L}
$$

for all $t$. Then, for all $\ell, t \geq 0$, we have

$$
\mathbb{E}[\|x_{t+1}^\ell - 1 (x_t^\ell)\|^2] \leq \left( 1 - \frac{\rho_w^2}{2} \right) \mathbb{E}[\|x_t^\ell - 1 (x_t^\ell)\|^2] + 30\alpha_t^2 nL^2 \left( \frac{1 - \rho_w^2}{1 - \rho_w^2} \right) \mathbb{E}[\|x_t^\ell - x_t^\ell\|^2] + 15\alpha_t^2 \frac{\rho_w^2 \alpha_t^2}{2} \left( 1 + \frac{2L\sigma_{\text{shuffle}}^2}{n} \right).
$$

**Proof:** See Appendix A-B.

If the error term $\mathbb{E}[\|x_t^1 - x_t^1\|^2]$ is assumed to be bounded, Lemma 7 implies that the expected consensus error $\mathbb{E}[\|x_t^1 - 1 (x_t^1)\|^2]$ decreases as fast as $O(\alpha_t^2)$ given a stepsize sequence $\{\alpha_t\}$. Combining this result with Lemma 5, we can then obtain convergence of D-RR with an overall complexity similar to C-RR. This observation corroborates our intuitive idea in Remark 1. Our remaining discussions will make this argument rigorous.

The main difficulty for formalizing our previous discussions about Lemmas 5 and 7 is that these two recursions are coupled. As a result, it is hard to unroll them directly to relate the errors corresponding to the inner loop and the outer loop of Algorithm 1. To handle this issue, we first define a Lyapunov function $H_t^f$ in (12) based on the decomposition (8):

$$
H_t^f := \mathbb{E}[\|x_t^1 - x_t^1\|^2] + \omega_t \mathbb{E}[\|x_t^1 - 1 (x_t^1)\|^2],
$$

where $\omega_t$ is specified in (13). Constructing an appropriate recursion for $H_t^f$ in Lemma 8 allows to finish Step (i). Note that the proof of Lemma 8 is similar to [12, Lemma 12].

**Lemma 8:** Under Assumption 1 and 2, let

$$
\omega_t := \frac{16\alpha_t \bar{L}^2}{n \mu (1 - \rho_w^2)}
$$

and suppose $\alpha_t$ satisfies

$$
\alpha_t \leq \min \left\{ \frac{\sqrt{2 - \rho_w^2}}{24 \rho_w^2 (5 - \rho_w^2)} \frac{1 - \rho_w^2}{L}, \frac{1 - \rho_w^2}{2 \mu}, \frac{1 - \rho_w^2}{8\sqrt{30}L^2} \right\}.
$$

We have the following relation between $H_t^f$ and $H_t := H_t^f$:

$$
H_t^f \leq \left( 1 - \frac{\alpha_t \mu}{4} \right)^t H_t^0 + 2 \left( \frac{\alpha_t \sigma_{\text{shuffle}}^2}{4} + \frac{\rho_w^2 \alpha_t^2}{4} \frac{L^2}{\mu (1 - \rho_w^2)} + 120 \alpha_t^2 \rho_w^2 L^2 \sigma_{\text{shuffle}}^2 \right) \sum_{k=0}^{t-1} \left( 1 - \frac{\alpha_t \mu}{4} \right)^k.
$$

(15)
In addition, 
\[ H_{t+1} \leq \left( 1 - \frac{\alpha_t \mu}{m} \right) H_t + 2 \left[ \alpha_t \sigma^2_{\text{stoch}} \left( 1 + \frac{240 \rho^2 L^3}{\mu (1 - \rho^2_w)^2} \right) \right] + \frac{120 \rho^2 L^2}{\mu (1 - \rho^2_w)^2} \sum_{k=0}^{m-1} \left( 1 - \frac{\alpha_t \mu}{4} \right)^k , \tag{16} \]
where \( H_{t+1} \) is defined as
\[ H_{t+1} := \mathbb{E}\left[ \left\| \tilde{x}^m_t - x^* \right\|^2 \right] + \omega_t \mathbb{E}\left[ \left\| x_t^m - 1 (\tilde{x}^m_t) \right\|^2 \right] \]
\[ = \mathbb{E}\left[ \left\| \tilde{x}_{t+1} - x^* \right\|^2 \right] + \omega_t \mathbb{E}\left[ \left\| x_{t+1} - 1 \tilde{x}^T_{t+1} \right\|^2 \right] . \]

**Proof:** See Appendix A-C. \( \square \)

Lemma 8 plays a key role in decoupling the recursions in Lemmas 5 and 7. It can also be used to bound \( \mathbb{E}\left[ \left\| \tilde{x}_t^\ell - \tilde{x}^\ell_t \right\|^2 \right] \) from the definition of \( H^\ell_t \) in (12).

### B. Preliminary Results

In this section, we consider two specific stepsize choices to finish Step (ii)-(iii): a decreasing stepsize sequence and a constant stepsize. Specifically, the decreasing stepsize is given by
\[ \alpha_t = \frac{\theta}{m \mu (t + K)} , \quad \forall t > 0 , \tag{17} \]
for some \( \theta, K > 0 \).

**Remark 2:** The chosen stepsize policy (17) is common in centralized RR algorithms when the objective function is smooth and strongly convex; see for example, [22].

Note that relation (16) provides a recursion with respect to the epoch-wise error for \( H_t \). We unroll the inequality in light of Lemma 4 to obtain Lemma 9.

**Lemma 9:** Under Assumption 1 and 2, let \( K \) be chosen such that \( K \geq \frac{\theta}{2m} \) and \( \alpha_t = \frac{\theta}{m \mu (t + K)} \) satisfies (14) for all \( t \geq 0 \), and \( \theta > 12 \). Then we have
\[ H_t \leq \left( \frac{K}{t+K} \right)^{\frac{9}{2}} H_0 + \left( \frac{240 \rho^2 L^2}{\mu (1 - \rho^2_w)^2} \right) + \frac{80 \rho^2}{m^2 \mu^{3} \left( \theta - 8 \right) (t + K)^2} . \]

In addition, under the constant stepsize \( \alpha_t = \alpha \) that satisfies (14), we have
\[ H_t \leq \left( 1 - \frac{\alpha \mu}{4} \right)^{mt} H_0 + \frac{4 \alpha^2}{\mu} \left( m \mu + \frac{240 \rho^2 L^2}{\mu (1 - \rho^2_w)^2} \right) . \]

**Proof:** See Appendix B-C in the Supplementary Material. \( \square \)

We also obtain the bound for \( H^\ell_t \) according to Lemmas 8 and 9 by repeating the procedures in the proof of Lemma 9.

