Communication-efficient local SGD with age-based worker selection

Feng Zhu · Jingjing Zhang · Xin Wang

Accepted: 12 March 2023 / Published online: 30 March 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

A major bottleneck of distributed learning under parameter server (PS) framework is communication cost due to frequent bidirectional transmissions between the PS and workers. To address this issue, local stochastic gradient descent (SGD) and worker selection have been exploited by reducing the communication frequency and the number of participating workers at each round, respectively. However, partial participation can be detrimental to convergence rate, especially for heterogeneous local datasets. In this paper, to improve communication efficiency and speed up the training process, we develop a novel worker selection strategy named AgeSel. The key enabler of AgeSel is utilization of the ages of workers to balance their participation frequencies. The convergence of local SGD with the proposed age-based partial worker participation is rigorously established. Simulation results demonstrate that the proposed AgeSel strategy can significantly reduce the number of training rounds needed to achieve a targeted accuracy, as well as the communication cost. The influence of the algorithm hyper-parameter is also explored to manifest the benefit of age-based worker selection.

Keywords  Local SGD · Communication efficiency · Age-based worker selection · Distributed learning
1 Introduction

As machine learning tasks nowadays are becoming increasingly challenging and the sizes of datasets are multiplying at an astonishing speed, distributed machine learning has become the main workhorse nowadays to relieve the burden of the central server and to achieve speedup in the training process [1, 2]. It has to be pointed out that distributed learning is a promising and enabling technology in various fields such as edge computing [3], fog computing [4], Internet of Things [5] and sensor networks [6].

Parameter server (PS) setting is one of the most popular paradigms in distributed machine learning. In this setting, the PS broadcasts the current global model parameter to the workers for gradient computation and aggregates the computed gradients to update the global model. The operation repeats until some targeted convergence criterion is reached [1, 2, 7–10]. However, the massive communication overhead between the PS and the workers has become the main bottleneck of the overall system performance as the sizes of the neural networks grow exponentially enormous [11].

In dealing with the communication issue, the idea of local SGD is proposed. To be specific, each worker is allowed to perform local iterations to update its local model and the communication frequency between the PS and the workers is effectively reduced, leading to the improvement in communication efficiency. One of the pioneering works is FedAvg [12], based on which a number of researches are developed to further improve the performance of local SGD and the communication efficiency of the system.

Besides performing local updates, instead of random selection as in FedAvg, adaptive worker selection technique has attracted increasing attention recently [13, 14]. Via adaptive selection in each training round instead of selecting all the workers, the per-round communication load is reduced, while the performance as full worker participation is also preserved, as a result of which the system communication efficiency is further improved.

However, most of the worker selection methods require information such as gradient norm [13, 14] and local loss [15] that entails high computation complexity. In contrast, in this work we simply count on the ages of each worker, i.e., the number of consecutive rounds the worker has not communicated with the PS, as the defining metric and develop the novel AgeSel worker selection strategy with local SGD to further improve the communication efficiency of the system. The contributions of this work are summarized as follows.

Our Contributions In this paper, we propose a novel age-based worker selection strategy for local SGD. In what we call AgeSel scheme, a simple age-based mechanism is implemented such that the workers are forced to be selected if they have not been involved for a certain number of consecutive rounds. Different from existing adaptive worker selection strategies, the proposed AgeSel relies on the age information that is readily available at the PS without additional communication and computation overhead. The convergence of AgeSel is also rigorously established to justify the benefit of age-based partial worker participation.
The simulation results corroborate the superiority of AgeSel in terms of communication efficiency and required number of training rounds (to achieve a targeted accuracy) over state-of-the-art schemes. The impact of the algorithm hyperparameter is also explored through experiments.

2 Related works

In this section, we comprehensively review previous researches that are most related to our work, i.e., local SGD variants, worker selection strategies and the age-based mechanism.

Local SGD The idea of local SGD is first introduced in [16]. To be specific, all the workers perform a certain number of local updates, and the PS aggregates the latest models of the workers as the final output. However, the authors of [17] prove that the method proposed in [16] is no better than the single-worker algorithm under the worst-case scenario because of the divergence of local optima across workers. To iron this issue out, [12] proposes a widely studied algorithm, named federated averaging (FedAvg), where the PS randomly selects a subset of workers and sends the global model to the selected workers for a certain number of local updates at each round. The PS then collects the latest local models to update the global model and re-sends it to the re-selected workers for local updates again. By increasing the communication frequency compared to [16] and decreasing the number of participating workers, FedAvg can achieve improved communication efficiency, with its convergence analysis extensively studied. Under homogenous data distribution, FedAvg is proved to achieve sublinear convergence rate and a linear speedup concerning the number of participating workers [18–20]. Its convergence under heterogeneous data distribution is also explored in [21–27]. [28] succeeds to establish a linear speedup with respect to the number of workers with non-convex objective functions. Additionally, [29] proposes a free local SGD method with parallel synchronization to reduce the time and communication load.

