Front. Comput. Sci., 2023, 17(2): 172311
https://doi.org/10.1007/s11704-021-1018-0

RESEARCH ARTICLE

Accelerating local SGD for non-IID data using variance reduction

Xianfeng LIANG[1], Shuheng SHEN[2], Enhong CHEN (✉)[1], Jinchang LIU[3], Qi LIU[1], Yifei CHENG[1], Zhen PAN[1]

1 Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China, Hefei 230027, China
2 Ant Financial Services Group, Hangzhou 310000, China
3 Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China

Abstract Distributed stochastic gradient descent and its variants have been widely adopted in the training of machine learning models, which apply multiple workers in parallel. Among them, local-based algorithms, including Local SGD and FedAvg, have gained much attention due to their superior properties, such as low communication cost and privacy-preserving. Nevertheless, when the data distribution on workers is non-identical, local-based algorithms would encounter a significant degradation in the convergence rate. In this paper, we propose Variance Reduced Local SGD (VRL-SGD) to deal with the heterogeneous data. Without extra communication cost, VRL-SGD can reduce the gradient variance among workers caused by the heterogeneous data, and thus it prevents local-based algorithms from slow convergence rate. Moreover, we present VRL-SGD-W with an effective warm-up mechanism for the scenarios, where the data among workers are quite diverse. Benefiting from eliminating the impact of such heterogeneous data, we theoretically prove that VRL-SGD achieves a *linear iteration speedup* with lower communication complexity even if workers access non-identical datasets. We conduct experiments on three machine learning tasks. The experimental results demonstrate that VRL-SGD performs impressively better than Local SGD for the heterogeneous data and VRL-SGD-W is much robust under high data variance among workers.

Keywords distributed optimization, variance reduction, local SGD, federated learning, non-IID data

1 Introduction

For large-scale machine learning problems, stochastic gradient descent (SGD) [1] is a fundamental tool. However, with the expansion of data and model scale, the training of machine learning model, especially deep learning models has become increasingly time-consuming. To accelerate the training process, synchronous stochastic gradient descent (S-SGD), a parallelized version of SGD, has been widely adopted recently, which encourages multiple workers to optimize the model cooperatively. At each iteration, N workers calculate the gradients based on their local data and then communicate the gradients with the parameter server. However, in practice S-SGD suffers from a major drawback: the communication cost is expensive when the number of workers is large. That prevents S-SGD from achieving a *linear time speedup*, which means the total training time is reduced by N times with N workers. Therefore, it is crucial to overcome the communication bottleneck.

In recent years, deep learning has been successfully applied in many fields, such as image recognition [2,3], natural language processing [4], recommender systems [5], intelligent education [6] and finance [7–9]. However, the training of deep learning models has become increasingly time-consuming. To reduce communication cost, several studies [10–14] have managed to lower the communication frequency, which are the so-called local-based algorithms. Among them, Local SGD [12] (also called FedAvg [15]) is a representative local-based algorithm, where workers conduct SGD locally and average model with each other every k iterations as shown in Fig. 1. Compared with S-SGD, local-based algorithms reduce the communication rounds from $O(T)$ to $O(T/k)$, where T is the total number of iterations, and hence accelerate the training process. However, the convergence rate of local-based algorithms has a strong dependence on the extent of non-IID (not independent and identically distributed). They can only exhibit superior performance if the data distribution on workers is identical, which is the so-called identical case. Nevertheless, the identical data assumption is not valid in general, especially in federated learning [16–19], which aims at training on heterogeneous data. When the data distribution is non-identical, which is the so-called non-identical case, the optimization tasks on workers will be different. Specifically, the local model would move towards their local optima and away from the global optima as shown in Fig. 1, hence local-based algorithms would encounter a significant degradation in

Received January 15, 2021; accepted May 27, 2021

E-mail: cheneh@ustc.edu.cn

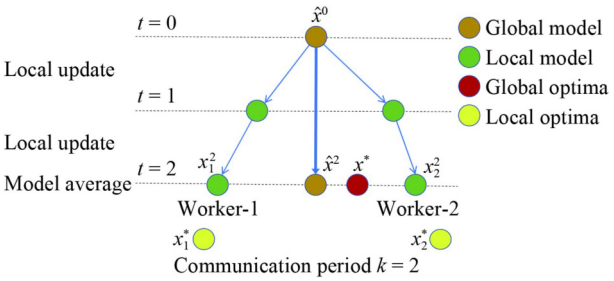


Fig. 1 Procedure of Local SGD with $N = 2$ workers and $k = 2$ local udaptes in the non-identical case. The local models x_i^t (green) move towards their local optima x_i^* (yellow) and away from the global optima x^*(red)

the convergence rate or fail to converge in some cases[1]. Therefore, heterogeneous data has become a fundamentally challenging problem in machine learning. We seek to remove the impact of heterogeneous data, which would make the algorithms converge much faster than the vanilla Local SGD [12].

In this paper, we propose Variance Reduced Local SGD (VRL-SGD), a novel distributed optimization algorithm to accelerate convergence. Benefiting from an additional variance reduction component, VRL-SGD can reduce the the gradient variance among workers, which helps Local SGD to converge faster. For some practical scenarios with high data variance, we present an effective warm-up mechanism, VRL-SGD-W, to eliminate the impact of the high data variance among workers. Consequently, the communication complexity of Local SGD can be reduced from $O(T^{\frac{3}{4}}N^{\frac{3}{4}})$ to $O(T^{\frac{1}{2}}N^{\frac{3}{2}})$[2] in the *non-identical case*, which is crucial for overcoming the communication bottleneck. Therefore, VRL-SGD is more suitable than Local SGD in practice. Contributions are summarized as follows:

- We propose VRL-SGD, a novel distributed optimization algorithm with better communication complexity. Specifically, the communication complexity is reduced from $O(T^{\frac{3}{4}}N^{\frac{3}{4}})$ to $O(T^{\frac{1}{2}}N^{\frac{3}{2}})$ in the *non-identical case*. Meanwhile, VRL-SGD also achieves the same communication complexity $O(T^{\frac{1}{2}}N^{\frac{3}{2}})$ as *Local SGD* in the *identical case*.
- We present VRL-SGD-W, an effective warm-up mechanism deal with some situations, where the data among workers are quite diverse. And the effect of warm-up mechanism is guaranteed both theoretically and experimentally.
- We provide a more intuitive explanation to improve the convergence rate of local-based algorithms and theoretical analysis for VRL-SGD. Besides, our method does not require the extra assumptions, e.g., the gradient variance across workers is bounded.
- We validate the effectiveness of VRL-SGD on standard machine learning tasks. And experimental results show that the proposed algorithm performs significantly better than Local SGD if data distribution among workers is different. Besides, an additional numerical experiment validates the robustness of VRL-SGD-W under high data variance among workers.

2 Related work

Synchronous stochastic gradient descent (S-SGD) is a parallelized version of SGD and is theoretically proved to achieve a *linear iteration speedup* with respect to the number of workers [20,21]. Nevertheless, due to the communication bottleneck, it is hard to achieve *linear time speedup* in practice. To eliminate communication bottlenecks, many distributed SGD-based methods are proposed, such as lossy compression methods [22–27], which use approximations or partial data to represent the gradients, and methods [10,12,13] based on the lower communication frequency.

Among them, *Local SGD* [12], a representative method to lower the communication frequency, has been widely used to train large-scale machine learning models, and its superior performance is verified in several tasks [28–30]. In Local SGD, each worker conducts SGD updates locally and averages its model parameters with others periodically. Previous studies have proven that Local SGD can achieve a linear iteration speedup for both strongly convex [12] and non-convex [13] problems. To fully utilize hardware resources, a variant of Local SGD, called CoCoD-SGD [14], is proposed with decoupling computation and communication. Furthermore, Yu et al. [31] provided a clear linear speedup analysis for *Local SGD* with momentum. However, the rate of convergence for the above algorithms has a poor dependence on the extent of non-IID, which leads to a slow convergence rate for the non-identical case and limits the further reduction of communication cost. Haddadpour et al. [32] verified that the utility of redundant data can lead to lower communication complexity and accelerate training. The redundant data can help reduce the data variance among workers, thus it prevents the slow convergence rate. Nevertheless, this method may be constrained in some cases. For instance, it could not be applied in federated learning [16,17] as data cannot be exchanged between workers for privacy-preserving. Some recent studies [33,34] analyzed the convergence of local-based algorithms on heterogeneous data.

Although there are many studies proposed to reduce the variance in SGD, e.g., SVRG [35], EMGD [36], SAGA [37], and SARAH [38], they could not directly deal with the gradient variance among workers in distributed optimization. In recent years, several studies [39–41] have been proposed to eliminate the gradient variance among workers in the decentralized setting. Among them, a novel decentralized

[1] Under certain settings, *Local SGD* would get a new model $\hat{x}^2 = \hat{x}^0$ after one period, which means that *Local SGD* gets stuck in $\hat{x}^0$ and can not converge to the global optima x^*. A specific case is provided in Appendix A.

[2] The upper bound of k in *Local SGD* is $O(T^{\frac{1}{4}}/N^{\frac{3}{4}})$. By setting $k = O(T^{\frac{1}{4}}/N^{\frac{3}{4}})$, we can observe that the communication complexity $O(T^{\frac{1}{2}}N^{\frac{3}{2}}) = O(T/k^2)$ of *VRL-SGD* is less than that $O(T^{\frac{1}{4}}N^{\frac{3}{4}}) = O(T/k)$ of *Local SGD*.

algorithm, EXTRA [39], provides an ergodic convergence rate for convex problems and a linear convergence rate for strongly convex problems. The D^2 [41] algorithm further applies variance reduction on non-convex stochastic decentralized optimization problems.

To accelerate the training process, we incorporate the variance reduction technique into Local-SGD, which reduce the gradient variance among workers, and hence avoid the extra assumptions, e.g., the bounded variance among workers, in the theoretical analysis. For a better comparison with related algorithms in terms of communication complexity and assumptions, we summarize the results in Table 1. It presents that our algorithm achieves better communication complexity compared with the existing algorithms in the *non-identical case* and does not need extra assumptions.

In a concurrent work, SCAFFOLD [42] is proposed to adopt two learning rates and to communicate an extra variable for variance reduction. However, our algorithm does not require an extra variable, and hence has less communication cost per round. Moreover, we present a warm-up mechanism to remove the impact of high data variance among workers on the convergence rate.

3 Preliminary

In this section, we introduce the problem definition, notations and assumptions used in this paper.

3.1 Problem definition

We focus on data-parallel distributed training, where N workers collaboratively train a machine learning model, and each worker may have its data with different distributions, which is the non-identical case. We use $\mathcal{D}_i$ denote the local data distribution in the ith worker. Specifically, we consider the following finite-sum optimization:

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

where $f_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(x,\xi_i)]$ is the local loss function of the ith worker.

3.2 Notations

First of all, we summarize the key notations as follows.

- $\|\cdot\|$ denotes the ℓ_2 norm of a vector.
- $\mathbb{E}$ denotes that the expectation is taken with respect to all random indexes sampled to calculate stochastic gradients in all iterations.
- x_i^t denotes the local model of the ith worker at the tth iteration.
- $\hat{x}^t$ denotes the average of local models over all N workers, and that is $\hat{x}^t = \frac{1}{N}\sum_{i=1}^{N} x_i^t$.
- $\nabla f_i(x_i^t, \xi_i^t)$ is a stochastic gradient of the ith worker at the tth iteration.
- t' represents the iteration of the last communication, and that is $t' = \left\lfloor \frac{t}{k} \right\rfloor k$.
- t'' represents the iteration of the penultimate communication, and that is $t'' = \left(\left\lfloor \frac{t}{k} \right\rfloor - 1\right)k$.

3.3 Assumptions

Throughout this paper, we make the following assumptions, which are commonly adopted in the theoretical analysis of distributed algorithms [12,31].

Assumption 1

(1) **Lipschitz gradient**: All local functions f_i have L-Lipschitz gradients

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leqslant L\|x-y\|, \forall i, \forall x, y \in \mathbb{R}^d.$$

(2) **Bounded variance within each worker**: There exists a constant σ such that

$$\mathbb{E}_{\xi\sim\mathcal{D}_i}\|\nabla f_i(x,\xi) - \nabla f_i(x)\|^2 \leqslant \sigma^2, \ \ \forall x \in \mathbb{R}^d, \forall i.$$

(3) **Dependence of random variables**: ξ_i^t are independent random variables, where

$$t \in \{0,1,\ldots,T-1\} \text{ and } i \in \{1,2,\ldots,N\}.$$

Please note that previous local-based studies assume that the gradient variance among workers is bounded, or even depend on a stronger assumption, e.g., an upper bound for gradients or identical data distribution on workers, while ours do not require these assumptions.