**Lemma 10:** Let the conditions in Lemma 9 hold. Under the decreasing stepsize policy, we have
\[ H^\ell_t \leq \left( \frac{K}{t+K} \right)^{\frac{9}{2}} H_0 + \left( \frac{240 \rho^2 L^2}{\mu (1 - \rho^2_w)^2} \right) \left( \frac{1}{(t + K)^2} \right) , \quad \forall t, \ell . \]

Under the constant stepsize, we obtain
\[ H^\ell_t \leq \left( 1 - \frac{\alpha \mu}{4} \right)^{mt} H_0 + \frac{8 \alpha^2}{\mu} \left( m \mu + \frac{240 \rho^2 L^2}{\mu (1 - \rho^2_w)^2} \right) . \]

**Proof:** See Appendix B-D in the Supplementary Material. \( \square \)

From the definition of \( H^\ell_t \), we have \( \mathbb{E}\left[ \left\| \tilde{x}_t^\ell - \tilde{x}^\ell_t \right\|^2 \right] \leq H^\ell_t \). Noticing that the right-hand sides of the inequalities in Lemma 10 do not involve \( \ell \), we can directly substitute the above bound into Lemma 7 and unroll the recursion with respect to \( \ell \). Then, we can obtain a decoupled recursion for the consensus error in the following lemma.

**Lemma 11:** Let conditions in Lemma 9 hold and define
\[ \hat{X}_0 := H_0 + \frac{28 (m \mu + L) \rho^2}{\mu L^2} , \]
\[ \hat{X}_1 := \frac{30 \rho^2 L^2}{1 - \rho^2_w} \hat{X}_0 + \frac{15 \rho^2}{1 - \rho^2_w} \sigma^2 + \frac{m \mu (1 - \rho^2_w)}{8 L} . \]
Under both constant and decreasing stepsizes, we have for all \( t, \ell \) that
\[ \mathbb{E}\left[ \left\| \tilde{x}_t^\ell - 1 (\tilde{x}_t^\ell) \right\|^2 \right] \leq \left( 1 + \frac{\rho^2_w}{2} \right)^{\ell+1} \mathbb{E}\left[ \left\| \tilde{x}_t^0 - 1 (\tilde{x}_t^0) \right\|^2 \right] + \frac{\alpha^2 \hat{X}_1}{1 - \rho^2_w} . \]
Moreover,
\[ \mathbb{E}\left[ \left\| \tilde{x}_t^0 - 1 (\tilde{x}_t^0) \right\|^2 \right] \leq \left( 1 + \frac{\rho^2_w}{2} \right)^{mt} \mathbb{E}\left[ \left\| \tilde{x}_0^0 - 1 (\tilde{x}_0^0) \right\|^2 \right] + \frac{4 \hat{X}_1}{(1 - \rho^2_w)^2} \alpha^2 . \]

**Proof:** See Appendix A-D. \( \square \)

Lemma 11 verifies our previous discussion that the consensus error \( \mathbb{E}\left[ \left\| \tilde{x}_t^0 - 1 (\tilde{x}_t^0) \right\|^2 \right] \) decreases as fast as \( O(\alpha^2) \). Combining Lemma 11 and Lemma 5, we can also derive a decoupled and refined bound for the optimization error \( \mathbb{E}\left[ \left\| \tilde{x}_t^0 - x^* \right\|^2 \right] \) in Lemma 12.

**Lemma 12:** Let the conditions in Lemma 9 hold. Under the decreasing stepsize policy, we have
\[ \mathbb{E}\left[ \left\| \tilde{x}_t^0 - x^* \right\|^2 \right] \leq \left( \frac{K}{t+K} \right)^{\frac{9}{2}} \left\| \tilde{x}_0^0 - x^* \right\|^2 + \frac{2 \theta^2 L \sigma^2}{m \mu^2 (\theta - 4) (t + K)^2} \mathbb{E}\left[ \left\| \tilde{x}_0^0 - 1 (\tilde{x}_0^0) \right\|^2 \right] \]
\[ + \frac{9 \theta}{m \mu^2 (1 - \rho^2_w)} \left\| \tilde{x}_0^0 - 1 (\tilde{x}_0^0) \right\|^2 , \quad \forall t \geq 0 , \ell . \]

**Proof:** See Appendix B-E in the Supplementary Material. \( \square \)

### C. Main Results: Strongly-Convex Case

We now present the main convergence results of D-RR for the strongly convex case under both decreasing stepsizes (17) in Theorem 1 and a constant stepsize \( \alpha \) in Theorem 2 by combining Lemmas 11 and 12.
Under decreasing stepsizes, we have the following theorem.

**Theorem 1:** Under Assumptions 1 and 2, let \( \alpha_t = \frac{\theta}{m\mu(t+K)} \) with \( \theta > 12 \) and \( K \) be chosen as

\[
K \geq \max \left\{ \frac{\theta}{2}, \frac{24\rho_w^2(5-\rho_w^2)\theta^2}{(2-\rho_w^2)(1-\rho_w^2)m^2\mu^2}, \frac{8\sqrt{3}L^2\theta}{(1-\rho_w^2)m\mu^2} \right\}.
\]

Then for Algorithm 1, we have

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \|x_{i,t}^0 - x^*\|^2 \right] \leq \left( \frac{K}{t+K} \right)^{\frac{\theta}{2}} \|x_0^0 - x^*\|^2
\]

\[
\left( C_1 + \left( \frac{1 + \rho_w^2}{2} \right)^{mt} \right) \frac{\|x_0^0 - (x_0^0)^T\|^2}{n} + \frac{C_2}{(t+K)^2}.
\]

where

\[
C_1 := \left( \frac{K}{t+K} \right)^{\frac{\theta}{2}} \frac{96L^2}{\mu^3(1-\rho_w^2)},
\]

\[
C_2 := \frac{2\theta^2\rho_w^2 L^2}{m\mu^3(\theta - 4)} + \frac{96\theta^3 L^2 \hat{X}_1}{nm^2\mu^5(1-\rho_w^2)\theta - 4}
\]

\[
+ \frac{4\theta^2}{np^2 m^2(1-\rho_w^2)^2}.
\]

and \( \hat{X}_1 \) is defined in Lemma 11.

**Proof:** Combining Lemmas 11 and 12 leads to the result. \( \square \)

**Remark 3:** From Theorem 1, D-RR enjoys the \( O(1/mT^2) \) rate of convergence under a decreasing stepsize policy. Compared with DSGD whose convergence rate is \( O(1/\min T) \), D-RR is more favorable when \( T \) is relatively large compared to the number of agents \( n \).

The convergence result of D-RR under the constant stepsize \( \alpha \) is stated in the next theorem.

**Theorem 2:** Under Assumptions 1 and 2, let \( \alpha_t = \alpha \) satisfy (14) and \( \hat{X}_1 \) be defined as in Lemma 11. We have for Algorithm 1 that

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \|x_{i,t}^0 - x^*\|^2 \right] \leq \left( 1 - \frac{\alpha\mu}{4} \right)^{mt} H_0
\]

\[
+ \frac{4\alpha^2}{\mu} \left( mL + \frac{240\rho_w^2 L^2}{\mu(1-\rho_w^2)^2} \right) \sigma^2
\]

\[
+ \frac{1 + \rho_w^2}{2} \left( \frac{mL}{(1-\rho_w^2)^2} \right)^{mt} \frac{\|x_0^0 - (x_0^0)^T\|^2}{n} + \frac{4\hat{X}_1}{n(1-\rho_w^2)^2} \alpha^2.
\]

**Proof:** Let \( \alpha_t = \alpha \) in (43), we obtain

\[
\sum_{k=0}^{t-1} \left( \frac{1 + \rho_w^2}{2} \right)^{m(t-k)} \alpha^2 \leq \frac{2\alpha^2}{1 - \rho_w^2}.
\]