Worker Selection Instead of random selection as in FedAvg, adaptive worker selection has become another feasible technique in improving communication efficiency. One of the representative works on adaptive selection in SGD-based distributed learning is communication-adaptive distributed Adam (CADA) proposed in [13], in which a worker is required to upload its gradient only if its difference (i.e., the change to its former local model) is large enough or its local gradient has become overly stale. The asserted optimal client selection strategy (OCS) is further developed in [14] by selecting the workers with larger gradient norms to minimize the variance of the global update. In addition, another metric of “contribution,” i.e., local loss of workers, is also explored as a criterion for worker selection design [15, 30–33].

Age-Based Mechanism On the other hand, imbalanced participation could render the training unstable or slow, especially under heterogeneous data distribution, since the influence of some parts of the data on the overall training process is weakened [34, 35]. Hence, the age of each worker, indicating the number of consecutive rounds where it has not participated in the computation, could be taken into consideration
by adaptive selection techniques. To this end, [36] explores the influence of ages in gradient descent (GD) scenario. For distributed learning over wireless connections, [37] jointly leverages the channel quality and the ages of workers to improve the communication efficiency. Additionally, if with perfect lossless channel, the scheme reduces to the Round Robin (RR) policy [38]. Moreover, [39] finds the age-optimal number of workers to select at each round, where the age is defined as the sum of computation time and uplink transmission time.

The rest of the paper is organized as follows. Section 3 delineates the system model and performance metrics of interest. The development of the proposed Age-Sel algorithm along with its convergence analysis is described in Sect. 4. Section 5 provides numerical results, followed by the conclusions and appendices.

Notations $\mathbb{R}$ denote the real number fields; $\mathbb{E}$ denotes the expectation operator; $\| \cdot \|$ denotes the $\ell_2$ norm; $\nabla F$ denotes the gradient of function $F$; $\bigcup$ denotes the union of sets; $A \subseteq B$ represents that set $A$ is a subset of set $B$; and $|A|$ denotes the size of set $A$.

### 3 Problem formulation

#### 3.1 System model

Consider the PS-based framework of distributed learning with heterogeneous data distribution. There are $M$ distributed workers in set $\mathcal{M} \triangleq \{1, \ldots, M\}$. Each worker $m$ maintains a local dataset $\mathcal{D}_m$ of size $N_m$. It is drawn from the global dataset $\mathcal{D} = \{z_i\}_{i=1}^N$, i.e., we have $\mathcal{D} = \bigcup_{m \in \mathcal{M}} \mathcal{D}_m$. Our objective is to minimize the weighted loss function

$$\mathcal{L}(\theta) = \sum_{m=1}^M \frac{N_m}{N} \mathcal{L}_m(\theta),$$

(1)

where $\theta \in \mathbb{R}^d$ is the $d$-dimension parameter to be optimized, and we have the definition $\mathcal{L}_m(\theta) \triangleq \mathbb{E}[\mathcal{L}_m(\theta ; z_m)]$, with $\mathcal{L}_m(\theta)$ being the local loss function of worker $m$ and $z_m$ being the sample drawn randomly from its local dataset $\mathcal{D}_m$.

To elaborate on local SGD, we define $\theta^i$ as the global model parameter at training round $j$, and $\theta^{i,0}_m$ as the local model of worker $m$ before operating local updates. At each training round $j$, a subset $\mathcal{M}_D^j \subseteq \mathcal{M}$ of workers are selected to download the global model $\theta^j$ from the PS prior to computation. After that, each selected worker $m$ in $\mathcal{M}_D^j$ sets its local model as $\theta^{j,0}_m = \theta^j$ and starts operating $U$ local iterations, with the updating formula given as

$$\theta^{i,u+1}_m = \theta^{i,u}_m - \frac{\eta}{B} \sum_{b=1}^B \nabla \mathcal{L}(\theta^{i,u}_m ; z^{i,u}_{m,b}),$$

(2)

for any local iteration $u = 0, \ldots, U - 1$. Here $\eta$ is the stepsize and $\frac{1}{B} \sum_{b=1}^B \nabla \mathcal{L}(\theta^{i,u}_m ; z^{i,u}_{m,b})$ is the minibatch gradient to be computed by worker $m$ at local iteration $u$, where $B$
is the minibatch size and \(z_{m,b}^j\) is a sample drawn independently across iterations from the local dataset \(D_m\).