4 Algorithm

In this section, we first introduce the proposed algorithm and a warm-up mechanism. Then we give an intuitive explanation from the perspective of variance reduction.

4.1 Variance reduced local SGD

We propose VRL-SGD, a variant of Local SGD. VRL-SGD allows locally updating in each worker to reduce the communication cost. But there are a few more steps in VRL-SGD to eliminate the gradient variance among workers. And in VRL-SGD, a worker

1. Communicates with others to get the average of all local models $\hat{x}^t = \frac{1}{N}\sum_{i=1}^{N} x_i^t$.
2. Calculates $\Delta_i^{t'}$, which denotes the average deviation of gradient between the local gradients and the global

Table 1 Comparisons of the communication complexity for different algorithms. The second column and the third column show communication complexity for identical and non-identical datasets respectively. Here, we regard the following assumptions as extra assumptions: (1) an upper bound for gradients; (2) the bounded gradient variance among workers

Reference	Identical data	Non-identical data	Extra assumptions
SGD [20]	T	T	NO
PR-SGD [13]	$O(N^{\frac{3}{4}}T^{\frac{3}{4}})$	$O(N^{\frac{3}{4}}T^{\frac{3}{4}})$	(1)
CoCoD [14]	$O(N^{\frac{3}{2}}T^{\frac{1}{2}})$	$O(N^{\frac{3}{4}}T^{\frac{3}{4}})$	(2)
VRL-SGD	$O(N^{\frac{3}{2}}T^{\frac{1}{2}})$	$O(N^{\frac{3}{2}}T^{\frac{1}{2}})$	NO

gradients in the previous period. And it is defined as

$$\Delta_i^{t'} = \Delta_i^{t''} + \frac{1}{k\gamma}(\hat{x}^t - x_i^t), \tag{2}$$

where k is the communication period and γ is the learning rate.

3. Updates local model k times with a stochastic approximation gradient v_i^t in the form of

$$x_i^{t+1} = x_i^t - \gamma v_i^t. \tag{3}$$

The essential part v_i^t is formed by

$$v_i^t = \nabla f_i(x_i^t, \xi_i^t) - \Delta_i^{t'}. \tag{4}$$

The complete procedure of VRL-SGD is summarized in Algorithm 1. VRL-SGD allows each worker to maintain its local model x_i^t and get the average of all local models every k steps. And VRL-SGD only communicates the local model x_i^t for averaging.

Algorithm 1 Variance reduced local SGD (VRL-SGD)

Input: Initialize $x_i^0 = x^0 \hat{\in} \mathbb{R}^d$, $\Delta_i^0 = \mathbf{0} \in \mathbb{R}^d$, $\forall i$ and $t = 0$. Set learning rate $\gamma > 0$ and communication period $k > 0$.
while $t < T$ **do**
  Worker W_i **does**:
  Communicate with other workers to get the average of all local models: $\hat{x}^t = \frac{1}{N}\sum_{i=1}^N x_i^t$.
  $\Delta_i^t = \Delta_i^{t''} + \frac{1}{k\gamma}(\hat{x}^t - x_i^t)$.
  Update local model $\hat{x}_i^t = x^t$.
  for $\tau = t$ to $t + k - 1$ **do**
    Calculate a stochastic gradient $\nabla f_i(x_i^\tau, \xi_i^\tau)$.
    $v_i^\tau = \nabla f_i(x_i^\tau, \xi_i^\tau) - \Delta_i^t$.
    Each worker updates its local model: $x_i^{\tau+1} = x_i^\tau - \gamma v_i^\tau$.
  end for
  $t = t + k$.
end while

To achieve a linear iteration speedup, Local SGD requires that the communication period k is less than $O(T^{\frac{1}{4}})$. Notice that a better communication period bound $O(T^{\frac{1}{2}})$ can be attained in the *identical case* according to the previous studies [14,31]. Nevertheless, VRL-SGD can attain the better communication period bound $O(T^{\frac{1}{2}})$ in both the *identical case* and the *non-identical case*.

4.2 VRL-SGD with warm-up

Note that VRL-SGD is equivalent to Local SGD in the first period if Δ_i is initialized to 0. To remove the impact of non-IID, we propose an effective warm-up mechanism. We set the first communication period k to 1 in VRL-SGD, which is VRL-SGD with a warm-up (VRL-SGD-W). Essentially, this is equivalent to conducting one S-SGD update and initialize $\Delta_i = \nabla f_i(x_i^0, \xi_i^0) - \frac{1}{N}\sum_{j=1}^N \nabla f_j(x_j^0, \xi_j^0)$. Therefore, the convergence result is not related to the extent of non-IID. This is guaranteed both theoretically and experimentally. Warm-up mechanism is effective for the scenarios, where the data among workers are quite diverse.

4.3 Variance reduction interpretation

Now we illustrate why VRL-SGD can improve the convergence rate compared with Local SGD. VRL-SGD uses an inexact variance reduction technique to reduce the variance among workers. To better understand the intuition of VRL-SGD, let us see the update of Δ_i in Eq. (2). By summing up all Δ_i from 0 to t' and using the fact that $\Delta_i^0 = 0$, we have

$$\Delta_i^{t'} = \frac{1}{k\gamma}\sum_{s=0}^{\lfloor \frac{t}{k} \rfloor}\left(\hat{x}^{ks} - x_i^{ks}\right). \tag{5}$$

Then summing up the equality above over all workers, e.g., $i = 1, 2, \ldots, N$, we can obtain the following equality

$$\begin{aligned}\sum_{i=1}^N \Delta_i^{t'} &= \frac{1}{k\gamma}\sum_{i=1}^N\sum_{s=0}^{\lfloor \frac{t}{k} \rfloor}\left(\hat{x}^{ks} - x_i^{ks}\right) \\ &= \frac{1}{k\gamma}\left(N\sum_{s=0}^{\lfloor \frac{t}{k} \rfloor}\hat{x}^{ks} - \sum_{i=1}^N\sum_{s=0}^{\lfloor \frac{t}{k} \rfloor} x_i^{ks}\right) = 0.\end{aligned} \tag{6}$$

It shows that the expectation of $\Delta_i^{t'}$ over all workers euqals zero. Thus we can obtain the new update formula with respect to $\hat{x}^t$:

$$\begin{aligned}\hat{x}^t &= \hat{x}^{t-1} - \gamma\frac{1}{N}\sum_{i=1}^N v_i^{t-1} \\ &= \hat{x}^{t-1} - \gamma\frac{1}{N}\sum_{i=1}^N\left(\nabla f_i(x_i^t, \xi_i^t) - \Delta_i^{t'}\right) \\ &= \hat{x}^{t-1} - \gamma\frac{1}{N}\sum_{i=1}^N \nabla f_i(x_i^t, \xi_i^t).\end{aligned} \tag{7}$$

It can be noticed that the update of $\hat{x}^t$ in Eq. (7) is in the form of the generalized stochastic gradient descent. In addition, we can obtain a new representation of $\Delta_i^{t'}$:

$$\begin{aligned}\Delta_i^{t'} &= \Delta_i^{t''} + \frac{1}{k\gamma}\left(\hat{x}^{t''} - \gamma\sum_{\tau=t''}^{t'-1}\frac{1}{N}\sum_{j=1}^N v_j^\tau - \hat{x}^{t''} + \gamma\sum_{\tau=t''}^{t'-1} v_i^\tau\right) \\ &= \Delta_i^{t''} + \frac{1}{k\gamma}\left(-\gamma\sum_{\tau=t''}^{t'-1}\frac{1}{N}\sum_{j=1}^N\left(\nabla f_j(x_j^\tau, \xi_j^\tau) - \Delta_j^{t''}\right)\right. \\ &\quad \left. + \gamma\sum_{\tau=t''}^{t'-1}\left(\nabla f_i(x_i^\tau, \xi_i^\tau) - \Delta_i^{t''}\right)\right) \\ &= \frac{1}{k}\sum_{\tau=t''}^{t'-1}\left(\nabla f_i(x_i^\tau, \xi_i^\tau) - \frac{1}{N}\sum_{j=1}^N \nabla f_j(x_j^\tau, \xi_j^\tau)\right).\end{aligned} \tag{8}$$

Substituting Eq. (8) into Eq. (4), we have

$$\begin{aligned}v_i^t &= \nabla f_i(x_i^t, \xi_i^t) - \frac{1}{k}\sum_{\tau=t''}^{t'-1}\nabla f_i(x_i^\tau, \xi_i^\tau) \\ &\quad + \frac{1}{Nk}\sum_{\tau=t''}^{t'-1}\sum_{j=1}^N \nabla f_j(x_j^\tau, \xi_j^\tau).\end{aligned} \tag{9}$$

The representation of v_i^t in Eq. (9) can be regarded as the form of the generalized variance reduction, which is similar to

SVRG [35] and SAGA [37]. To observe that the variance among workers is reduced, we assume that the gradient variance within each worker is zero, which means that we calculate $\nabla f_i(x_i^t)$ in line 8 of Algorithm 1. When all local model x_i^t, x_i^τ and the average model $\hat{x}^t$ converge to the minimum x^*, it holds that

$$\begin{aligned} v_i^t &= \nabla f_i(x_i^t) - \frac{1}{k}\sum_{\tau=t''}^{t'-1}\nabla f_i(x_i^\tau) + \frac{1}{Nk}\sum_{\tau=t''}^{t'-1}\sum_{j=1}^{N}\nabla f_j(x_j^\tau) \\ &\to \quad \nabla f_i(x^*) - \frac{1}{k}\sum_{\tau=t''}^{t'-1}\nabla f_i(x^*) + \frac{1}{Nk}\sum_{\tau=t''}^{t'-1}\sum_{j=1}^{N}\nabla f_j(x^*) \\ &\to \quad \frac{1}{Nk}\sum_{\tau=t''}^{t'-1}\sum_{j=1}^{N}\nabla f_j(x^*) \to \nabla f(x^*) \to 0. \end{aligned} \tag{10}$$

Therefore, v_i^t can converge to zero when the variance within each worker is zero, which helps VRL-SGD converge faster. Note that the local gradient in all workers should be close to 0 while local model x_i^t converge to the optima x^*. In addition, the gradient $\nabla f_i(x_i^t, \xi_i^t)$ in Local SGD is biased and it cannot converge to zero, which prevents the local model x_i^τ from converging to the optima x^*. Therefore it is hard to converge for Local SGD. That is why VRL-SGD performs better than Local SGD in the non-identical case, where the gradient variance among workers is not zero. We provide a specific bad case of Local SGD in Appendix and conduct numerical experiments in Section 6 to support our viewpoint.

4.4 Comparison with previous studies

In this subsection, we compare our proposed method VRL-SGD with related algorithms.

- Comparison with S-SGD [20,21]
  - In S-SGD, N workers communicate the model (or gradients) with the parameter server at each iteration, which indicates the local model x_i^t is always consistent with the global model $\hat{x}^t$. The update formula of $\hat{x}^t$ in *S-SGD* can be written as $\hat{x}^t = \hat{x}^{t-1} - \gamma\sum_{i=1}^{N}\nabla f_i(\hat{x}^t, \xi_i^t)$.
  - **When** $k = 1$. The local model x_i^t has the same update formula as that (Eq.(7)) in $\hat{x}^t$ if we set $k = 1$. Therefore, *VRL-SGD* with $k = 1$ is equivalent to *S-SGD*.
  - **When** $k > 1$. VRL-SGD reduces the number of communication rounds by k times compared with S-SGD.
- Comparison with Local SGD [12]
  - **When** $\Delta_i = 0$. *VRL-SGD* degenerates to *Local SGD* if we set Δ_i to 0 in line 5 of Algorithm 1 all the time.
  - **Under the non-IID setting** In *Local SGD*, the local gradient $\nabla f_i(x_i^\tau, \xi_i^\tau)$ will be biased and hence the local model x_i^t would go away from the global model $\hat{x}^t$, which causes significant degradation in the convergence rate. v_i^τ approximates the global gradient using variance reduction as stated in Eq.(9), which accelerates Local SGD. *Therefore, VRL-SGD is superior to Local SGD in non-identical case.*c>*Local SGD in non-identical case*.
- Comparison with SCAFFOLD [42]
  - SCAFFOLD also uses the idea of variance reduction to prevent local-based algorithms from slow convergence rate, which is quite similar to VRL-SGD. However, they use an extra variable to track local gradients and communicate it for variance reduction, which is not efficient in practice. VRL-SGD can reduce the gradient variance among workers using predictive gradient deviation Δ_i, where Δ_i can be recovered without communication.
  - **Communication cost per round** SCAFFOLD communicates an extra variable for variance reduction. Therefore, the communication cost of SCAFFOLD is at least twice as much as that of VRL-SGD per round.
  - **Warm-up** We also present an effective warm-up mechanism that helps to eliminate the impact of variance among workers. VRL-SGD-W is more robust under high data variance among workers. The effectiveness is guaranteed both theoretically and experimentally.