Then, applying Lemma 11, it follows

\[
\mathbb{E} \left[ \|x_t^0 - (x_t^0)^T\|^2 \right] \leq \left( \frac{1 + \rho_w^2}{2} \right)^{mt} \|x_0^0 - (x_0^0)^T\|^2 + \frac{4\hat{X}_1}{(1-\rho_w^2)^2} \alpha^2.
\]

According to (9) and (12), we have

\[
\mathbb{E} \left[ \|x_t^0 - x^*\|^2 \right] \leq \left( 1 - \frac{\alpha\mu}{4} \right)^{mt} H_0
\]

\[
+ \frac{4\alpha^2}{\mu} \left( mL + \frac{240\rho_w^2 L^2}{\mu(1-\rho_w^2)^2} \right) \sigma^2.
\]

Combining (19) and (20) finishes the proof. \( \square \)

**Remark 4:** In light of Theorem 2, under a constant stepsize policy, the expected error of D-RR decreases exponentially fast to a neighborhood of 0 with size being of order \( O(\alpha n) \). By comparison, the expected error of DSGD decreases to a neighborhood of 0 with size being of order \( O(\alpha/n) \). Therefore, if \( \alpha \) is relatively small, e.g., when higher accuracy is desirable, then D-RR is more favorable than DSGD.

To better compare the convergence results of D-RR with those of C-RR, we present the convergence result of C-RR (Algorithm 2) presented in [23] in Theorem 3 which considers a constant stepsize depending on the number of epochs \( T \).

**Theorem 3 ([23, Corollary 1]):** Consider Algorithm 2, and let Assumption 2 hold and the stepsize be chosen as

\[
\alpha \leq \min \left\{ \frac{1}{L}, \frac{1}{\mu mT} \log \left( \frac{x_0 - x^*}{\mu^2 mT^2} \right) \frac{\kappa \sigma^2}{\kappa \sigma^2} \right\}, \text{ where } \kappa = \frac{L}{\mu},
\]

then the final iterate of Algorithm 2 \( x_T \) satisfies

\[
\mathbb{E} \left[ ||x_T - x^*||^2 \right] \leq \exp \left( -\alpha \mu mT \right) ||x_0 - x^*||^2 + \tilde{O} \left( \frac{\kappa \sigma^2}{\mu^2 mT^2} \right).
\]

In Corollary 1, we state convergence of D-RR under a similar setting as in Theorem 3. It can be seen that the first two terms in (22) are comparable with those in (21). In particular, if we consider a complete graph where \( 1 - \rho_w^2 = 1 \) and use the same initialization for all the agents, then the convergence result of D-RR reduces to that given in Theorem 3.

**Corollary 1:** Let the conditions in Theorem 2 hold. Furthermore, let the stepsize \( \alpha_t = \alpha \) satisfy

\[
\alpha \leq \frac{4}{L mT} \log \left( \frac{H_0 \mu^2 mT^2}{\kappa \sigma^2} \right), \text{ where } \kappa = \frac{L}{\mu}.
\]
Then, the final iterate $x^0_{i,T}$, for all $i \in [n]$, satisfies
\begin{align*}
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| x^0_{i,T} - x^* \right\|^2 \right] &= \exp \left( -\frac{\alpha \mu m T}{4} \right) \| H_0 \|
+ \hat{O} \left( \frac{\kappa \sigma^2}{\mu^2 mT^2} \right) + \hat{O} \left( n \left\| x^0_{i,T} - x^* \right\|^2 + \left\| x^0 - 1(x^0)_T \right\|^2 \right) \left( 1 - \frac{\sigma^2}{(1 - \rho_w^2)\mu^2 mT^2} \right) \\
&\quad + \frac{\alpha \mu m T}{4} \left[ \left\| x^0_{i,T} - x^* \right\|^2 + \hat{O} \left( \frac{\kappa \sigma^2}{\mu^2 mT^2} \right) \right].
\end{align*}

In addition, if the network topology is a complete graph and we initialize $x^0_{i,0} = x_0$, for all $i \in [n]$, with the further assumption $\| x_0 - x^* \|^2 = O(m)$, then it holds that
\begin{align*}
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \left\| x^0_{i,T} - x^* \right\|^2 \right] &= \exp \left( -\frac{\alpha \mu m T}{4} \right) \| x_0 - x^* \|^2 \\
&\quad + \hat{O} \left( \frac{\kappa \sigma^2}{\mu^2 mT^2} \right).
\end{align*}

Remark 5: The assumption $\| x_0 - x^* \|^2 = O(m)$ is not restrictive. In fact, $\| x_0 - x^* \|^2$ is usually far less than $O(m)$.

Proof: See Appendix B-F in the Supplementary Material.

IV. CONVERGENCE ANALYSIS: NONCONVEX CASE

In this section, we consider the case where the objective functions are smooth nonconvex. Under Assumption 3, we derive a convergence rate result of D-RR, which is comparable to that of centralized RR. Note that Assumption 3 is standard for studying distributed nonconvex optimization algorithms; see, e.g., [51].

We first provide Lemma 13, which states an approximate descent property of D-RR under the general smooth nonconvex setting. The error terms in this approximate descent property consist of the consensus errors and the algorithmic errors. In order to establish iteration complexity of D-RR, we have to further bound these two types of errors. In Lemma 14, we present an upper bound for the consensus errors in terms of the graph structure $\rho_0$ and the stepsize. The algorithmic errors can be easily bounded under Assumption 3. Finally, by invoking the bounds for these two types of errors in Lemma 13, we can derive the convergence result for D-RR; see Theorem 4.

A. Supporting Lemmas

Lemma 13: Let Assumptions 1 and 3 be valid. Suppose further $\alpha_i = \alpha \leq 1/mL$. Then, the following holds for all $t \geq 1$:
\begin{align*}
f(x^0_{t+1}) &\leq f(x^0_t) - \frac{\alpha m}{2} \left\| \nabla f(x^0_t) \right\|^2 \\
&\quad + \frac{L^2 \alpha}{n} \sum_{t=0}^{m-1} \left\| x^t - 1(x^t)_T \right\|^2 + \alpha L^2 \sum_{t=0}^{m-1} \left\| x^t - x^0 \right\|^2.
\end{align*}

Proof: See Appendix C-B in the Supplementary Material.

Lemma 13 is an approximate descent property for D-RR. Next, we estimate the consensus error $\sum_{t=0}^{m-1} \left\| x^t - (1(x^t)_T \right\|^2$ and the algorithmic error $\sum_{t=0}^{m-1} \left\| x^t - x^0 \right\|^2$. In Lemma 14, we bound the last two terms in the inequality of Lemma 13.