After all the workers in set \(M_D^j\) have completed their local computations, a subset of \(S\) workers \(M_U^j \subseteq M_D^j\) are selected to upload their latest models, and the PS aggregates the models to update the global model

\[
\theta^{j+1} = \frac{1}{S} \sum_{m \in M_U^j} \theta_{m}^{j,U}.
\]

Note that here we employ the unbiased aggregation which is also adopted in [14], since the weight of each worker is reflected in the worker selection process, as will be specified later.

The training process ends when some stopping criterion is satisfied, with the total number of rounds denoted as \(J\).

### 3.2 Preliminaries

Many local SGD variants utilizing worker selection from previous works can be established under the aforementioned PS framework.

**Federated Averaging (FedAvg)** We first introduce the so-called FedAvg algorithm [12]. FedAvg applies a random worker sampling strategy at the PS side. At each training round, each randomly selected worker computes \(U\) iterations of local updates and then sends the local parameter back to the PS, i.e., with \(M_U^j = M_D^j\).

**Optimal Client Sampling (OCS)** In contrast to FedAvg with random selection, [14] proposes Optimal Client Sampling (OCS) relying on gradient norms to make the selection decision. The idea is to choose the workers with larger gradient norms, such that the variance of the global update is reduced, yielding a faster convergence rate than FedAvg. However, the selection requires additional communication cost. More precisely, in the first iteration of each round, all the workers download the global model from the PS and compute their gradient norms. By choosing the largest ones, the PS decides subset \(M_U^j\) to upload their local models.

**Round Robin (RR)** The Round Robin (RR) policy [38] selects the workers in a circular order with \(S\) workers being selected at each round. With RR we also have \(M_U^j = M_D^j\).

### 3.3 Performance Metrics

To gauge the efficiency of the proposed scheme, we are interested in two performance metrics, i.e., the number of training rounds and the communication cost required to reach a targeted training accuracy.
Firstly, the number of training rounds $J$ of the algorithm to achieve a targeted test accuracy is used to reflect the training speed, as well as the communication cost\(^1\) of the algorithm.

Secondly, we define the communication cost $C^j$ at training round $j$ as the number of communication rounds between the PS and the workers, given as

$$C^j = |\mathcal{M}_D^j| + S. \quad \text{(4)}$$

This is due to the fact that the size of the global parameter downloaded by the workers is the same as that of the latest parameter uploaded by each selected worker. As a result, the total communication cost is given as $C = \sum_{j=0}^{J-1} C^j$.

These two metrics will be used in Sect. 5 to compare the performance of different schemes.

4 Adaptive selection in local SGD (AgeSel)

In this section, we first develop the novel AgeSel scheme that aims at improving communication efficiency by utilizing the age-based worker selection. Then the convergence of AgeSel is rigorously established.

As mentioned in Sect. 1, most of the counterparts in worker selection depend heavily on information that requires computing gradient norm and local loss, etc. However, the computation of these information consumes additional time and might slow down the training process. As a result, we introduce the proposed AgeSel strategy that merely depends on the ages and the sizes of local datasets of the workers, which hardly need complex computation. The simulation results in Sect. 5 show that the proposed AgeSel can achieve even better performance than state-of-the-art schemes in terms of number of training rounds and communication efficiency.

4.1 Algorithm description

To elaborate on worker selection strategy in the proposed AgeSel, we define an $M$-length vector $\tau_M$ to collect the ages of the workers, as in [13, 14]. Note that $\tau_M$ is maintained by the PS and initialized to be a zero vector. Each element $\tau_m$ of vector $\tau_M$ measures the number of consecutive rounds that worker $m$ has not been selected by the PS. Particularly, at each round $j$, $\tau_m$ is updated as follows:

$$\tau_m = \begin{cases} 0, & \text{if } m \text{ is selected}, \\ \tau_m + 1, & \text{otherwise}. \end{cases} \quad \text{(5)}$$

To identify the workers with low participation frequency, we pre-define a threshold $\tau_{\text{max}}$. Accordingly, worker $m$ would be forced to participate when having its $\tau_m \geq \tau_{\text{max}}$. The specific procedure of AgeSel is delineated next.