5 Theoretical analysis

In this section, we provide a theoretical analysis of VRL-SGD. We bound the expected squared gradient norm of the average model, which is the commonly used metric to prove the convergence rate for non-convex problems [20,31,41].

Theorem 1 Under Assumption 1, if the learning rate and the communication period both satisfy that $\gamma \leqslant \frac{1}{2L}$ and $k \leqslant \frac{1}{6\gamma L}$, we have the following inequality for VRL-SGD in Algorithm 1:

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}^t)\|^2 \leqslant& \frac{3(f(\hat{x}^0) - f^*)}{T\gamma} + \frac{3\gamma L\sigma^2}{2N} \\ &+ 11k\gamma^2\sigma^2L^2 + \frac{9k^3\gamma^2L^2C}{T}, \end{aligned}$$

where C is defined as

$$C := \frac{1}{kN}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\left\|\nabla f_i(\hat{x}^t) - \nabla f(\hat{x}^t)\right\|^2. \tag{11}$$

Note that C is a constant related to the extent of non-IID. We can regard C as the gradient variance among workers in the first period. By setting a suitable learning rate γ, we have the following corollary.

Corollary 1 Under Assumption 1, when the learning rate is set as $\gamma = \frac{\sqrt{N}}{\sigma\sqrt{T}}$, and the communication period is set as $k = \min\left\{\frac{\sigma\sqrt{T}}{6LN^{\frac{3}{2}}}, \frac{\sqrt{T}}{\sqrt{N}}\right\}$, we have the following convergence result for Algorithm 1:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}^t)\right\|^2 \leqslant \frac{3\sigma(f(\hat{x}^0) - f^* + L)}{\sqrt{NT}} + \frac{C}{4\sqrt{NT}},$$

where C is defined in Theorem 1.

The detailed proof of Corollary 1 is given in Appendix. Note that the constant C will be 0 when $k = 1$ according to Eq. (11). It is consistent with the fact that when $k = 1$ VRL-SGD is equivalent to S-SGD, where the convergence of S-SGD is not related to the variance among workers.

Corollary 2 VRL-SGD-W If we set the first communication period k to 1 in Corollary 1, which is VRL-SGD with a warm-up mechanism, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}^t)\right\|^2 \leqslant \frac{3\sigma(f(\hat{x}^0)-f^*+L)}{\sqrt{NT}}+\frac{\sigma^2}{4\sqrt{NT}}.$$

In the setting above, the constant in Corollary 1 will become σ^2. Therefore, the convergence result of VRL-SGD-W is not related to the extent of non-IID. The detailed proof is given in Appendix. We also conduct additional experiments in Section 6 to verify this conclusion. Next, we establish the linear iteration speedup and show the communication complexity of VRL-SGD.

Remark 1 Linear speedup For non-convex optimization, if there are N workers training a model collaboratively, according to Corollary 1, VRL-SGD converges at the rate $O(1/\sqrt{NT})$, which is consistent with S-SGD and Local SGD. To achieve the ϵ-optimal solutuioin, $O\left(\frac{1}{N\epsilon^2}\right)$ iterations are needed. Thus, VRL-SGD has a linear iteration speedup with respect to the number of workers.

Remark 2 Communication complexity By Corollary 1, to achieve the convergence rate $O(1/\sqrt{NT})$, we can set the communication period k as $O(T^{\frac{1}{2}}/N^{\frac{3}{2}})$. Consequently, VRL-SGD reduce communication complexity to $O(T^{\frac{1}{2}}N^{\frac{3}{2}})$. However, for the *non-identical case*, previous local-based algorithms only reduce communication complexity to $O(T^{\frac{3}{4}}N^{\frac{3}{4}})$.

6 Experiments

In this section, we will validate the effectiveness of *VRL-SGD* in two cases, the *non-identical case* and the *identical case*. Then we evaluate our algorithm with different communication periods. In the end, we conduct additional experiments to analyze the effect of warm-up.

6.1 Experimental settings

6.1.1 Experimental environment

We implement algorithms with Pytorch 1.1 [43]. And we use a machine with 8 Nvidia Geforce GTX 1080Ti GPUs, 2 Xeon(R) E5-2620 cores and 256 GB RAM Memory. Each GPU is regarded as one worker in experiments.

6.1.2 Baselines

We compare our proposed algorithm *VRL-SGD* with Local SGD [12], SCAFFOLD [42], EASGD [44] and S-SGD [20]. For SCAFFOLD, we use Option II in all experiments, which is consistent with their experiment setting.

6.1.3 Data partitioning

To validate the effectiveness of VRL-SGD in various scenarios, we consider two cases: the non-identical case and the identical case. Under the non-identical case, each worker can only access a subset of data. For example, when five workers are used to train a model on the dataset of 10 classes, each worker can only access to two classes of data. In the *identical case*, we allow each worker to access all data.

6.1.4 Datasets and models

We consider three typical tasks with the most popular methods: (1) LeNet [45] on MNIST [46]; (2) TextCNN [47] on DBPedia [48]; (3) transfer learning on tiny ImageNet, which is a subset of the ImageNet dataset [49]. When training TextCNN on DBPedia, we retain the first 50 words and use a GloVe [50] pre-trained model to extract 50 features for word representation. In transfer learning, we use an Inception V3 [51] pre-trained model as the feature extractor to extract 2,048 features for each image. Then we train a multilayer perceptron with one fully-connected hidden layer of 1,024 nodes, 200 output nodes, and relu activation. All datasets are summarized in Table 2. A lot of deep learning models use batch normalization [52], which assumes that the mini-batches are sampled from the same distribution. Applying batch normalization to the non-identical case may lead to some other issues, which is beyond the scope of this paper.

6.1.5 Parameter setting

For all tasks, we set the weight decay as 10^{-4}. We initialize parameters of all models by performing 2 epoch SGD iterations. Other hyper-parameters can be found in Table 2.

6.1.6 Evaluation metrics

In this paper, we mainly focus on the convergence rate of different algorithms. Local SGD has a more superior training speed performance than S-SGD, which has been empirically observed in various machine learning tasks [28,29]. Besides, VRL-SGD has only a minor change over Local SGD. So VRL-SGD and Local SGD have the same training time in one epoch and both of them have a faster training speed compared with S-SGD. Note that local-based algorithms would have the same communication complexity under the same period k. Therefore, we will compare the convergence rate (the training loss with regard to epochs) of different algorithms. To verify the superiority of VRL-SGD in communication complexity, we will also compare the training loss and test accuracy of different algorithms with regard to the communication size.

6.2 Overall performance on the non-identical case

This paper seeks to address the problem of poor convergence for Local SGD when the variance among workers is high. Therefore, we focus on comparing the convergence rate of all algorithms in the *non-identical case*, where the data variance among workers is maximized.

Figure 2 shows the training loss with regard to epochs on the three tasks: image classification, text classification and

Table 2 Parameters used in experiments and a summary of datasets. N denotes the number of workers, b denotes batch size on each worker, γ is the learning rate, k is the communication period, n represents the number of data samples and m represents the number of data categories

Model	N	b	γ	k	Dataset	n	m
LeNet	8	32	0.005	20	MNIST	60,000	10
TextCNN	8	64	0.01	50	DBPedia	560,000	14
Transfer Learning	8	32	0.025	20	Tiny ImageNet	100,000	200

transfer learning. The results are indicative of the strength of VRL-SGD in the non-identical case. Local SGD converges slowly compared with S-SGD when the communication period k is relatively large, while VRL-SGD enjoys the same convergence rate as that of S-SGD. This is consistent with theoretical analysis that VRL-SGD has a better communication period bound compared to Local SGD. Under the non-IID setting, Local SGD requires $k \leqslant O(T^{\frac{1}{4}}/N^{\frac{3}{4}})$. However, benefiting from variance reduction, VRL-SGD can attain a better communication period bound $O(T^{\frac{1}{2}}/N^{\frac{3}{2}})$ than Local SGD as shown in Corollary 1. Therefore, under the same communication period, VRL-SGD can achieve a linear iteration speedup and converges much faster than Local SGD. To maintain the same convergence rate, Local SGD needs to set a smaller communication period, which will result in higher communication cost. EASGD converges the worst under the same communication period in the non-identical case.

Although both VRL-SGD and SCAFFOLD converge as fast as S-SGD, VRL-SGD is more simpler and has lower communication cost per round. Therefore, VRL-SGD can obtain lower training loss and higher test accuracy under the same communication size as showed in Figs. 4 and 5.

6.3 Overall Performance on the Identical Case

In addition to the above extreme case, we also validate the effectiveness of VRL-SGD in the identical case. As shown in Fig. 3, all algorithms have a similar convergence rate. All local-based algorithms converge as fast as S-SGD when the data distribution on workers is identical.

6.4 Analysis of the communication Period k

In this subsection, we evaluate our algorithm with different communication period k.

As shown in Fig. 6, VRL-SGD converges as fast as S-SGD, while Local SGD, converge slowly even if we set the period k to half of it in Fig. 2. The results show that k in Local SGD should be smaller, such as $k = 2$ or $k = 5$ in transfer learning, which is in line with $\frac{T^{\frac{1}{4}}}{N^{\frac{3}{4}}} = \frac{117,187^{\frac{1}{4}}}{8^{\frac{3}{4}}} \approx 3.9$. However, we can set k to $\frac{T^{\frac{1}{2}}}{N^{\frac{3}{2}}} = \frac{117,187^{\frac{1}{2}}}{8^{\frac{3}{2}}} \approx 15$ in VRL-SGD. We also compare the convergence of different algorithms with a larger k, and observe that the convergence of VRL-SGD will be affected with k, but VRL-SGD is still faster than Local SGD, which is consistent with our theoretical analysis.

6.5 Effectiveness of VRL-SGD-W

In this subsection, we evaluate the effect of warm-up on different variances among workers.

To verify that the convergence of VRL-SGD with warm-up (VRL-SGD-W) is not related to the extent of non-IID, we compare the convergence rate of algorithms in different variance. Therefore, we consider the following optimization

$$\min_{x\in\mathbb{R}} f(x) := \frac{1}{2}(f_1(x) + f_2(x)) = 3x^2 + 6b^2, \tag{12}$$

where $f_1(x) := (x+2b)^2$ and $f_2(x) := 2(x-b)^2$ denote the local loss function of the first and the second worker. In such a setting, the variance among workers is large with a large b.

Figure 7 shows the gap with regard to iteration on different k and b. We can see that Local SGD converges slowly

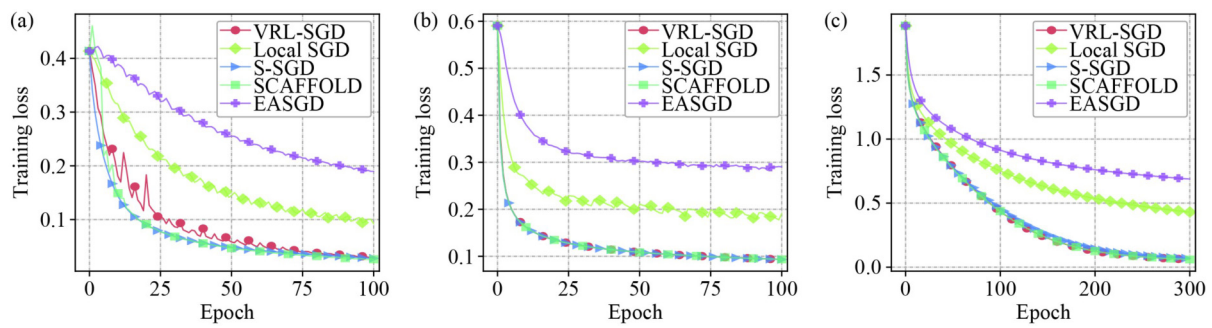


Fig. 2 Epoch loss for the *non-identical case*. VRL-SGD and SCAFFOLD converge as fast as S-SGD, and Local SGD, EASGD converge slowly or even cannot converge. (a) LeNet, MNIST; (b) TextCNN, DBPedia; (c) Transfer Learning, Tiny ImageNet

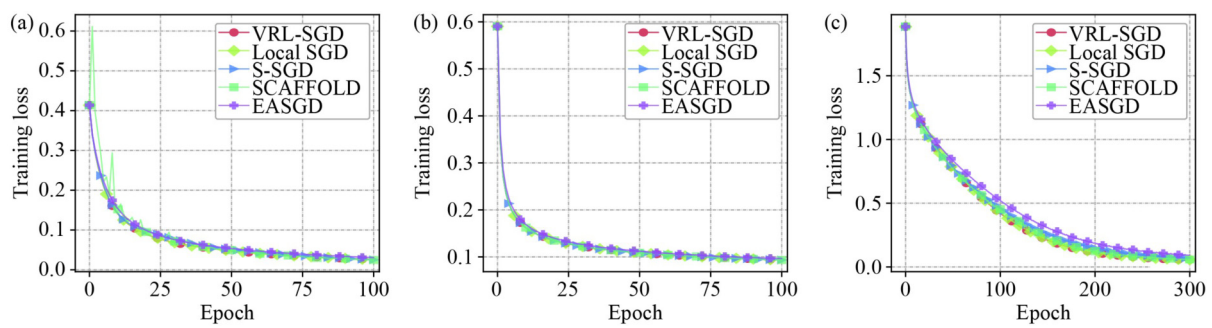


Fig. 3 Epoch loss for the *identical case*. All of the algorithms have a similar convergence rate. (a) LeNet, MNIST; (b) TextCNN, DBPedia; (c) Transfer Learning, Tiny ImageNet