Lemma 14: Suppose Assumptions 1 and 3 are valid. Let the stepsize $\alpha_t = \alpha$ satisfy
\begin{align*}
\alpha &\leq \min \left\{ \frac{1}{2 \sqrt{6mL}}, \frac{1 - \rho_0^2}{4 \sqrt{6L}} \right\}.
\end{align*}

Then, the following holds for all $t \geq 1$:
\begin{align*}
L_t &\leq \frac{4}{n(1 - \rho_0^2)} \left\| x^0_t - (1(x^0)_T \right\|^2 \\
&\quad + 6 m^2 \alpha^2 B^2 (m + 4) + 6 \alpha^2 m (m + 2) \left\| \nabla f(x^0_t) \right\|^2 \\
&\quad + 12 m^2 \alpha^2 A (m + 1) \left( f(x^0_t) - \bar{f} \right),
\end{align*}
where $L_t := \frac{1}{n} \sum_{t=0}^{m-1} \left\| x^t - 1(x^t)_T \right\|^2 + \sum_{t=0}^{m-1} \left\| x^t - x^0 \right\|^2$.

In addition, we have
\begin{align*}
\left\| x^0_{t+1} - 1(x^0_{t+1})_T \right\|^2 &\leq \left( \frac{1 + \rho_0^2}{2} \right) \left\| x^0_t - 1(x^0)_T \right\|^2 \\
&\quad + \frac{12 \alpha^2 n L^2 L_t}{1 - \rho_0^2} + \frac{6 \alpha^2 mnL^2 B^2 (m + 4) + 6 \alpha^2 mnL^2 (1 - \rho_0^2)}{1 - \rho_0^2} \left\| \nabla f(x^0_t) \right\|^2 \\
&\quad + \frac{12 \alpha^2 mnL^2 (1 - \rho_0^2)}{1 - \rho_0^2} \left( f(x^0_t) - \bar{f} \right).
\end{align*}

Proof: See Appendix C-C in the Supplementary Material.

The extra term $\left\| x^0_t - 1(x^0)_T \right\|^2$ in $L_t$ inspires us to consider the Lyapunov function $Q_t$ in (24):
\begin{align*}
Q_t := f(x^0_t) - \bar{f} + \frac{16 \alpha L^2}{n(1 - \rho_0^2)^2} \left\| x^0_t - 1(x^0)_T \right\|^2.
\end{align*}

Lemma 15: Suppose Assumptions 1 and 3 are valid. Let the stepsize $\alpha_t = \alpha$ satisfy
\begin{align*}
\alpha &\leq \min \left\{ \frac{1 - \rho_0^2}{4 \sqrt{3L(m + 1)}}, \frac{1 - \rho_0^2}{16 \sqrt{6L}} \right\}.
\end{align*}

Then, we have
\begin{align*}
Q_{t+1} &\leq \left[ 1 + \frac{12 m^2 \alpha^2 L^2 A (m + 1)}{(1 - \rho_0^2)^2} + \frac{384 \alpha \alpha^3 L^2 m}{1 - \rho_0^2} \right] Q_t \\
&\quad - \frac{ma}{4} \left\| \nabla f(x^0_t) \right\|^2 + \frac{6 \alpha \alpha^2 L^2 B^2 (m + 4) + 32}{1 - \rho_0^2}.
\end{align*}

Proof: See Appendix C-D in the Supplementary Material.

Lemma 16 in [23, Lemma 6] provides a direct link connecting Lemma 15 to Theorem 4 in the next subsection.

Lemma 16 ([23, Lemma 6]): Suppose that there exist constants $a, b, c \geq 0$ and nonnegative sequences $(s_t)^T_{t=0}, (q_t)^T_{t=0}$ such that for any $t$ satisfying $0 \leq t \leq T$, we have the recursion
\begin{align*}
s_{t+1} &\leq (1 + a)s_t - bq_t + c.
\end{align*}

Then, the following holds:
\begin{align*}
\min_{t=0, \ldots, T-1} q_t &\leq \frac{(1 + a)^T}{bT} s_0 + \frac{c}{b}.
\end{align*}
B. Main Results: Nonconvex Case

Equipped with Lemmas 15 and 16, we are ready to derive the convergence result of D-RR for smooth nonconvex optimization problems over the networks. The main result is given in the following theorem.

**Theorem 4:** Suppose Assumptions 1 and 3 are valid. Suppose further that \( \alpha = \frac{\eta}{m^2} \) with \( \gamma \in (0, 1) \), where \( \eta > 0 \) is some constant such that
\[
0 < \alpha \leq \min \left\{ \frac{1 - \rho_w^2}{4\sqrt{3}L(m + 2)}, \frac{(1 - \rho_w^2)^{3/2}}{16\sqrt{6}L}, \left( \frac{12 m^2 L^2 A(4 + m)}{1 - \rho_w^2} + \frac{384 L^2 A m}{(1 - \rho_w^2)^3} \right)^{-1/3} \right\}
\]
where \( T \) denotes the total number of iterations. Then, we have
\[
\min_{t=0,\ldots,T-1} \| \nabla f(\tilde{x}^0_t) \|^2 
\leq \frac{12}{\eta T^{1-\gamma}} \left( (f(\tilde{x}^0_0) - \bar{f}) + \frac{16\alpha L^2}{m(1 - \rho_w^2)} \| x^0_0 - 1(x^0_0)^T \|^2 \right) + \frac{18\alpha^2 B^2 L^2 [m(m + 4) + 32]}{m^2(1 - \rho_w^2)^{3/2} T^{2-\gamma}}.
\]
Consequently, the optimal rate is attained when \( \gamma = 1/3 \) and it follows that
\[
\min_{t=0,\ldots,T-1} \| \nabla f(\tilde{x}^0_t) \|^2 
\leq \frac{12}{\eta T^{1/3}} \left( (f(\tilde{x}^0_0) - \bar{f}) + \frac{16\alpha L^2}{m(1 - \rho_w^2)} \| x^0_0 - 1(x^0_0)^T \|^2 \right) + \frac{18\alpha^2 B^2 L^2 [m(m + 4) + 32]}{m^2(1 - \rho_w^2)^{3/2} T^{2/3}}.
\]

**Proof:** Set
\[
b := \frac{m\alpha}{4},
\]
\[
a := \frac{12 m^2 \alpha^3 L^2 A(4 + m)}{1 - \rho_w^2} + \frac{384 \alpha^3 L^2 A m}{(1 - \rho_w^2)^3},
\]
\[
c := \frac{6\alpha^3 B^2 L^2 [m(m + 4) + 32]}{(1 - \rho_w^2)^3},
\]
\[
st := (f(\tilde{x}^0_0) - \bar{f}) + \frac{16\alpha L^2}{m(1 - \rho_w^2)^2} \| x^0_0 - 1(x^0_0)^T \|^2
\]
\[
q_t := \| \nabla f(\tilde{x}^t) \|^2.
\]
Invoking Lemma 16, using the inequality \((1 + a)T \leq \exp(aT) \leq 3\) for all \( \alpha \) satisfying
\[
0 < \alpha \leq \left( \frac{12 m^2 L^2 A(4 + m)}{1 - \rho_w^2} + \frac{384 L^2 A m}{(1 - \rho_w^2)^3} \right)^{-1/3} \frac{1}{T^{1/3}},
\]
it follows that
\[
\min_{t=0,\ldots,T-1} \| \nabla f(\tilde{x}^0_t) \|^2 
\leq \frac{12}{mA T} \left( (f(\tilde{x}^0_0) - \bar{f}) + \frac{16\alpha L^2}{m(1 - \rho_w^2)^2} \| x^0_0 - 1(x^0_0)^T \|^2 \right) + \frac{18\alpha^2 B^2 L^2 [m(m + 4) + 32]}{m^2(1 - \rho_w^2)^3 T^{2-\gamma}}.
\]