---

\(^1\) Since more training rounds also indicate more rounds of communication between the PS and the workers.
At each round \( j \), the PS selects \( S \) workers to perform computation by checking the vector \( \tau_M \). More precisely, with vector \( \tau_M \), we can identify all the infrequent workers with their ages greater than \( \tau_{\text{max}} \), which would be considered first. Let \( S' \) denote the number of these workers. The set \( \mathcal{M}_D^j \) of selected workers can be determined as follows. For the first case with \( S' \geq S \), the PS simply picks the \( S \) workers in an age-descending order. Note that the workers with larger sizes of local datasets are prioritized when there are ties in ages. For the second case with \( S' < S \), the PS first picks all the \( S' \) infrequent workers and then chooses the rest \( S - S' \) without replacement from the workers having ages smaller than \( \tau_{\text{max}} \) with the probabilities proportional to the sizes of their datasets.

Once the set \( \mathcal{M}_D^j \) is determined, the PS then broadcasts the global parameter \( \theta^j \) to all the selected workers. By initiating its local model with \( \theta_{m,0}^j = \theta^j \), each worker \( m \) in \( \mathcal{M}_D^j \) then starts \( U \) iterations of local updates through (2) and sends its latest model \( \theta_{m}^{j,U} \) to the PS, i.e., we have \( \mathcal{M}_U^j = \mathcal{M}_D^j \). Eventually, the PS updates the vector \( \tau_M \), along with the global model aggregated via (3).

**Algorithm 1 AgeSel**

**Require:** number of workers \( M \), number of workers to select \( S \), age vector \( \{\tau_m\}_{m=1}^M \), stepsize \( \alpha > 0 \), batch size \( B \), age threshold \( \tau_{\text{max}} \geq 1 \)

1. **Initiate** \( \theta^0 \), \( \{\tau_m\}_{m=1}^M = 0 \)

2. **for** \( j = 0, 1, \ldots, J - 1 \) **do**

3. **PS checks** the age vector \( \{\tau_m\}_{m=1}^M \) and determines the \( S' \) workers with \( \tau_m \geq \tau_{\text{max}} \)

4. **if** \( S' \geq S \) **then**

5. **PS adds** the \( S \) workers in an age-descending order into the set \( \mathcal{M}_D^j \)

6. **else**

7. **PS first adds** the \( S' \) workers with \( \tau_m \geq \tau_{\text{max}} \) into the set \( \mathcal{M}_D^j \)

8. **PS randomly picks** the rest \( S - S' \) without replacement from the rest of the workers with the probabilities proportional to the sizes of their datasets

9. **end if**

10. **PS sets** \( \tau_m = 0 \) for \( m \in \mathcal{M}_D^j \) and sets \( \tau_m = \tau_m + 1 \) for \( m \in \mathcal{M}_D \setminus \mathcal{M}_D^j \)

11. **PS sends** the global model \( \theta^j \) to all the workers in \( \mathcal{M}_D^j \)

12. **for** \( m \in \mathcal{M}_D^j \) **do**

13. **Worker** \( m \) sets \( \theta_{m,0}^j = \theta^j \) and performs \( U \) local updates via (2)

14. **Worker** \( m \) sends the local model \( \theta_{m}^{j,U} \) back to the PS

15. **end for**

16. **PS updates** the global parameter via (3)

17. **end for**

Note that when \( \tau_{\text{max}} \) is very large, we barely have infrequent workers. Then the set \( \mathcal{M}_D^j \) is indeed chosen by weights (i.e., the sizes of datasets), as with the FedAvg algorithm [12]; i.e., in this case, AgeSel is reduced to FedAvg.
Merits of AgeSel: The benefit from the age-based mechanism used in AgeSel is twofold. First, it has been shown that less participation of some workers can be detrimental to the convergence rate due to the lack of gradient diversity [34], especially for heterogeneous local datasets. Our age-based selection strategy can balance the worker participation, thereby preserving gradient diversity and ensuring the fast convergence. Second, generation of age information \( \tau_M \) here incurs no extra communication and computation cost, in contrast to other information such as the (costly) norm of updates [14], used in existing alternatives. Hence, AgeSel is more communication- and computation-efficient.