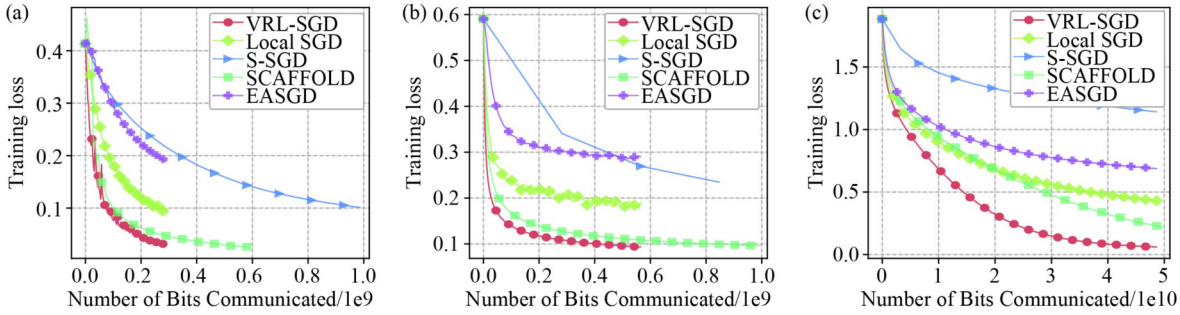


Fig. 4 Train loss in terms of communication size. VRL-SGD outperform other algorithms in terms of communication size. (a) LeNet, MNIST; (b) TextCNN, DBPedia; (c) Transfer Learning, Tiny ImageNet

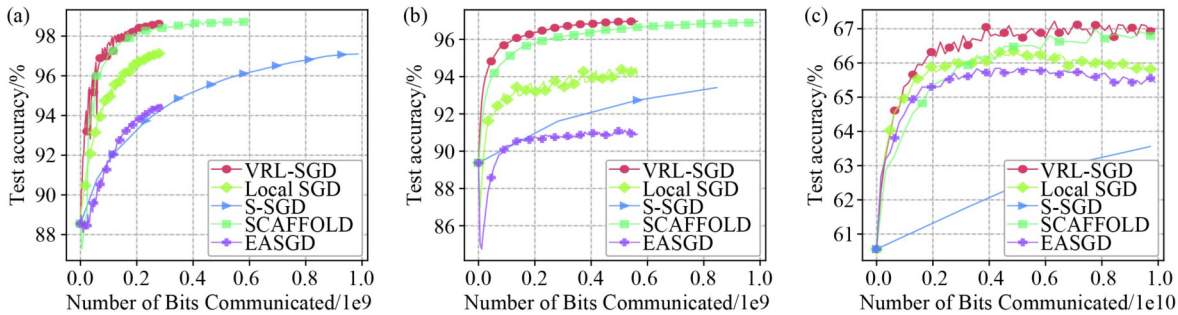


Fig. 5 Test accuracy in terms of communication size. VRL-SGD outperform other algorithms in terms of communication size. (a) LeNet, MNIST; (b) TextCNN, DBPedia; (c) Transfer Learning, Tiny ImageNet

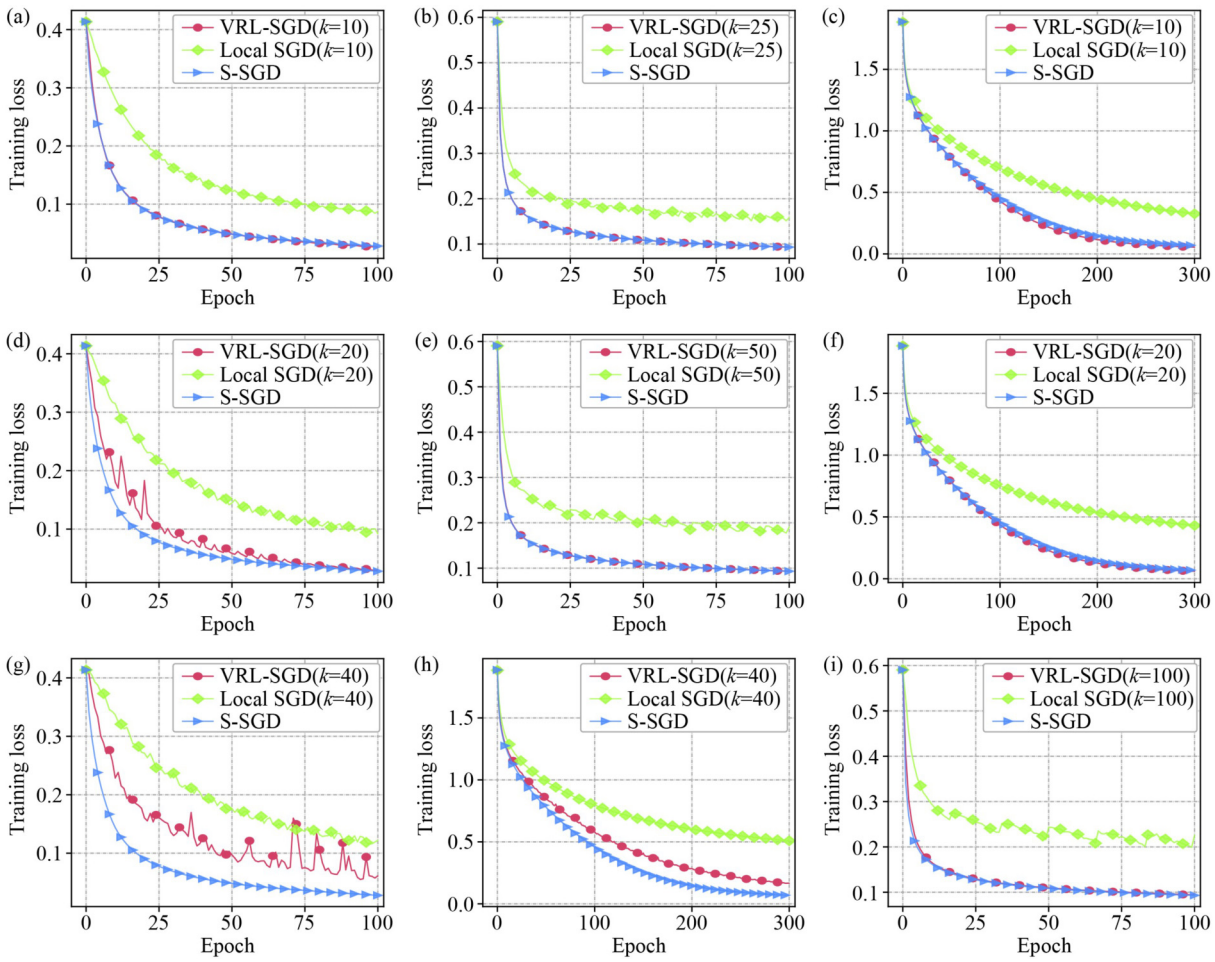


Fig. 6 Epoch loss for the *non-identical case* with different communication period k. (a) LeNet, MNIST; (b) TextCNN, DBPedia; (c) Transfer Learning, Tiny ImageNet; (d) LeNet, MNIST; (e) TextCNN, DBPedia; (f) Transfer Learning, Tiny ImageNet; (g) LeNet, MNIST; (h) Transfer Learning, Tiny ImageNet; (i) TextCNN, DBPedia

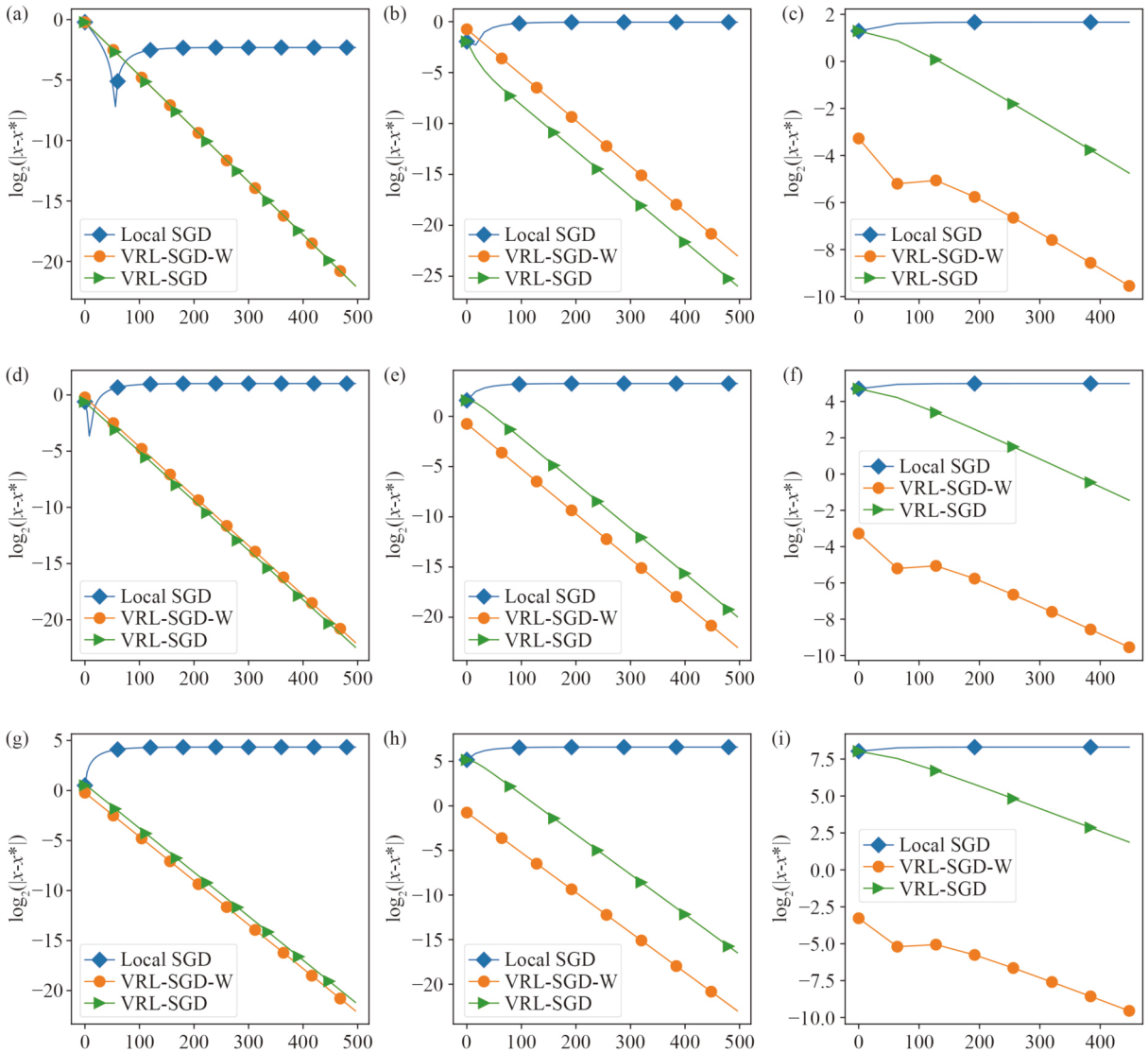


Fig. 7 Logarithm of distance to the global optima for different b and communication period k. (a) b=10, k=4; (b) b=10, k=16; (c) b=10, k=64; (d) b=100, k=4; (e) b=100, k=16; (f) b=100, k=64; (g) b=1000, k=4; (h) b=1000, k=16; (i) b=1000, k=64

compared with VRL-SGD-W and VRL-SGD when the communication period k is relatively large. And the convergence of VRL-SGD is related to b while VRL-SGD-W is not sensitive to b. The experimental results verify our conclusion that VRL-SGD has a better convergence rate compared with Local SGD in the non-identical case, and the convergence of VRL-SGD is not related to the extent of non-IID. Consequently, warm-up is an effective and crucial mechanism for the heterogeneous data, which has been a fundamentally challenging problem in federated learning.