Substituting \( \alpha = \frac{\eta}{m^2} \), we have
\[
\min_{t=0,\ldots,T-1} \| \nabla f(\tilde{x}^0_t) \|^2 
\leq \frac{12}{\eta T^{1-\gamma}} \left( (f(\tilde{x}^0_0) - \bar{f}) + \frac{16\alpha L^2}{m(1 - \rho_w^2)} \| x^0_0 - 1(x^0_0)^T \|^2 \right) + \frac{18\alpha^2 B^2 L^2 [m(m + 4) + 32]}{m^2(1 - \rho_w^2)^{3/2} T^{2-\gamma}}.
\]
This completes the proof. \(\Box\)

**Remark 6:** It can be seen from Theorem 4 that D-RR enjoys the \( O(1/T^{2/3}) \) rate of convergence for solving smooth nonconvex problems. Similar to the strongly convex case, noticing that DSGD with constant stepsize has \( O(1/(mn)^{1/2}T^{1/2}) \) rate of convergence [9], D-RR outperforms DSGD if \( T \) is relatively large (related to the sample size \( m \) in each agent and the number of agents \( n \)).

V. EXPERIMENTAL RESULTS

In this section, we provide two numerical examples which illustrate the performance of D-RR. For both the strongly convex problem (25) and the nonconvex problem (26), we show the proposed D-RR algorithm outperforms SGD and DSGD when \( T \) is large enough. All the results in the following experiments are averaged over ten repeated runs if not otherwise specified. We consider the heterogeneous data setting for all the experiments, where the data samples are first sorted according to their labels and are then partitioned among the agents. Some codes are from [52].

A. Logistic Regression

We consider a binary classification problem using logistic regression (25) and the MNIST dataset [53]. Each agent possesses a distinct local dataset \( \mathcal{S}_i \) selected from the whole dataset \( \mathcal{S} \). The classifier can then be obtained by solving the following optimization problem using all the agents’ local datasets \( \mathcal{S}_i, i = 1, 2, \ldots, n \):
\[
\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x), \quad (25a)
\]
\[
f_i(x) := \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \log [1 + \exp(-x^T u_j v_j)] + \frac{\rho}{2} \| x \|^2, \quad (25b)
\]
where \( \rho \) is set as \( \rho = 0.2 \).

We compare D-RR (Algorithm 1) with SGD, DGD, DPGRR [47], and centralized RR (Algorithm 2) for classifying handwritten digits 2 and 6 on the MNIST dataset over a grid graph, an exponential graph, and an Erdős-Rényi graph (all with \( n = 16 \), see Fig. 1) respectively. We consider both constant stepizes (Fig. 2) and decreasing stepsizes (Fig. 3) for all the methods. The methods use the same initialization in each figure.

For both constant (Fig. 2) and decreasing (Fig. 3) stepsizes, the errors decay at the same rate for all the algorithms during
Fig. 1. Illustration of three graph topologies. The spectral gaps for the three graphs increase from left to right. Each node in the exponential graph is connected to its $2^0, 2^1, \ldots$ neighbors, and the probability for edge creation in the Erdős-Rényi graph is set as 0.8.

Fig. 2. Comparison among D-RR, DSGD, SGD, DPG-RR, and centralized RR for solving Problem (25) on the MNIST dataset using constant stepsize. The stepsize is set as $\frac{1}{8000}$ for all the methods.

Fig. 3. Comparison among D-RR, DSGD, SGD, DPG-RR, and centralized RR for solving Problem (25) on the MNIST dataset using decreasing stepsizes. The stepsize is set as $\alpha_t = \frac{1}{50t + 400}$ for all the methods.

the starting epochs. After the starting epochs, DSGD and SGD achieve less accuracy compared to D-RR and C-RR. Comparing the two random reshuffling methods, the performance of D-RR is worse than C-RR since the convergence result of D-RR is affected by the connectivity of the graph topology. When the network topology becomes better-connected (from left to right), the performance of D-RR tends to be more comparable to that of C-RR. Regarding the stepsize policy, decreasing stepsizes are more favorable which allows larger stepsizes at the starting epochs.

We also compare the performance of D-RR and DPG-RR with respect to the number of communication rounds for each node in Fig. 4. Both methods utilize the same stepsize and the underlying graph is a grid graph with $n = 16$. Note that D-RR conducts one round of communication per gradient computation while DPG-RR only communicates epoch-wisely. However, as
can be seen from Fig. 4, although DPG-RR saves communication cost and proceeds faster at the beginning, the error can not be controlled as well as in D-RR; see Remark 1 for discussion.

### B. Nonconvex Logistic Regression

Nonconvex regularizers are also widely used in statistical learning such as approximating sparsity. In this part, we consider a nonconvex binary classification problem (\(26\)) classifying airplanes and trucks in CIFAR-10 [54] dataset and compare the proposed D-RR method (Algorithm 1) with DSGD, SGD, and centralized RR (Algorithm 2) over a grid graph, an exponential graph, and an Erdős–Rényi graph, respectively. The optimization problem is

\[
\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x),
\]

\[
f_i(x) := \frac{1}{|S_i|} \sum_{j \in S_i} \log \left[ 1 + \exp(-x^T u_j v_j) \right] + \frac{\eta}{2} \sum_{q=1}^{p} x_q^2 \left( 1 + \frac{1}{x_q^2} \right).
\]

Here, \(x_q\) denotes the \(q\)-th element of \(x \in \mathbb{R}^p\). We choose \(\eta = 0.2\) and use constant stepsize for all the epochs. All the methods use the same initialization with the same stepsize.

From Fig. 5, we also observe that the performance of the two random reshuffling methods outperform DSGD and SGD and achieve higher accuracy after the starting epochs. By comparing the performance from Fig. 5(a)-(c) where the spectral gap increases, we can infer that D-RR performs better when the spectral gap becomes larger (i.e., the graph connectivity becomes better).

**Remark 7:** Note that for both problems above, the performance gap between D-RR and DSGD becomes obvious only when the optimization errors are small, which may lead to similar testing performance for the two algorithms. Thus it is of future interest to further explore the conditions under which D-RR outperforms DSGD in the testing accuracy, especially for training large-scale machine learning models.

### VI. Conclusion

This paper is concerned with solving the distributed optimization problem over networked agents. Inspired by the classical distributed gradient descent (DGD) method and Random Reshuffling (RR), we propose a distributed random reshuffling (D-RR) algorithm and show the convergence results of D-RR match those of centralized RR (up to constant factors) for both smooth strongly convex and smooth nonconvex objective functions.