4.2 Convergence analysis

We next establish the convergence of the proposed AgeSel algorithm under heterogeneous data distribution, with a general (not necessarily convex) objective function. Our analysis is based on the following two assumptions, which are widely adopted in related works such as [21, 28].

**Assumption 1** (Smoothness and Lower Boundedness) Each local function \( \mathcal{L}_m(\theta) \) is \( L \)-smooth, i.e.,

\[
\| \nabla \mathcal{L}_m(\theta_1) - \nabla \mathcal{L}_m(\theta_2) \| \leq L \| \theta_1 - \theta_2 \|
\]

\( \forall \theta_1, \theta_2 \in \mathbb{R}^d \). We also assume that the objective function \( \mathcal{L} \) is bounded below by \( \mathcal{L}^* \).

**Assumption 2** (Unbiasedness and Bounded Variance) For the given model parameter \( \theta \), the local gradient estimator is unbiased, i.e.,

\[
\mathbb{E}[\nabla \mathcal{L}_m(\theta; z)] = \nabla \mathcal{L}_m(\theta).
\]

Moreover, both the variance of the local gradient estimator and the variance of the local gradient from the global one are bounded, i.e., there exist two constants \( \sigma_L, \sigma_G > 0 \), such that

\[
\mathbb{E}[\| \nabla \mathcal{L}_m(\theta; z) - \nabla \mathcal{L}_m(\theta) \|^2] \leq \sigma_L^2, \forall m
\]

\[
\mathbb{E}[\| \nabla \mathcal{L}_m(\theta) - \nabla \mathcal{L}(\theta) \|^2] \leq \sigma_G^2, \forall m.
\]

With these assumptions, we can derive an upper bound for the expectation of the average squared gradient norm \( \frac{1}{J} \mathbb{E}\left[ \sum_{j=0}^{J-1} \| \nabla \mathcal{L}(\theta_j) \|^2 \right] \), to prove the convergence of the proposed AgeSel. We start by presenting the following lemma.

**Lemma 1** Under Assumptions 1 and 2 and \( \eta \) is chosen such that \( \eta \leq \frac{1}{8LU} \), there exists a positive constant \( c < \frac{1}{2} - 15U^2\eta^2L^2 - Ln(90U^3L^2\eta^2 + 3U) \), such that
\[ \mathbb{E}[\mathcal{L}(\Theta^{t+1})] \leq \mathcal{L}(\Theta^t) - c\eta U \left\| \nabla \mathcal{L}(\Theta^t) \right\|^2 + \frac{\eta^2 UL}{2SB} \sigma_L^2 \]
\[ + Z_1 + L\eta^2 Z_2 + \frac{2\eta L^2 (A^t)^2}{S^2} Z_2 - \frac{2\eta^2 A^t}{S} Z_2, \]

where we have defined \( Z_1 = \frac{5U\eta^2 U}{2} (\sigma_L^2 + 6U\sigma_U^2) \), \( Z_2 = 15U^3 L^2 \eta^2 (\sigma_L^2 + 6U\sigma_U^2) + 3U^2 \sigma_U^2 \), and \( A^t \) is the cardinality of the set \( A^t \) of workers selected by age at round \( j \), i.e., \( A \)= \( \min \{S, S'\} \).

**Proof** The proof can be found in Appendix A.

Lemma 1 depicts the one-step difference of the objective function, from which we can see the impact of the age. Particularly, as \( A^t \) grows larger, the variance term \(- \frac{2\eta L^2}{S} A^t \sigma_L^2 \) is reduced, while the variance term \( \frac{2\eta L^2 (A^t)^2}{S^2} Z_1 \) is increased. Therefore, the values of \( \tau_{\text{max}} \) and \( S \), deciding \( A^t \) jointly, have a significant impact on training speed, that is, the value of \( A^t \) should neither be too small nor too large to achieve a balance between age and dataset size, which will be further demonstrated in Sect. 5. Based on Lemma 1, we are able to arrive at the final convergence result.

**Theorem 1** With the same conditions as in Lemma 1, we have

\[ \frac{1}{J} \mathbb{E} \left[ \sum_{j=0}^{J-1} \left\| \nabla \mathcal{L}(\Theta^j) \right\|^2 \right] \leq \frac{\mathcal{L}(\Theta^0) - \mathcal{L}^*}{c\eta UJ} + V, \]

where we have defined the constant \( V = \frac{1}{c} \left[ \frac{UL}{2SB} \sigma_L^2 + \frac{Z_1}{\eta U} + (3\eta L - \frac{2\eta M}{SR}) \frac{Z_2}{U} \right] \), and \( R \) denotes the maximum number of rounds to traverse all the workers in \( M \) due to the age-based mechanism.