7 Conclusion

In this paper, we proposed a novel distributed algorithm VRL-SGD for accelerating the training of machine learning models. VRL-SGD incorporated the variance reduction technique into Local SGD (FedAvg) to fix the issues of poor convergence for heterogeneous data. Therefore, VRL-SGD further accelerated Local SGD and reduced communication complexity. Compared to SCAFFOLD, VRL-SGD was much simpler and did not require any extra communication for variance reduction. Besides, we presented VRL-SGD-W, an effective warm-up mechanism to remove the impact of the large variance among workers. For non-convex functions, we theoretically proved that VRL-SGD can achieve a linear iteration speedup with lower communication complexity $O(T^{\frac{1}{2}}N^{\frac{3}{2}})$ compared to Local SGD. Moreover, our method did not require the extra assumption, e.g., the gradient variance across workers is bounded. Experimental results demonstrated VRL-SGD was significantly better than traditional Local SGD for the non-identical case and VRL-SGD-W was much robust under high data variance among workers.

Acknowledgements This research was partially supported by grants from the National Key Research and Development Program of China (No. 2018YFC0832101), and the National Natural Science Foundation of China (Grant Nos. U20A20229 and 61922073).

Appendixes

Appendix A:

A bad case of Local SGD

Consider the such optimization

$$\min_{x\in\mathbb{R}} f(x) := \frac{1}{2}(f_1(x)+f_2(x)) = 3x^2+6b^2, \tag{A1}$$

where $f_1(x) := (x+2b)^2$ and $f_2(x) := 2(x-b)^2$ denote the local loss function of the first and the second worker. Set learning rate $\gamma = \frac{1}{3}$, communication period $k=2$, and initialize $x_i^0 = \hat{x}^0 = -\frac{b}{2}$. In such a setting, Local SGD would get a new model $\hat{x}^2 = \hat{x}^0$ after one period, which means that Local SGD can not converge to the global optima $x^* = 0$. The detailed steps are as follows:

For the first worker,

$$\begin{aligned} t=1, x_1^1 &= \hat{x}^0 - \gamma\nabla f_i(\hat{x}^0) \\ &= \hat{x}^0 - 2\gamma(\hat{x}^0+2b) \\ &= -\frac{b}{2} - \frac{2}{3}\left(-\frac{b}{2}+2b\right) = -\frac{3}{2}b, \\ t=2, x_1^2 &= x_1^1 - \gamma\nabla f_i(x_1^1) \\ &= x_1^1 - 2\gamma(x_1^1+2b) \\ &= -\frac{3}{2}b - \frac{2}{3}\left(-\frac{3}{2}b+2b\right) = -\frac{11}{6}b. \end{aligned}$$

For the second worker,

$$\begin{aligned} t=1, x_2^1 &= \hat{x}^0 - \gamma\nabla f_i(\hat{x}^0) \\ &= \hat{x}^0 - 4\gamma(\hat{x}^0-b) \\ &= -\frac{b}{2} - \frac{4}{3}\left(-\frac{b}{2}-b\right) = \frac{3}{2}b, \\ t=2, x_2^2 &= x_2^1 - \gamma\nabla f_i(x_2^1) \\ &= x_2^1 - 4\gamma(x_2^1-b) \\ &= \frac{3}{2}b - \frac{4}{3}\left(\frac{3}{2}b-b\right) = \frac{5}{6}b. \end{aligned}$$

By averaging local models x_1^2 and x_2^2, we can get $\hat{x}^2 = \frac{x_1^2+x_2^2}{2} = -\frac{b}{2} = \hat{x}^0$. However, the global optima x^* of $f(x) = 3x^2+6b^2$ is obviously 0, which means that Local SGD gets stuck in $\hat{x}^0 = -\frac{b}{2}$ and can not converge in such case. Even if γ is set small enough, Local SGD could also meet above problems.

Appendix B: Linear iteration speed up

In this section, we conduct some experiments to verify that VRL-SGD achieves the linear iteration speedup. Figure A1 shows the iteration speedup of VRL-SGD with regard to the training loss for three models, by varying the number of workers N. The speedup is over a single worker iteration for reaching the target loss. For all tasks, we set the target loss as 0.1 and the learning rate can be found in Table 2. The experimental results show that with the increase of the number of workers, VRL-SGD can achieve the linear iteration speedup.

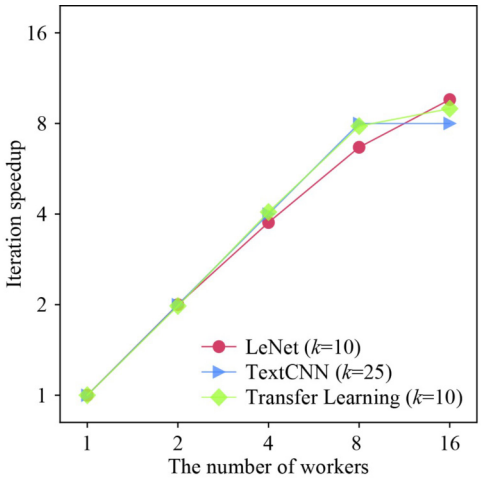


Fig. A1 The iteration speedup of *VRL-SGD* by varying the number of workers N

Appendix C: Proof of auxiliary lemma

Lemma 1 z_i are independent random variables, where $i \in \{1,2,\ldots,m\}$. If their variance is bounded $\mathbb{E}\|z_i - \mathbb{E}[z_i]\|^2 \leqslant \sigma^2$, then we have

$$\mathbb{E}\left\|\sum_{i=1}^m c_i z_i\right\|^2 \leqslant \left\|\sum_{i=1}^m c_i \mathbb{E}[z_i]\right\|^2 + \sum_{i=1}^m c_i^2\sigma^2. \tag{A2}$$

Proof

$$\begin{aligned} &\mathbb{E}\left\|\sum_{i=1}^m c_i z_i\right\|^2 \\ =&\mathbb{E}\left\|\sum_{i=1}^m c_i (z_i - \mathbb{E}[z_i]) + \sum_{i=1}^m c_i \mathbb{E}[z_i]\right\|^2 \\ =&\sum_{i=1}^m c_i^2 \mathbb{E}\|z_i - \mathbb{E}[z_i]\|^2 + \left\|\sum_{i=1}^m c_i \mathbb{E}[z_i]\right\|^2 \\ &+ 2\mathbb{E}\langle \sum_{i=1}^m c_i (z_i - \mathbb{E}[z_i]), \sum_{i=1}^m c_i \mathbb{E}[z_i]\rangle \\ &+ \sum_{0\leqslant i<j\leqslant m} 2\mathbb{E}\langle c_i (z_i - \mathbb{E}[z_i]), c_j \left(z_j - \mathbb{E}[z_j]\right)\rangle \\ =&\left\|\sum_{i=1}^m c_i \mathbb{E}[z_i]\right\|^2 + \sum_{i=1}^m c_i^2 \mathbb{E}\|z_i - \mathbb{E}[z_i]\|^2 \\ \leqslant&\left\|\sum_{i=1}^m c_i \mathbb{E}[z_i]\right\|^2 + \sum_{i=1}^m c_i^2\sigma^2. \end{aligned} \tag{A3}$$

Setting $c_i = 1$ for $i = 1,\ldots,m$, then we have

$$\mathbb{E}\left\|\sum_{i=1}^m z_i\right\|^2 \leqslant \left\|\sum_{i=1}^m \mathbb{E}[z_i]\right\|^2 + m\sigma^2. \tag{A4}$$

Appendix D: Proof of partially accumulated local gradients

In this section, we present Lemma 2, Lemma 4 and Lemma 3 to bound the partially accumulated local gradients in the sth communication stage.

It is defined as, for $t < k_1$ using Option 1,

$$v_i^t = \nabla f(x_i^t, \xi_i^t), \tag{A5}$$

for $t < k_1$ using Option 2,

$$v_i^t = \nabla f(x_i^t, \xi_i^t) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(x_j^0, \xi_j^0) - \nabla f_i(x_i^0, \xi_i^0), \tag{A6}$$

for $t \geqslant k_1$,

$$v_i^t = \nabla f(x_i^t, \xi_i^t) + \frac{1}{k_{s-1}} \sum_{\tau'=t''}^{t'-1} \left(\frac{1}{N}\sum_{j=1}^{N} \nabla f_j(x_j^{\tau'}, \xi_j^{\tau'}) - \nabla f_i(x_i^{\tau'}, \xi_i^{\tau'}) \right), \tag{A7}$$

where k_s denotes the s-th communication stage, $t' = t'' + k_{s-1}$.

In the first communication stage, Option 1 is equivalent to Local SGD, and Option 2 (Warm-up) is consistent with $k_{s-1} = 1$.

Lemma 2 Under Assumption 1, we have the following inequality for $t \geqslant k$

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\| \sum_{\tau=t'}^{t-1} v_i^\tau \|^2 \\
\leqslant& \frac{6L^2}{N} \sum_{i=1}^{N} \Bigg(k_s \sum_{\tau=t'}^{t-1} \mathbb{E}\|x_i^\tau - \hat{x}^\tau\|^2 + \frac{2k_s}{k_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \mathbb{E}\|\hat{x}^\tau - \hat{x}^{\tau'}\|^2 \\
&+ \frac{2k_s^2}{k_{s-1}} \sum_{\tau'=t''}^{t'-1} \mathbb{E}\|\hat{x}^{\tau'} - x_i^{\tau'}\|^2 \Bigg) + 6k_s \sum_{\tau=t'}^{t-1} \|\nabla f(\hat{x}^\tau)\|^2 + \frac{3k_s^2}{k_{s-1}}\sigma^2,
\end{aligned}$$

where $t' \leqslant t < t' + k_s$, $t' = t'' + k_{s-1}$ and $k_{s-1} \leqslant k_s$.

Proof By the definition of v_i^t in Eq. (A7), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\| \sum_{\tau=t'}^{t-1} v_i^\tau \right\|^2 \\
=& \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Bigg\| \sum_{\tau=t'}^{t-1} \nabla f_i(x_i^\tau, \xi_i^\tau) - \frac{1}{k_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \nabla f_i(x_i^{\tau'}, \xi_i^{\tau'}) \\
&+ \frac{1}{Nk_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \sum_{j=1}^{N} \nabla f_j(x_j^{\tau'}, \xi_j^{\tau'}) \Bigg\|^2 \\
=& \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\Bigg\| \sum_{\tau=t'}^{t-1} \nabla f_i(x_i^\tau, \xi_i^\tau) + \sum_{\tau'=t''}^{t'-1} \sum_{j\neq i}^{N} \frac{t-t'}{Nk_{s-1}} \nabla f_j(x_j^{\tau'}, \xi_j^{\tau'}) \\
&- \sum_{\tau'=t''}^{t'-1} \frac{(N-1)(t-t')}{Nk_{s-1}} \nabla f_i(x_i^{\tau'}, \xi_i^{\tau'}) \Bigg\|^2 \\
\leqslant& \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}\left\| \sum_{\tau=t'}^{t-1} \left(\nabla f_i(x_i^\tau) + \frac{1}{k_{s-1}} \sum_{\tau'=t''}^{t'-1} \left(\frac{1}{N}\sum_{j=1}^{N} \nabla f_j(x_j^{\tau'}) - \nabla f_i(x_i^{\tau'}) \right) \right) \right\|^2}_{T_2} \\
&+ \sum_{\tau=t'}^{t-1} \sigma^2 + \sum_{\tau'=t''}^{t'-1} \sum_{j\neq i}^{N} \frac{(t-t')^2}{N^2 k_{s-1}^2} \sigma^2 \\
&+ \sum_{\tau'=t''}^{t'-1} \frac{(N-1)^2 (t-t')^2}{N^2 k_{s-1}^2} \sigma^2 \\
\leqslant& \frac{1}{N}\sum_{i=1}^{N} T_2 + \frac{3k_s^2}{k_{s-1}}\sigma^2
\end{aligned} \tag{A8}$$

where the first inequality comes from Lemma 1 and the second inequality follows from $t - t' < k_s$, $k_{s-1} \leqslant k_s$. We next bound T_2 as