### APPENDIX A

### APPENDIX PARTS OF PROOFS FOR THE STRONGLY-CONVEX CASE

**A. Proof of Lemma 5**

Proof: As discussed in Remark 1, \(\frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_{i,t}^f)\) is an approximation of \(\frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t(x_{i,t}^f)\), hence, the core difference between our analysis and the one in [23] mainly lies in this approximation. According to (7), we have

\[
\tilde{x}_{i+1}^f = x_i^* - \frac{\alpha_t}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_i^*).
\]

This yields

\[
\mathbb{E} \left[ \left\| \tilde{x}_{i+1}^f - x_{i+1}^f \right\|^2 \right] = \mathbb{E} \left[ \left\| \tilde{x}_i^f - \tilde{x}_i^f - \left( \frac{\alpha_t}{n} \sum_{i=1}^{n} \left[ \nabla f_i,\pi_t^i(x_{i,t}^f) - \nabla f_i,\pi_t^i(x_i^*) \right] \right) \right\|^2 \right]
\]

\[
= \mathbb{E} \left[ \left\| \tilde{x}_i^f - \tilde{x}_i^f \right\|^2 + \alpha_t^2 \left\| \sum_{i=1}^{n} \left[ \nabla f_i,\pi_t^i(x_{i,t}^f) - \nabla f_i,\pi_t^i(x_i^*) \right] \right\|^2 \right] - 2\alpha_t \left( \tilde{x}_i^f - \tilde{x}_i^f, \frac{1}{n} \sum_{i=1}^{n} \left[ \nabla f_i,\pi_t^i(x_{i,t}^f) - \nabla f_i,\pi_t^i(x_i^*) \right] \right).
\]

We now divide the the inner product in (27) into two parts:

\[
\left( \tilde{x}_i^f - \tilde{x}_i^f, \frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_{i,t}^f) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_i^*) \right) = \left( \tilde{x}_i^f - \tilde{x}_i^f, \frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_{i,t}^f) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_i^*) \right)
\]

\[
+ \left( \tilde{x}_i^f - \tilde{x}_i^f, \frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_{i,t}^f) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_i,\pi_t^i(x_i^*) \right).
\]

Introducing \(\bar{s}_t := \frac{1}{n} \sum_{i=1}^{n} f_i,\pi_t\) and recalling \(D_{\bar{s}_t}(y, x) = \bar{s}_t(y) - \bar{s}_t(x) - \langle \nabla \bar{s}_t(x), y - x \rangle\), we have

\[
A = D_{\bar{s}_t}(\tilde{x}_{i,t}^f, x_t^f) + D_{\bar{s}_t}(\tilde{x}_{i,t}^f, x_t^*) - D_{\bar{s}_t}(\tilde{x}_{i,t}^f, x_t^f).
\]

According to Assumption 2, \(\bar{s}_t = \frac{1}{n} \sum_{i=1}^{n} f_i,\pi_t\) is also \(\mu\)-strongly convex and \(L\)-smooth. Thus, applying (3) and (6),
we obtain
\[
\frac{1}{L} \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \nabla f_{i,i} (x^*) - f_{i,i} (\bar{x}_i) \right) \right\|^2 \leq D_{\delta_i} (\bar{x}_i, x^*),
\]
(28)

The last term in A can be bounded by shuffling variance \(\sigma^2_{\text{shuffle}}\) introduced in Definition 1. Note that the definition of \(\sigma^2_{\text{shuffle}}\) is different from that in [23] and it does not include the factor \(1/\alpha_i\) (since we use decreasing stepsizes). We have
\[
\mathbb{E}[D_{\delta_i} (\bar{x}_i, x^*)] \leq \sigma^2_{\text{shuffle}}.
\]
(30)

For B, we apply Cauchy’s inequality, Young’s inequality and invoke (5),
\[
|B| \leq \frac{c}{2} \left\| \ddot{x}_i - \dddot{x}_i \right\|^2 + \frac{L^2}{2nc} \sum_{i=1}^{n} \left\| x_{i,t} - \ddot{x}_i \right\|^2 \forall c > 0.
\]
(31)

Next, we bound the gradient term in (27).
\[
\frac{1}{2} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,i} (x_{i,t}) - \nabla f_{i,i} (x^*) \right\|^2 \\
\leq \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,i} (x_{i,t}) - \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,i} (x^*) \right\|^2 \\
+ \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,i} (x^*) \right\|^2 \\
\leq L^2 \left\| x_{i,t} - \ddot{x}_i \right\|^2 + \left\| \frac{1}{n} \sum_{i=1}^{n} \left( \nabla f_{i,i} (\ddot{x}_i) - \nabla f_{i,i} (x^*) \right) \right\|^2.
\]

The second term would get absorbed combining (29) when the stepsize is small, i.e., \(\alpha_i \leq 1/(2L)\). Finally, choosing \(c = \mu/2\) in (31), we obtain the result.

B. Proof of Lemma 7

Proof: Lemma 17 first bounds a specific term in our derivation for Lemma 7.

Lemma 17: We have
\[
\mathbb{E}[\| \nabla F_{\pi_i} (x_i^*) \|^2] \leq 6 L^2 n \mathbb{E}[\| x_i^* - \dddot{x}_i \|^2] \\
+ 6 L^2 \mathbb{E}[\| x_i^* - (\dddot{x}_i)^\top \|^2] + 3n \sigma^2 + 6nL \sigma^2_{\text{shuffle}}.
\]

Proof: It holds that
\[
\mathbb{E}[\| \nabla F_{\pi_i} (x_i^*) \|^2] \leq 3 \mathbb{E}[\| \nabla F_{\pi_i} (x_i^*) - \nabla F_{\pi_i} (1 (x^*)^\top) \|^2] \\
+ 3 \mathbb{E}[\| \nabla F_{\pi_i} (1 (x^*)^\top) - \nabla F_{\pi_i} (1 (x^*)^\top) \|^2] \\
\leq 6 L^2 n \mathbb{E}[\| x_i^* - \dddot{x}_i \|^2] + 6 L^2 \sum_{i=1}^{n} \mathbb{E}[\| x_i^* - \dddot{x}_i \|^2] \\
+ 3 \sum_{i=1}^{n} \mathbb{E}[\| \nabla f_{i,i} (x_i^*) \|^2] \\
+ 3 \mathbb{E}[\| \nabla F_{\pi_i} (1 (\dddot{x}_i)^\top) - \nabla F_{\pi_i} (1 (x^*)^\top) \|^2] \\
\]
(32)

We use \(\sigma^2_{\text{shuffle}}\) to bound the last term in (32). From (6):
\[
\mathbb{E}[\| \nabla F_{\pi_i} (1 (\dddot{x}_i)^\top) - \nabla F_{\pi_i} (1 (x^*)^\top) \|^2] \\
= \mathbb{E} \left[ \sum_{i=1}^{n} \| \nabla f_{i,i} (\dddot{x}_i) - \nabla f_{i,i} (x^*) \|^2 \right] \\
\leq \mathbb{E} \left[ 2L \sum_{i=1}^{n} f_{i,i} (\dddot{x}_i) - f_{i,i} (x^*) - \langle \nabla f_{i,i} (x^*), \dddot{x}_i - x^* \rangle \right] \\
= 2nL \mathbb{E} [D_{\delta_i} (\dddot{x}_i, x^*)] \leq 2nL \sigma^2_{\text{shuffle}}.
\]

Next, we bound \(\sum_{i=1}^{n} \mathbb{E}[\| \nabla f_{i,i} (x_i^*) \|^2]\) using \(\sigma^2\):
\[
\sum_{i=1}^{n} \mathbb{E}[\| \nabla f_{i,i} (x_i^*) \|^2] = \sum_{i=1}^{n} \frac{1}{m} \sum_{j=1}^{m} \| \nabla f_{i,j} (x_i^*) \|^2 = n \sigma^2.
\]