**Proof** The proof can be found in Appendix B.

Theorem 1 states that the proposed AgeSel scheme can achieve a sublinear convergence rate \( O\left(\frac{1}{J}\right) \) as local SGD with partial worker participation [21, 28]. The advantage of age-based worker selection is reflected in the expression of \( V \), i.e., a smaller \( \tau_{\text{max}} \) leads to a smaller \( R \), and then implies a reduced \( V \).

## 5 Simulation results

In this section, we evaluate the effectiveness of the proposed AgeSel against state-of-the-art schemes including FedAvg [12], OCS [14] and RR [38]. Note that AgeSel with \( \tau_{\text{max}} \) being large reduces to FedAvg. Furthermore, when \( \tau_{\text{max}} \) goes to zero, it is obvious that AgeSel is equivalent to RR. Since both \( \tau_{\text{max}} \) and \( S \) determine the value of \( A^t \), we will also explore the impact of the hyper-parameter \( S \) on the performance of AgeSel.
**Metric Specification** We start by measuring the per-round metrics for FedAvg, OCS, RR and AgeSel.

The communication cost of FedAvg $C_{\text{FedAvg}}^j$ at round $j$ can be expressed as:

$$C_{\text{FedAvg}}^j = 2|\mathcal{M}_D^j| = 2S,$$

(12)

since $S$ workers are randomly selected at each round to carry out the computation, i.e., we have $\mathcal{M}_D^j = \mathcal{M}_U^j$.

The communication cost of OCS is given as

$$C_{\text{OCS}}^j = M + S,$$

(13)

since all the workers in $\mathcal{M}$ download the global parameter from the PS and $S$ workers are finally selected to upload their local models.

The communication cost of RR $C_{\text{RR}}^j$ at round $j$ can be expressed as:

$$C_{\text{RR}}^j = 2|\mathcal{M}_D^j| = 2S,$$

(14)

since all the workers are selected in a circular order and at each round $S$ workers are selected to carry out the computation, i.e., we have $\mathcal{M}_D^j = \mathcal{M}_U^j$. Also, the aggregation of the updates is weighted for fair comparison.

For the proposed AgeSel scheme, the per-round communication cost is

$$C_{\text{AgeSel}}^j = 2|\mathcal{M}_D^j| = 2S,$$

(15)

since in AgeSel, $S$ workers are selected in each round.

**Simulation Setting** The dataset $\mathcal{D}$ considered here is the EMNIST dataset. We aim to solve the image classification task with a two-layer fully connected neural network, with the number of hidden units being 500. There are $M = 20$ workers in total, and the data are heterogeneously distributed among them. Particularly, the samples of the dataset $\mathcal{D}$ are sorted according to their labels and allocated to each worker in order with different sizes. The stepsize $\eta$ is 0.1; the batchsize $B$ is 100 and the number of local updates performed per round $U$ is set to be 5. All the schemes stop training once the test accuracy reaches 80%. For comparison, the number $S$ of workers selected to upload their local parameters at each round in all three schemes is designated to be $S = 5$. Moreover, we set $\tau_{\text{max}} = 4$ for AgeSel.

**AgeSel Outperforms the State-of-Art Schemes in Both Performance Metrics**

The performances of FedAvg, OCS, RR and AgeSel in terms of training rounds and communication cost with 10 Monte Carlo runs are depicted in Figs. 1 and 2, respectively.

As shown in Fig. 1, by either considering the weights (i.e., the sizes of datasets) or the ages only, FedAvg and RR require more training rounds than OCS and AgeSel in the presence of data heterogeneity. With larger norms for worker selection, OCS can converge faster. Clearly, AgeSel performs the best by achieving a desirable balance between ages and weights.
As illustrated in Fig. 2, OCS has the largest communication cost because all the $M$ workers download the global model, while $S$ of them is eventually selected. As a result, the reduced training rounds cannot offset the increase of the per-round communication overhead, yielding a larger total cost than the other schemes. Since AgeSel requires fewer training rounds than FedAvg and RR, and they all have the same per-round communication cost, AgeSel achieves better communication efficiency than both FedAvg and RR. Overall, AgeSel is the most communication-efficient selection strategy among all these schemes.