$$\begin{aligned}
&T_2 \\
=& \mathbb{E}\left\| \frac{1}{Nk_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \sum_{j=1}^{N} \left(\nabla f_i(x_i^\tau) + \nabla f_j(x_j^{\tau'}) - \nabla f_i(x_i^{\tau'}) \right) \right\|^2 \\
\leqslant& \frac{k_s}{k_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \mathbb{E}\left\| \nabla f_i(x_i^\tau) - \nabla f_i(x_i^{\tau'}) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(x_j^{\tau'}) \right\|^2 \\
=& \frac{k_s}{k_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \mathbb{E}\Bigg\| \nabla f_i(x_i^\tau) - \nabla f_i(x_i^{\tau'}) + \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(x_j^{\tau'}) \\
&- \nabla f_i(\hat{x}^\tau) + \nabla f_i(\hat{x}^\tau) - \nabla f_i(\hat{x}^{\tau'}) + \nabla f_i(\hat{x}^{\tau'}) \\
&- \frac{1}{N}\sum_{j=1}^{N} \nabla f_j(\hat{x}^{\tau'}) + \nabla f(\hat{x}^{\tau'}) - \nabla f(\hat{x}^\tau) + \nabla f(\hat{x}^\tau) \Bigg\|^2 \\
\leqslant& \frac{6k_s}{Nk_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \sum_{j=1}^{N} \Bigg(\mathbb{E}\left\| \nabla f_i(x_i^\tau) - \nabla f_i(\hat{x}^\tau) \right\|^2 \\
&+ \mathbb{E}\left\| \nabla f_i(\hat{x}^\tau) - \nabla f_i(\hat{x}^{\tau'}) \right\|^2 + \mathbb{E}\left\| \nabla f_i(\hat{x}^{\tau'}) - \nabla f_i(x_i^{\tau'}) \right\|^2 \\
&+ \mathbb{E}\left\| \nabla f_j(x_j^{\tau'}) - \nabla f_j(\hat{x}^{\tau'}) \right\|^2 + \mathbb{E}\left\| \nabla f(\hat{x}^{\tau'}) - \nabla f(\hat{x}^\tau) \right\|^2 \\
&+ \mathbb{E}\left\| \nabla f(\hat{x}^\tau) \right\|^2 \Bigg) \\
\leqslant& \frac{6k_s L^2}{Nk_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \sum_{j=1}^{N} \Bigg(\mathbb{E}\left\| x_i^\tau - \hat{x}^\tau \right\|^2 + 2\mathbb{E}\left\| \hat{x}^\tau - \hat{x}^{\tau'} \right\|^2 \\
&+ \mathbb{E}\left\| \hat{x}^{\tau'} - x_i^{\tau'} \right\|^2 + \mathbb{E}\left\| x_j^{\tau'} - \hat{x}^{\tau'} \right\|^2 \Bigg) \\
&+ 6k_s \sum_{\tau=t'}^{t-1} \mathbb{E}\left\| \nabla f(\hat{x}^\tau) \right\|^2,
\end{aligned} \tag{A9}$$

where the first two inequalities follow from Cauchy's inequality, and the last inequality follows from the Lipschitz gradient assumption. According to Eq. (A9), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} T_2 \\
\leqslant& \frac{6L^2}{N} \sum_{i=1}^{N} \Bigg(k_s \sum_{\tau=t'}^{t-1} \mathbb{E}\left\| x_i^\tau - \hat{x}^\tau \right\|^2 + \frac{2k_s}{k_{s-1}} \sum_{\tau=t'}^{t-1} \sum_{\tau'=t''}^{t'-1} \mathbb{E}\left\| \hat{x}^\tau - \hat{x}^{\tau'} \right\|^2 \\
&+ \frac{2k_s^2}{k_{s-1}} \sum_{\tau'=t''}^{t'-1} \mathbb{E}\left\| \hat{x}^{\tau'} - x_i^{\tau'} \right\|^2 \Bigg) + 6k_s \sum_{\tau=t'}^{t-1} \mathbb{E}\left\| \nabla f(\hat{x}^\tau) \right\|^2.
\end{aligned} \tag{A10}$$

Substituting Eq. (A10) into Eq. (A8), we obtain Lemma 2.

Lemma 3 Under Assumption 1, we have the following inequality for $t < k$ using Option 1,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left\| \sum_{\tau=t'}^{t-1} v_i^\tau \right\|^2 \\
\leqslant& \frac{3kL^2}{N} \sum_{i=1}^{N} \sum_{\tau=t'}^{t-1} \mathbb{E}\|x_i^\tau - \hat{x}^\tau\|^2 + 3k \sum_{\tau=t'}^{t-1} \|\nabla f(\hat{x}^\tau)\|^2 + k\sigma^2 \\
&+ \frac{3k}{N} \sum_{i=1}^{N} \sum_{\tau=t'}^{t-1} \left\| \nabla f_i(\hat{x}^\tau) - \nabla f(\hat{x}^\tau) \right\|^2.
\end{aligned} \tag{A11}$$

Proof By the definition of v_i^t in Eq. (A5), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}v_i^\tau\right\|^2\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\nabla f_i(x_i^\tau,\xi_i^\tau)\right\|^2\\
\leqslant&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\nabla f_i(x_i^\tau)\right\|^2+k\sigma^2\\
\leqslant&\frac{k}{N}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f_i(x_i^\tau)\right\|^2+k\sigma^2\\
=&\frac{k}{N}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\nabla f_i(x_i^\tau)-\nabla f_i(\hat{x}^\tau)+\nabla f_i(\hat{x}^\tau)\right.\\
&\left.-\nabla f(\hat{x}^\tau)+\nabla f(\hat{x}^\tau)\right\|^2+k\sigma^2\\
\leqslant&\frac{3k}{N}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\left(\mathbb{E}\left\|\nabla f_i(x_i^\tau)-\nabla f_i(\hat{x}^\tau)\right\|^2\right.\\
&\left.+\mathbb{E}\left\|\nabla f_i(\hat{x}^\tau)-\nabla f(\hat{x}^\tau)\right\|^2+\mathbb{E}\left\|\nabla f(\hat{x}^\tau)\right\|^2\right)+k\sigma^2\\
\leqslant&\frac{3k}{N}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\left(L^2\mathbb{E}\left\|x_i^\tau-\hat{x}^\tau\right\|^2+\mathbb{E}\left\|\nabla f_i(\hat{x}^\tau)-\nabla f(\hat{x}^\tau)\right\|^2\right.\\
&\left.+\mathbb{E}\left\|\nabla f(\hat{x}^\tau)\right\|^2\right)+k\sigma^2,
\end{aligned}\tag{A12}$$

where the first inequality comes from Lemma 1, the second and third inequalities follow from Cauchy's inequality, and the last inequality follows from the Lipschitz gradient assumption. Rerrangeing the inequality, we obtain Lemma 3.

Lemma 4 Under Assumption 1, we have the following inequality for $t < k$ using Option 2:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|\sum_{\tau=t'}^{t-1}v_i^\tau\|^2\\
\leqslant&\frac{6L^2}{N}\sum_{i=1}^{N}\left(k\sum_{\tau=t'}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2+2k\sum_{\tau=t'}^{t-1}\mathbb{E}\|\hat{x}^\tau-\hat{x}^0\|^2\right)\\
&+6k\sum_{\tau=t'}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2+3k^2\sigma^2.
\end{aligned}\tag{A13}$$

Proof According to Lemma 2, we can derive Lemma 4 by setting $k_{s-1}=1$.

Appendix E: Proof of VRL-SGD without Warm-up

In this section, we give the proof of VRL-SGD. First, we introduce Lemma 5, which bounds the difference between the local model x_i^t and the average model $\hat{x}^t$.

Lemma 5 Under Lemma 2 and Lemma 3, when the learning rate γ and the communication period k satisfy that $\gamma\leqslant\frac{1}{2L}$ and $6k\gamma L\leqslant 1$, we have the following inequality

$$\begin{aligned}
&\frac{1}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\|x_i^t-\hat{x}^t\|^2\\
\leqslant&12k^2\gamma^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2+96k^4\gamma^4L^2\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
&+\frac{22k\gamma^2\sigma^2T}{3}+6k^3\gamma^2C,
\end{aligned}\tag{A14}$$

where $C:=\frac{1}{kN}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\left\|\nabla f_i(\hat{x}^t)-\nabla f(\hat{x}^t)\right\|^2$.

Proof According to the updating scheme in Algorithms 1, x_i^t can be represented as

$$x_i^t=\hat{x}^{t'}-\gamma\sum_{\tau=t'}^{t-1}v_i^\tau.\tag{A15}$$

By the definition of $\hat{x}^t$, we can represent it as

$$\hat{x}^t=\hat{x}^{t'}-\frac{\gamma}{N}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}v_i^\tau.\tag{A16}$$

Substituting Eq. (A15) and Eq. (A16) into the left side of Eq. (A14), we have

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{x}^t-x_i^t\right\|^2\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\left(\hat{x}^{t'}-\frac{\gamma}{N}\sum_{\tau=t'}^{t-1}\sum_{j=1}^{N}v_j^\tau\right)-\left(\hat{x}^{t'}-\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau\right)\right\|^2\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau-\frac{\gamma}{N}\sum_{\tau=t'}^{t-1}\sum_{j=1}^{N}v_j^\tau\right\|^2\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau\right\|^2+\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\frac{\gamma}{N}\sum_{\tau=t'}^{t-1}\sum_{j=1}^{N}v_j^\tau\right\|^2\\
&-2\sum_{i=1}^{N}\frac{1}{N}\mathbb{E}\langle\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau,\frac{\gamma}{N}\sum_{\tau=t'}^{t-1}\sum_{j=1}^{N}v_j^\tau\rangle\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau\right\|^2+\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{1}{N}v_j^\tau\right\|^2\\
&-2\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma\sum_{j=1}^{N}\frac{1}{N}v_j^\tau\right\|^2\\
=&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau\right\|^2-\mathbb{E}\left\|\frac{\gamma}{N}\sum_{\tau=t'}^{t-1}\sum_{j=1}^{N}\nabla f_j(x_i^\tau,\xi_j^\tau)\right\|^2\\
\leqslant&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\sum_{\tau=t'}^{t-1}\gamma v_i{}^\tau\right\|^2.
\end{aligned}\tag{A17}$$

According to the result in Lemma 2 and Lemma 3, for $t\geqslant k$, we have

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{x}^t-x_i^t\right\|^2\\ \leqslant&\frac{6\gamma^2L^2}{N}\sum_{i=1}^{N}\left(k\sum_{\tau=t'}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2+2\sum_{\tau=t'}^{t-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^\tau-\hat{x}^{\tau'}\|^2\right.\\ &\left.+2k\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^{\tau'}-x_i^{\tau'}\|^2\right)+6k\gamma^2\sum_{\tau=t'}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2+3k\gamma^2\sigma^2,\end{aligned}\tag{A18}$$

and for $t<k$, we have

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{x}^t-x_i^t\right\|^2\\ \leqslant&\frac{3k\gamma^2L^2}{N}\sum_{i=1}^{N}\sum_{\tau=0}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2+3k\gamma^2\sum_{\tau=0}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2\\ &+k\gamma^2\sigma^2+\frac{3k\gamma^2}{N}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\left\|\nabla f_i(\hat{x}^\tau)-\nabla f(\hat{x}^\tau)\right\|^2.\end{aligned}\tag{A19}$$

Summing up Eq. (A18) and Eq. (A19) from $t=0$ to $T-1$, we obtain

$$\begin{aligned}&\frac{1}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{x}^t-x_i^t\right\|^2\\ \leqslant&\frac{6\gamma^2L^2}{N}\sum_{t=k}^{T-1}\sum_{i=1}^{N}\left(k\sum_{\tau=t'}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2\right.\\ &\left.+2\sum_{\tau=t'}^{t-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^\tau-\hat{x}^{\tau'}\|^2+2k\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^{\tau'}-x_i^{\tau'}\|^2\right)\\ &+3k\gamma^2\sigma^2(T-k)+6k\gamma^2\sum_{t=k}^{T-1}\sum_{\tau=t'}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2\\ &+\frac{3k\gamma^2L^2}{N}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2+3k\gamma^2\sum_{t=0}^{k-1}\sum_{\tau=t'}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2\\ &+k^2\gamma^2\sigma^2+\frac{3k\gamma^2}{N}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\left\|\nabla f_i(\hat{x}^\tau)-\nabla f(\hat{x}^\tau)\right\|^2\\ \leqslant&\frac{18\gamma^2k^2L^2}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\|x_i^t-\hat{x}^t\|^2\\ &+12k\gamma^2L^2\underbrace{\sum_{t=k}^{T-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^t-\hat{x}^{\tau'}\|^2}_{T_6}+6k^2\gamma^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2\\ &+3k\gamma^2\sigma^2T+3k^3\gamma^2C,\end{aligned}\tag{A20}$$

where the second and the third inequalities can be obtained by using a simple counting argument and C is defined as $C:=\frac{1}{kN}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\left\|\nabla f_i(\hat{x}^t)-\nabla f(\hat{x}^t)\right\|^2$.