Combining the last two steps and (32) finishes the proof of Lemma 17.

\[\qed\]
With the help of Lemma 17, we prove Lemma 7. Let us set $\nabla F_{\pi}(x'_t) := \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,\pi}(x'_{t,i})$. By Lemma 1, we have

$$E[\|x'_{t+1} - 1 (x'_{t})^T\|^2] \leq \rho_{w}^2 E[\|x'_{t} - 1 (x'_{t})^T\|^2 + \alpha^2 \|\nabla F_{\pi}(x'_{t}) - 1 (\nabla F_{\pi}(x'_{t}))^T\|^2 - 2\alpha t (x'_{t} - 1 (x'_{t})^T)]$$

$$\leq \rho_{w}^2 \left(1 + c\right) E[\|x'_{t} - 1 (x'_{t})^T\|^2] + \alpha^2 \rho_{w}^2 (1 + c^{-1}) E[\|\nabla F_{\pi}(x'_{t})\|^2].$$

The last step is due to Cauchy’s and Young’s inequality and holds for any $c > 0$. Invoking Lemma 17, we obtain

$$E[\|x'_{t+1} - 1 (x'_{t})^T\|^2] \leq \rho_{w}^2 \left(1 + c\right) E[\|x'_{t} - 1 (x'_{t})^T\|^2] + 3\alpha\rho_{w}^2 \alpha t (1 + c^{-1}) (\alpha^2 + 2\alpha \sigma_{w}^2) + 6\alpha^2 \rho_{w}^2 (1 + c^{-1}) E[\|x'_{t} - 1 (x'_{t})^T\|^2].$$

In order to guarantee a contractive behavior, we set $c = \left(1 - \rho_{w}^2\right)/4$, then we have $1 + c^{-1} \leq 5/\left(1 - \rho_{w}^2\right)$. In the case $\alpha_t \leq \sqrt{\frac{2 - \rho_{w}^2}{4\alpha\rho_{w}} \cdot \frac{1}{1 - \rho_{w}^2}} L^2$, we obtained the desired result. \[\square\]

C. Proof of Lemma 8

Proof: Step 1: Obtain a combined recursion $H_t^\ell$. Combining Lemmas 5 and 7, we obtain

$$H_{t+1}^\ell \leq \left(1 - \frac{\alpha t \mu}{4}\right) + \frac{30\alpha^2 L^2}{1 - \rho_{w}^2} \omega_t E[\|x_{t} - x_{t+1}\|^2]$$

$$+ \left[\frac{2\alpha \rho_{w}^2 L^2}{1 - \rho_{w}^2} + \frac{1 + \alpha}{2} \rho_{w}^2 \omega_t \right] E[\|x_{t} - 1 (x_{t})^T\|^2]$$

$$+ \frac{2\alpha_t \sigma_{w}^2}{4} \left[1 + \frac{15\alpha_t L \rho_{w}^2}{1 - \rho_{w}^2} \omega_t \right] + \frac{15\alpha_t^2 \rho_{w}^2 \sigma_{w}^2}{1 - \rho_{w}^2} \omega_t$$

where

$$\omega_t \text{ is chosen so that the following inequalities hold for all } t, \ell,$$

$$\left(1 - \frac{\alpha t \mu}{4}\right) + \frac{30\alpha^2 L^2}{1 - \rho_{w}^2} \omega_t \leq 1 - \frac{\alpha t \mu}{4}$$

$$2\frac{\alpha \rho_{w}^2 L^2}{1 - \rho_{w}^2} \left(1 + \alpha \right) + \frac{1 + \alpha}{2} \rho_{w}^2 \omega_t \leq \frac{1 + \alpha}{4} \omega_t$$

We verify the choice of $\omega_t$ in (13). Firstly, (34b) is equivalent to

$$\left(1 - \frac{\rho_{w}^2}{2} - \frac{\alpha t \mu}{4}\right) \omega_t \geq \frac{2\alpha t \omega_t}{n} \left(1 + \alpha t^{\frac{\mu}{4}} \right).$$

Noting $\alpha_t \leq \frac{1 - \rho_{w}^2}{2\mu} \leq \frac{1}{4}$, we have

$$\frac{1 - \rho_{w}^2}{2} \geq \frac{1 - \rho_{w}^2}{2} - \frac{1 - \rho_{w}^2}{8} = \frac{3(1 - \rho_{w}^2)}{4},$$

$$2\alpha t \frac{\frac{\alpha t \omega_t}{n} \left(1 + \alpha t^{\frac{\mu}{4}} \right)}{n \mu} \omega_t \leq \frac{4\omega_t}{n \mu}. $$

Thus, it is sufficient for $\omega_t \geq \frac{16\alpha t L^2}{n \mu(1 - \rho_{w}^2)}$ to satisfy (35).

Secondly, (34a) requires $\omega_t \leq (1 - \frac{\rho_{w}^2}{2\mu}) \frac{1}{\alpha t} \omega_t$ or $\frac{16\alpha t L^2}{n \mu(1 - \rho_{w}^2)} \leq (1 - \frac{\rho_{w}^2}{2\mu}) \frac{1}{\alpha t} \omega_t$. It is sufficient that $\alpha_t \leq \frac{1 - \rho_{w}^2}{4\mu}$. We thus obtain a recursion for $H_t^\ell$ according to (34) and (33):

$$H_{t+1}^\ell \leq \left(1 - \frac{\alpha t \mu}{4}\right) H_t^\ell + 2\alpha t \sigma_{w}^2 \left(1 + \frac{240\alpha^2 \rho_{w}^2 L^3}{\mu(1 - \rho_{w}^2)^2}\right)$$

$$+ \frac{240\alpha^2 \rho_{w}^2 L^2}{\mu(1 - \rho_{w}^2)^2} \sigma_{w}^2 \left(1 - \frac{\rho_{w}^2}{2\mu}\right) \frac{1}{\alpha t} \omega_t.$$
If we use a constant stepsize $\alpha_t = \alpha$, we have $\alpha < \frac{1}{\rho L^3} \leq \frac{1}{(1-\rho_w^2)\mu^2 \rho L^2}$, then

$$8\alpha^2 \sigma^2 \mu \left(\frac{mL + 24\rho_w^2 L^2}{\mu(1-\rho_w^2)^2}\right) \sigma^2 \mu \leq \frac{28(m \mu + L)\sigma^2}{\mu L^2}.$$ 

Therefore, we have an upper bound for the term $H_t$ for both decreasing and constant stepsizes,

$$E \left[ \left\| x_t - \bar{x}_t \right\|^2 \right] \leq H_t \leq H_0 + \frac{28(m \mu + L)\sigma^2}{\mu L^2} = \bar{x}_0. \quad (40)$$

Substitute (40) into Lemma 7 and let

$$\bar{X}_1 := \frac{30nL^2}{1-\rho_w^2} \bar{X}_0 + \frac{15nm \rho_w^2}{1-\rho_w^2} \sigma^2 + \frac{mnp(1-\rho_w^2)}{8L} \sigma^2, \quad (41)$$

we have,

$$E \left[ \left\| x_t - \bar{x}_t \right\|^2 \right] \leq \left(\frac{1 + \rho_w^2}{2}\right)^{t+1} E \left[ \left\| x_0 - 1 \bar{x}_0 \right\|^2 \right] + \alpha_t^2 \bar{X}_1 \leq \left(\frac{1 + \rho_w^2}{2}\right)^{t+1} E \left[ \left\| x_0 - 1 \bar{x}_0 \right\|^2 \right] + \alpha_t^2 \bar{X}_1 \quad (42)$$