Next, we bound T_6.

$$\begin{aligned}T_6=&\sum_{t=k}^{T-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\left\|\sum_{s=\tau'}^{t-1}\frac{\gamma}{N}\sum_{i=1}^{N}v_i^s\right\|^2\\ =&\frac{\gamma^2}{N^2}\sum_{t=k}^{T-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\left\|\sum_{s=\tau'}^{t-1}\sum_{i=1}^{N}\nabla f_i(x_i^s,\xi_i^s)\right\|^2\\ \leqslant&\frac{2k^2\gamma^2\sigma^2(T-k)}{N}+\sum_{t=k}^{T-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\left\|\frac{\gamma}{N}\sum_{s=\tau'}^{t-1}\sum_{i=1}^{N}\nabla f_i(x_i^s)\right\|^2\\ \leqslant&\frac{2k^2\gamma^2\sigma^2(T-k)}{N}+2k\gamma^2\sum_{t=k}^{T-1}\sum_{\tau'=t''}^{t'-1}\sum_{s=\tau'}^{t-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^s)\right\|^2\\ \leqslant&\frac{2k^2\gamma^2\sigma^2(T-k)}{N}+4k^3\gamma^2\sum_{t=k}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\ &+2k^3\gamma^2\sum_{t=0}^{k-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\ \leqslant&2k^2\gamma^2\sigma^2T+4k^3\gamma^2\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2,\end{aligned}\tag{A21}$$

where the first inequality comes from Lemma 1.

Substituting Eq. (A21) into Eq. (A20) and rerrangeing the inequality, we obtain

$$\begin{aligned}&(1-18k^2\gamma^2L^2)\frac{1}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\|x_i^t-\hat{x}^t\|^2\\ \leqslant&6k^2\gamma^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2+3k\gamma^2\sigma^2T+3k^3\gamma^2C\\ &+24k^3\gamma^4\sigma^2L^2T+48k^4\gamma^4L^2\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\ \leqslant&6k^2\gamma^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2+\frac{11k\gamma^2\sigma^2T}{3}+3k^3\gamma^2C\\ &+48k^4\gamma^4L^2\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2,\end{aligned}\tag{A22}$$

where the last equality holds because $36k^2\gamma^2L^2\leqslant 1$.

Dividing $(1-36k^2\gamma^2L^2)$ on both sides and using $1-18k^2\gamma^2L^2\geqslant\frac{1}{2}$ complete the proof.

Theorem 1 Under Assumption 1, if the learning rate and the communication period both satisfy that $\gamma\leqslant\frac{1}{2L}$ and $k\leqslant\frac{1}{6\gamma L}$, we have the following inequality for \$\textrm{VRL-SGD}\$ in Algorithm 1:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}^t)\|^2\leqslant&\frac{3(f(\hat{x}^0)-f^*)}{T\gamma}+\frac{3\gamma L\sigma^2}{2N}\\ &+11k\gamma^2\sigma^2L^2+\frac{9k^3\gamma^2L^2C}{T},\end{aligned}$$

where $C:=\frac{1}{kN}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\left\|\nabla f_i(\hat{x}^t)-\nabla f(\hat{x}^t)\right\|^2$.

Proof Since $f_i(\cdot), i=1,2,\ldots,N$ are L-smooth, it is easy to verify that $f(\cdot)$ is L-smooth. We have

$$
\begin{aligned}
&f(\hat{x}_{t+1})\\
\leqslant& f(\hat{x}^t)+\left\langle\nabla f(\hat{x}^t),\hat{x}^{t+1}-\hat{x}^t\right\rangle+\frac{L}{2}\left\|\hat{x}^{t+1}-\hat{x}^t\right\|^2\\
=& f(\hat{x}^t)-\gamma\left\langle\nabla f(\hat{x}^t),\frac{1}{N}\sum_{i=1}^{N}v_i^t\right\rangle+\frac{L\gamma^2}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}v_i^t\right\|^2\\
=& f(\hat{x}^t)-\gamma\left\langle\nabla f(\hat{x}^t),\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t,\xi_i^t)\right\rangle\\
&+\frac{L\gamma^2}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t,\xi_i^t)\right\|^2. \qquad \text{(A23)}
\end{aligned}
$$

By applying expectation with respect to all the random variables at step t and conditional on the past (denote by $\mathbb{E}_{t|\cdot}$), we have

$$
\begin{aligned}
&\mathbb{E}_{t|\cdot}f(\hat{x}_{t+1})\\
\leqslant& f(\hat{x}^t)-\gamma\left\langle\nabla f(\hat{x}^t),\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\rangle\\
&+\frac{L\gamma^2}{2}\mathbb{E}_{t|\cdot}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t,\xi_i^t)\right\|^2\\
\leqslant& f(\hat{x}^t)-\gamma\left\langle\nabla f(\hat{x}^t),\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\rangle\\
&+\frac{L\gamma^2}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2+\frac{L\gamma^2\sigma^2}{2N}\\
=& f(\hat{x}^t)-\frac{\gamma}{2}\left(\left\|\nabla f(\hat{x}^t)\right\|^2+\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\right.\\
&\left.-\left\|\nabla f(\hat{x}^t)-\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\right)+\frac{L\gamma^2}{2}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
&+\frac{L\gamma^2\sigma^2}{2N}\\
\leqslant& f(\hat{x}^t)-\frac{\gamma}{2}\left\|\nabla f(\hat{x}^t)\right\|^2-\frac{\gamma}{2}(1-L\gamma)\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
&+\frac{\gamma L^2}{2N}\sum_{i=1}^{N}\left\|\hat{x}^t-x_i^t\right\|^2+\frac{L\gamma^2\sigma^2}{2N}, \qquad \text{(A24)}
\end{aligned}
$$

where the second inequality comes from Lemma 1, the last inequality follow from Cauchy's inequality and Lipschitz gradient assumption, respectively.

Rearranging this inequality and summing up both sides from $t=0$ to $T-1$, we have

$$
\begin{aligned}
&\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}^t)\|^2\\
\leqslant& f(\hat{x}_0)-f^*-\frac{\gamma}{2}(1-L\gamma)\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
&+\frac{\gamma L^2}{2N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\mathbb{E}\|\hat{x}^t-x_i^t\|^2+\frac{T\gamma^2L\sigma^2}{2N}. \qquad \text{(A25)}
\end{aligned}
$$

Substituting Lemma 5 into Eq. (A25) and denoting $C=\frac{1}{N}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\left\|\nabla f_i(\hat{x}^t)-\nabla f(\hat{x}^t)\right\|^2$, we obtain

$$
\begin{aligned}
&\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}^t)\|^2\\
\leqslant& f(\hat{x}_0)-f^*+\frac{T\gamma^2L\sigma^2}{2N}x+6k^2\gamma^3L^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2\\
&-\frac{\gamma}{2}(1-L\gamma-96k^4\gamma^4L^4)\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
&+\frac{11k\gamma^3\sigma^2L^2T}{3}+3k^3\gamma^3L^2C\\
\leqslant& f(\hat{x}_0)-f^*+\frac{T\gamma^2L\sigma^2}{2N}+\frac{\gamma}{6}\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2\\
&+\frac{11k\gamma^3\sigma^2L^2T}{3}+3k^3\gamma^3L^2C, \qquad \text{(A26)}
\end{aligned}
$$

where the last equality holds because $96k^4\gamma^4L^4+L\gamma\leqslant\frac{96}{36^2}+\frac{1}{2}\leqslant\frac{2}{27}+\frac{1}{2}\leqslant 1$ and $6k^2\gamma^3L^2\leqslant\frac{6\gamma}{36}\leqslant\frac{\gamma}{6}$.

Rearranging this inequality and dividing both sides by $\frac{T\gamma}{3}$ complete the proof.

Corollary 1 Under Assumption 1, when the learning rate is set as $\gamma=\frac{\sqrt{N}}{\sigma\sqrt{T}}$, and the communication period is set as $k=min\left\{\frac{\sigma\sqrt{T}}{6LN^{\frac{3}{2}}},\frac{\sqrt{T}}{\sqrt{N}}\right\}$, we have the following convergence result for Algorithm 1:

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}^t)\right\|^2\leqslant\frac{3\sigma(f(\hat{x}^0)-f^*+L)}{\sqrt{NT}}+\frac{C}{4\sqrt{NT}},
$$

where C is defined in Theorem 1.

Proof Since $\gamma=\frac{\sqrt{N}}{\sigma\sqrt{T}}$ and $k=min\left\{\frac{\sigma\sqrt{T}}{6LN^{\frac{3}{2}}},\frac{\sqrt{T}}{\sqrt{N}}\right\}$, we have $T\geqslant max\left\{\frac{36N^3L^2k^2}{\sigma^2},Nk^2\right\}$, $\gamma\leqslant\frac{1}{2L}$ and $k^2\gamma^2L^2=\frac{1}{36N^2}\leqslant\frac{1}{36}$. Then we can have the result in Theorem 1 and get

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}^t)\|^2\\
\leqslant&\frac{3(f(\hat{x}^0)-f^*)}{T\gamma}+\frac{3\gamma L\sigma^2}{2N}+11k\gamma^2\sigma^2L^2+\frac{9k^3\gamma^2L^2C}{T}.
\end{aligned}
$$

Combing $\gamma=\frac{\sqrt{N}}{\sigma\sqrt{T}}$, $T\geqslant\frac{36N^3L^2k^2}{\sigma^2}$, $k\leqslant\frac{\sqrt{T}}{\sqrt{N}}$ and $k^2\gamma^2L^2\leqslant\frac{1}{36}$ we have

$$
\begin{aligned}
11k\gamma^2\sigma^2L^2 &= 11k\frac{N}{\sigma^2T}\sigma^2L^2\leqslant\frac{11kNL^2}{\sqrt{T}}\frac{\sigma}{6N^{\frac{3}{2}}Lk}\leqslant\frac{2\sigma L}{\sqrt{NT}},\\
\frac{3\gamma L\sigma^2}{2N} &= \frac{3\sigma L}{2\sqrt{NT}},\\
\frac{3(f(\hat{x}_0)-f^*)}{T\gamma} &= \frac{3\sigma(f(\hat{x}_0)-f^*)}{\sqrt{NT}},\\
\frac{9k^3\gamma^2L^2C}{T} &\leqslant \frac{9k^2\gamma^2L^2C}{\sqrt{NT}}\leqslant\frac{C}{4\sqrt{NT}}.
\end{aligned}
$$

We can get the final result

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}^t)\right\|^2 \leqslant \frac{3\sigma(f(\hat{x}^0)-f^*+L)}{\sqrt{NT}}+\frac{C}{4\sqrt{NT}},$$

which completes the proof.

Appendix F: Proof of VRL-SGD with warm-up

Lemma 6 Under Assumption 1, when the learning rate γ and the communication period k satisfy that $\gamma \leqslant \frac{1}{2L}$ and $6k\gamma L \leqslant 1$, we have the following inequality for Algorithm 1:

$$\begin{aligned}
&\frac{1}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\|x_i^t-\hat{x}^t\|^2\\
\leqslant&12k^2\gamma^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2+96k^4\gamma^4L^2\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
&+\frac{22k\gamma^2\sigma^2T}{3}+6k^3\gamma^2C',
\end{aligned}\tag{A27}$$

where C' is defined as $C'=\sigma^2$.

Proof According to Lemma 2 and Lemma 4, we can get a similar result like Eq. (A20).

$$\begin{aligned}
&\frac{1}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\left\|\hat{x}^t-x_i^t\right\|^2\\
\leqslant&\frac{6\gamma^2L^2}{N}\sum_{t=k}^{T-1}\sum_{i=1}^{N}\left(k\sum_{\tau=t'}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2\right.\\
&\left.+2\sum_{\tau=t'}^{t-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^\tau-\hat{x}^{\tau'}\|^2+2k\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^{\tau'}-x_i^{\tau'}\|^2\right)\\
&+6k\gamma^2\sum_{t=k}^{T-1}\sum_{\tau=t'}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2+3k\gamma^2\sigma^2(T-k)\\
&+\frac{6k\gamma^2L^2}{N}\sum_{t=0}^{k-1}\sum_{i=1}^{N}\sum_{\tau=t'}^{t-1}\mathbb{E}\|x_i^\tau-\hat{x}^\tau\|^2\\
&+6k\gamma^2\sum_{t=0}^{k-1}\sum_{\tau=t'}^{t-1}\|\nabla f(\hat{x}^\tau)\|^2+3k^3\gamma^2\sigma^2\\
&+12k\gamma^2L^2\sum_{t=0}^{k-1}\sum_{\tau=t'}^{t-1}\mathbb{E}\left\|\hat{x}^\tau-\hat{x}^0\right\|^2\\
\leqslant&\frac{18\gamma^2k^2L^2}{N}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\mathbb{E}\|x_i^t-\hat{x}^t\|^2+6k^2\gamma^2\sum_{t=0}^{T-1}\|\nabla f(\hat{x}^t)\|^2\\
&+12k\gamma^2L^2\left(\underbrace{\sum_{t=k}^{T-1}\sum_{\tau'=t''}^{t'-1}\mathbb{E}\|\hat{x}^t-\hat{x}^{\tau'}\|^2}_{T_6}+\underbrace{\sum_{t=0}^{k-1}\sum_{\tau=0}^{t-1}\mathbb{E}\|\hat{x}^\tau-\hat{x}^0\|^2}_{T_7}\right)\\
&+3k\gamma^2\sigma^2T+3k^3\gamma^2C',
\end{aligned}\tag{A28}$$

Next, we bound T_6+T_7.

$$\begin{aligned}
&T_6+T_7\\
=&T_6+\sum_{t=0}^{k-1}\sum_{\tau=0}^{t-1}\mathbb{E}\left\|\hat{x}^\tau-\hat{x}^0\right\|^2\\
=&T_6+\sum_{t=0}^{k-1}\sum_{\tau=0}^{t-1}\mathbb{E}\left\|\sum_{s=\tau}^{t-1}\frac{\gamma}{N}\sum_{i=1}^{N}v_i^s\right\|^2\\
=&T_6+\frac{\gamma^2}{N^2}\sum_{t=0}^{k-1}\sum_{\tau=0}^{t-1}\mathbb{E}\left\|\sum_{s=\tau}^{t-1}\sum_{i=1}^{N}\nabla f_i(x_i^s,\xi_i^s)\right\|^2\\
\leqslant&T_6+\frac{2k^3\gamma^2\sigma^2}{N}+\sum_{t=0}^{k-1}\sum_{\tau=0}^{t-1}\mathbb{E}\left\|\frac{\gamma}{N}\sum_{s=\tau}^{t-1}\sum_{i=1}^{N}\nabla f_i(x_i^s)\right\|^2\\
\leqslant&T_6+\frac{2k^2\gamma^2\sigma^2(T-k)}{N}+k^3\gamma^2\sum_{t=0}^{k-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2\\
\overset{(33)}{\leqslant}&2k^2\gamma^2\sigma^2T+4k^3\gamma^2\sum_{t=0}^{T-1}\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^t)\right\|^2,
\end{aligned}\tag{A29}$$

where the first inequality comes from Lemma 1. The rest proofs are the same as Lemma 5.

Corollary 2 Warm-up If we set the first communication period k to 1 in Corollary 1, which is *VRL-SGD* with a warm-up (*VRL-SGD-W*), we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}^t)\right\|^2 \leqslant \frac{3\sigma(f(\hat{x}^0)-f^*+L)}{\sqrt{NT}}+\frac{\sigma^2}{4\sqrt{NT}}.$$

Proof Note that Lemma 5 and Lemma 6 have the same results except the constants C and C'. Therefore, Theorem 1 and Corollary 1 are also true, which proves the convergence of VRL-SGD with warm-up. And we can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\hat{x}^t)\|^2\\
\leqslant&\frac{3(f(\hat{x}^0)-f^*)}{T\gamma}+\frac{3\gamma L\sigma^2}{2N}+11k\gamma^2\sigma^2L^2+\frac{9k^3\gamma^2L^2C'}{T},
\end{aligned}\tag{A30}$$

where $C'=\sigma^2$.mCombing $\gamma=\frac{\sqrt{N}}{\sigma\sqrt{T}}$, $T\geqslant\frac{36N^3L^2k^2}{\sigma^2}$, $k\leqslant\frac{\sqrt{T}}{\sqrt{N}}$ and $k^2\gamma^2L^2\leqslant\frac{1}{36}$,
we can get the final result similar to Corollary 1:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\hat{x}^t)\right\|^2 \leqslant \frac{3\sigma(f(\hat{x}^0)-f^*+L)}{\sqrt{NT}}+\frac{\sigma^2}{4\sqrt{NT}},$$

which completes the proof.

References

1. Robbins H, Monro S. A stochastic approximation method. The Annals of Mathematical Statistics, 1951, 22(3): 400–407
2. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016, 770–778
3. Chen J, Liu D, Xu T, Wu S, Cheng Y, Chen E. Is heuristic sampling necessary in training deep object detectors? 2019, arXiv preprint arXiv: 1909.04868
4. Devlin J, Chang M W, Lee K, Toutanova K. BERT: pre-training of deep

bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019
5. Cheng H T, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anderson G, Corrado G, Chai W, Ispir M, Anil R, Haque Z, Hong L, Jain V, Liu X, Shah H. Wide & deep learning for recommender systems. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. 2016, 7–10
6. Liu Q, Huang Z, Yin Y, Chen E, Xiong H, Su Y, Hu G. EKT: exercise-aware knowledge tracing for student performance prediction. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(1): 100–115
7. Wu L, Li Z, Zhao H, Pan Z, Liu Q, Chen E. Estimating early fundraising performance of innovations via graph-based market environment model. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6396–6403
8. Liu Y, Liu Q, Zhao H, Pan Z, Liu C. Adaptive quantitative trading: an imitative deep reinforcement learning approach. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(2): 2128–2135
9. Wang J, Zhang H, Liu Q, Pan Z, Tao H. Crowdfunding dynamics tracking: a reinforcement learning approach. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(4): 6210–6218
10. Wang J, Joshi G. Cooperative SGD: a unified framework for the design and analysis of communication-efficient SGD algorithms. 2019, arXiv preprint arXiv: 1808.07576
11. Zhou F, Cong G. On the convergence properties of a K-step averaging stochastic gradient descent algorithm for nonconvex optimization. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018, 3219−3227
12. Stich S U. Local SGD converges fast and communicates little. In: Proceedings of 2019 International Conference on Learning Representations. 2019
13. Yu H, Yang S, Zhu S. Parallel restarted SGD with faster convergence and less communication: demystifying why model averaging works for deep learning. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 5693–5700
14. Shen S, Xu L, Liu J, Liang X, Cheng Y. Faster distributed deep net training: computation and communication decoupled stochastic gradient descent. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. 2019, 4582−4589
15. McMahan H B, Moore E, Ramage D, Hampson S, Arcas B A. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017, 1273−1282
16. Konečný J, McMahan H B, Yu F X, Suresh A T, Bacon D, Richtárik P. Federated learning: strategies for improving communication efficiency. In: Proceedings of the ICLR 2018. 2018
17. Li T, Sahu A K, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Processing Magazine, 2020, 37(3): 50–60
18. Kairouz P, McMahan H B, Avent B, Bellet A, Bennis M, et al. Advances and Open Problems in Federated Learning. Foundations and Trends ® in Machine Learning, 2021, 14(1−2): 1−210
19. Yang J, Duan Y, Qiao T, Zhou H, Wang J, Zhao W. Prototyping federated learning on edge computing systems. Frontiers of Computer Science, 2020, 14: 146318
20. Ghadimi S, Lan G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 2013, 23(4): 2341–2368
21. Dekel O, Gilad-Bachrach R, Shamir O, Xiao L. Optimal distributed online prediction using mini-batches. The Journal of Machine Learning Research, 2012, 13: 165–202
22. Alistarh D, Grubic D, Li J, Tomioka R, Vojnovic M. QSGD: communication-efficient SGD via gradient quantization and encoding. In: Proceedings of the Advances in Neural Information Processing Systems. 2017, 1709−1720
23. Aji A F, Heafield K. Sparse communication for distributed gradient descent. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017, 440–445
24. Bernstein J, Zhao J, Azizzadenesheli K, Anandkumar A. SignSGD with majority vote is communication efficient and fault tolerant. In: Proceedings of the International Conference on Learning Representations. 2019
25. Lin Y, Han S, Mao H, Wang Y, Dally W J. Deep gradient compression: reducing the communication bandwidth for distributed training. In: Proceedings of the International Conference on Learning Representations. 2018
26. Karimireddy S P, Rebjock Q, Stich S U, Jaggi M. Error feedback fixes SignSGD and other gradient compression schemes. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 3252−3261
27. Tang H, Yu C, Lian X, Zhang T, Liu J. DoubleSqueeze: parallel stochastic gradient descent with double-pass error-compensated compression. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 6155−6165
28. Povey D, Zhang X, Khudanpur S. Parallel training of DNNs with natural gradient and parameter averaging. 2015, arXiv preprint arXiv: 1410.7455
29. Su H, Chen H. Experiments on parallel training of deep neural network using model averaging. 2018, arXiv preprint arXiv: 1507.01239
30. Lin T, Stich S U, Patel K K, Jaggi M. Don't use large mini-batches, use local SGD. In: Proceedings of the 8th International Conference on Learning Representations. 2020
31. Yu H, Jin R, Yang S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 7184−7193
32. Haddadpour F, Kamani M M, Mahdavi M, Cadambe V. Trading redundancy for communication: Speeding up distributed SGD for non-convex optimization. In: Proceedings of the 36th International Conference on Machine Learning. 2019, 2545−2554
33. Li X, Huang K, Yang W, Wang S, Zhang Z. On the convergence of FedAvg on non-IID data. In: Proceedings of the 8th International Conference on Learning Representations. 2020
34. Khaled A, Mishchenko K, Richtarik P. Tighter theory for local SGD on identical and heterogeneous data. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. 2020, 4519−4529
35. Johnson R, Zhang T. Accelerating stochastic gradient descent using predictive variance reduction. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems. 2013, 315–323
36. Zhang L, Mahdavi M, Jin R. Linear convergence with condition number independent access of full gradients. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013, 980–988
37. Defazio A, Bach F, Lacoste-Julien S. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014, 1646−1654
38. Nguyen L M, Liu J, Scheinberg K, Takáč M. SARAH: a novel method for machine learning problems using stochastic recursive gradient. In: Proceedings of the 34th International Conference on Machine Learning. 2017, 2613−2621
39. Shi W, Ling Q, Wu G, Yin W. EXTRA: an exact first-order algorithm for decentralized consensus optimization. SIAM Journal on Optimization, 2015, 25(2): 944–966
40. Mokhtari A, Ribeiro A. DSA: decentralized double stochastic averaging gradient algorithm. The Journal of Machine Learning Research, 2016, 17(1): 2165–2199
41. Tang H, Lian X, Yan M, Zhang C, Liu J. D^2: decentralized training over

decentralized data. In: Proceedings of the 35th International Conference on Machine Learning. 2018, 4855−4863
42. Karimireddy S P, Kale S, Mohri M, Reddi S, Stich S, Suresh A T. SCAFFOLD: stochastic controlled averaging for federated learning. In: Proceedings of the 37th International Conference on Machine Learning. 2020, 5132−5143
43. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 2017
44. Zhang S, Choromanska A, LeCun Y. Deep learning with elastic averaging SGD. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2015, 685–693
45. El-Sawy A, El-Bakry H, Loey M. CNN for handwritten Arabic digits recognition based on LeNet-5. In: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics. 2016, 566–575
46. LeCun Y. The MNIST database of handwritten digits. See Yann.lecun.com/exdb/mnist/ website, 1998
47. Kim Y. Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014
48. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes P N, Hellmann S, Morsey M, van Kleef P, Auer S, Bizer C. DBpedia−A large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web, 2015, 6(2): 167–195
49. Deng J, Dong W, Socher R, Li L J, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009, 248–255
50. Moschitti A, Pang B, Daelemans W. Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014, 1532−1543
51. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016, 2818−2826
52. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning. 2015, 448–456

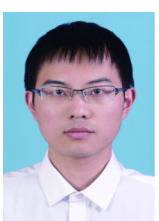
Xianfeng Liang is currently a MS student in the School of Computer Science and Technology at the University of Science and Technology of China (USTC), China. His major research interests include machine learning and optimization.

Shuheng Shen received the MS degree from University of Science and Technology of China (USTC), China in 2020. His major research interests include machine learning, stochastic optimization and distributed optimization.

Enhong Chen is a professor and vice dean of the School of Computer Science at University of Science and Technology of China (USTC), China. He received the PhD degree from USTC, China. His general area of research includes data mining and machine learning, social network analysis and recommender systems. He has published more than 100 papers in refereed conferences and journals, including IEEE Trans. KDE, IEEE Trans. MC, KDD, ICDM, NIPS, and CIKM. He was on program committees of numerous conferences including KDD, ICDM, SDM. He is a senior member of the IEEE.

Jingchang Liu received the MS degree from University of Science and Technology of China (USTC), China in 2019. His major research interests include machine learning, stochastic optimization and distributed optimization.

Qi Liu is a professor at University of Science and Technology of China (USTC), China. He received the PhD degree in computer science from USTC, China. His general area of research is data mining and knowledge discovery. He has published prolifically in refereed journals and conference proceedings, e.g., TKDE, TOIS, TKDD, TIST, KDD, IJCAI, AAAI, ICDM, SDM and CIKM. He has served regularly in the program committees of a number of conferences, and is a reviewer for the leading academic journals in his fields. He is a member of ACM and IEEE. Dr. Liu is the recipient of the KDD 2018 Best Student Paper Award (Research) and the ICDM 2011 Best Research Paper Award. He is supported by the Young Elite Scientist Sponsorship Program of CAST and the Youth Innovation Promotion Association of CAS.



Yifei Cheng is currently working toward the PhD degree in the School of Data Science, University of Science and Technology of China, China. His current research interests include machine learning, distributed optimization and federated learning.

Zhen Pan received the PhD degree from University of Science and Technology of China, China in 2020. His major research interests include machine learning and data mining.