Next, we derive recursion among epochs. From (42) and (38) in Supplementary Material, we have

$$E \left[ \left\| x_{t+1} - 1 \bar{x}_{t+1} \right\|^2 \right] \leq \left(\frac{1 + \rho_w^2}{2}\right)^{m} E \left[ \left\| x_0 - 1 \bar{x}_0 \right\|^2 \right] + \alpha_t^2 \bar{X}_1 \quad (43)$$

By similar induction of those in [12], we obtain

$$\sum_{k=0}^{t-1} \left(\frac{1 + \rho_w^2}{2}\right)^{m(t-1-k)} \alpha_t^2 \leq \frac{\alpha_t^2}{\alpha_t^2} \left(1 + \rho_w^2\right)^m \quad (44)$$

Invoking the choice of $\alpha_t$ and $K \geq \frac{24}{1-\rho_w^2}$, we obtain,

$$\frac{\alpha_{t+1}^2}{\alpha_t^2} = \left(\frac{1 + \rho_w^2}{2}\right)^m \left(1 - \frac{1}{t+K+1}\right)^2 \leq \left(\frac{1 + \rho_w^2}{2}\right)^m \geq \left(1 - \frac{1}{K}\right)^2 \geq \left(\frac{1 + \rho_w^2}{2}\right)^m.$$

Combining the above leads to the result. \( \square \)
[26] I. Lobel, A. Ozdaglar, and D. Feijer, “Distributed multi-agent optimization with state-dependent communication,” Math. Program., vol. 129, no. 2, pp. 255–284, 2011.

[27] D. Jakovetic, J. Xavier, and J. M. Moura, “Fast distributed gradient methods,” IEEE Trans. Autom. Control, vol. 59, no. 5, pp. 1131–1146, May 2014.

[28] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, “Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant step-sizes,” in Proc. IEEE 54th Conf. Decis. Control, 2015, pp. 2055–2060.

[29] S. S. Kia, J. Cortés, and S. Martínez, “Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication,” Automatica, vol. 55, pp. 254–264, 2015.

[30] W. Shi, Q. Ling, G. Wu, and W. Yin, “Extra: An exact first-order algorithm for decentralized consensus optimization,” SIAM J. Optim., vol. 25, no. 2, pp. 944–966, 2015.

[31] P. D. Lorenzo and G. Scutari, “NEXT: In-network nonconvex optimization,” IEEE Trans. Signal Inf. Process. Netw., vol. 2, no. 2, pp. 120–136, Jun. 2016.

[32] G. Qu and N. Li, “Harnessing smoothness to accelerate distributed optimization,” IEEE Trans. Control Netw. Syst., vol. 5, no. 3, pp. 1245–1260, Sep. 2018.

[33] A. Nedić, A. Olshevsky, and W. Shi, “Achieving geometric convergence for distributed optimization over time-varying graphs,” SIAM J. Optim., vol. 27, no. 4, pp. 2979–2933, 2017.

[34] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, “Convergence of asynchronous distributed gradient methods over stochastic networks,” IEEE Trans. Autom. Control, vol. 63, no. 2, pp. 434–448, Feb. 2018.

[35] S. Pu, W. Shi, J. Xu, and A. Nedić, “Push-pull gradient methods for distributed optimization in networks,” IEEE Trans. Autom. Control, vol. 66, no. 1, pp. 1–16, Jan. 2021.

[36] J. Chen and A. H. Sayed, “On the limiting behavior of distributed optimization strategies,” in Proc. IEEE 50th Ann. Allerton Conf. Commun., Control, Comput., 2012, pp. 1533–1542.

[37] J. Chen and A. H. Sayed, “On the learning behavior of adaptive networks—Part I: Transient analysis,” IEEE Trans. Inf. Theory, vol. 61, no. 6, pp. 3487–3517, Jun. 2015.

[38] K. Yuan S. A. Alghunaim, and X. Huang, “Removing data heterogeneity influence enhances network topology dependence of decentralized SGD,” 2021, arXiv:2105.08023.

[39] R. Xin, U. A. Khan, and S. Kar, “An improved convergence analysis for decentralized online stochastic non-convex optimization,” IEEE Trans. Signal Process., vol. 69, pp. 1842–1858, 2021.

[40] S. A. Alghunaim and K. Yuan, “A unified and refined convergence analysis for non-convex decentralized learning,” IEEE Trans. Signal Process., vol. 70, pp. 3264–3279, 2022.

[41] L. Bottou, “Curiously fast convergence of some stochastic gradient descent algorithms,” in Proc. Symp. Learn. Data Sci., Paris, vol. 8, 2009, pp. 2624–2633.

[42] L. Bottou, “Stochastic gradient descent tricks,” in Neural Networks: Tricks of the Trade. Berlin, Germany: Springer, 2012, pp. 421–436.

[43] D. Nagaraj, P. Jain, and P. Netrapalli, “SGD without replacement: Sharper rates for general smooth convex functions,” in Proc. Int. Conf. Mach. Learn., 2019, pp. 4703–4711.

[44] T. H. Tran, L. M. Nguyen, and Q. Tran-Dinh, “SMG: A shuffling gradient-based method with momentum,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 10379–10389.

[45] X. Li, A. Milzarek, and J. Qiu, “Convergence of random reshuffling under the Kurdyka-Lojasiewicz inequality,” 2021, arXiv:2110.04926.

[46] K. Yuan, B. Ying, J. Liu, and A. H. Sayed, “Variance-reduced stochastic learning by networked agents under random reshuffling,” IEEE Trans. Signal Process., vol. 67, no. 2, pp. 351–366, Jan. 2019.

[47] X. Jiang, X. Zeng, J. Sun, J. Chen, and L. Xie, “Distributed stochastic proximal algorithm with random reshuffling for nonsmooth finite-sum optimization,” IEEE Trans. Neural Netw. Learn. Syst., 2022.

[48] Y. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, vol. 87. Berlin, Germany: Springer, 2003.

[49] X. Zhang, M. Hong, S. Dhole, W. Yin, and Y. Liu, “FedPD: A federated learning framework with adaptivity to non-IID data,” IEEE Trans. Signal Process., vol. 69, pp. 6055–6070, 2021.

[50] G. Morral, P. Bianchi, and G. Fort, “Success and failure of adaptation-diffusion algorithms with decaying step size in multiagent networks,” IEEE Trans. Signal Process., vol. 65, no. 11, pp. 2798–2813, Jun. 2017.

[51] J. Zeng and W. Yin, “On nonconvex decentralized gradient descent,” IEEE Trans. Signal Process., vol. 66, no. 11, pp. 2834–2848, Jun. 2018.

[52] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, “S-ADDOPT: Decentralized stochastic first-order optimization over directed graphs,” IEEE Contr. Syst. Lett., vol. 5, no. 3, pp. 953–958, Jul. 2020.