Exploration of $S$ As illustrated in Fig. 3, when $S$ is larger than some value (e.g., $S = 5$), the performance of AgeSel in terms of training round with partial participation

\footnote{Here we do not perform Monte Carlo runs to illustrate the effect of $S$ on the fluctuation of the curves.}
becomes quite close to that of full participation (i.e., $S = M = 20$). This implies that the age can indeed accelerate the training process. However, as $S$ increases, the per-round communication cost also monotonically increases, as shown in Fig. 4. From Figs. 1, 2, 3 and 4, we can see that, with proper values of $S$ and $\tau_{\text{max}}$, AgeSel can strike a desirable balance in terms of convergence speed and communication overhead. More simulation results are provided in Appendix C.

6 Conclusions

We developed a novel AgeSel strategy to perform adaptive worker selection for PS-based local SGD under heterogeneous data distribution. The convergence of AgeSel scheme was rigorously proved. Simulation results showed that AgeSel is more communication-efficient and converges faster than state-of-the-art schemes. More
importantly, unlike most worker selection strategies that require gradient or loss information that needs massive computing, AgeSel only considers the age and the sizes of the datasets to achieve better performance. These merits of AgeSel add up to a good worker selection algorithm.

Note that AgeSel is compatible with other techniques to jointly improve the overall performance of distributed learning. To name a few, it can be combined with straggler-tolerant techniques to render the learning robust to stragglers; it can also be used together with variance-reduction algorithms to further accelerate the training process. These interesting directions will be pursued in future work.

Appendix A: Proof of Lemma 1

With the $L$-smoothness of the objective function $\mathcal{L}$ in Assumption 1, by taking expectation of all the randomness over $\mathcal{L}(\theta^{t+1})$ we have:

$$
\mathbb{E}[\mathcal{L}(\theta^{t+1})] \leq \mathcal{L}(\theta^t) + \left\langle \nabla \mathcal{L}(\theta^t), \mathbb{E}[\theta^{t+1} - \theta^t] \right\rangle + \frac{L}{2} \mathbb{E}[\|\theta^{t+1} - \theta^t\|^2]
$$

$$(a1) \quad \mathcal{L}(\theta^t) + \left\langle \nabla \mathcal{L}(\theta^t), \mathbb{E}[g^j + \eta U \nabla \mathcal{L}(\theta^t)] \right\rangle + \frac{L}{2} \mathbb{E}[\|g^j\|^2]$$

$$(b1) \quad \mathcal{L}(\theta^t) - \eta U \|\nabla \mathcal{L}(\theta^t)\|^2 + \frac{\|\nabla \mathcal{L}(\theta^t)\|^2}{T_1} + \frac{\mathbb{E}[\|g^j\|^2]}{T_2}$$

where $(a1)$ is due to the definitions $g^j = \frac{1}{S} \sum_{m \in \mathcal{M}_j} g^j_m$ where $p_m = N_m / N$ and $g^j_m = -\eta \sum_{u=0}^{u-1} \frac{1}{B} \sum_{j=1}^{j_u} \nabla \mathcal{L}(\theta^{u-1}_{j_u})$; and $(b1)$ can be obtained directly from $(a1)$.

We then bound the terms $T_1$ and $T_2$, respectively. For the term $T_1$, with the definition of the weighted global update $\tilde{g}^j = \sum_{m \in \mathcal{M}} p_m g^j_m$ and the unbiased global update $\bar{g}^j = \frac{1}{M} \sum_{m \in \mathcal{M}} g^j_m$, we have

$$T_1 = \left\langle \nabla \mathcal{L}(\theta^t), \mathbb{E}[g^j + \eta U \nabla \mathcal{L}(\theta^t)] \right\rangle$$

$$(a2) \quad \left\langle \nabla \mathcal{L}(\theta^t), \mathbb{E}\left[ \frac{1}{S} \left( \sum_{m \in \mathcal{M}_j \setminus \mathcal{A}_j} g^j_m + \sum_{m \in \mathcal{A}_j} g^j_m \right) + \frac{1}{S} \left( (S - A) \eta U \nabla \mathcal{L}(\theta^t) + A \eta U \nabla \mathcal{L}(\theta^t) \right) \right] \right\rangle$$

$$(b2) \quad \left\langle \nabla \mathcal{L}(\theta^t), \mathbb{E}\left[ \frac{S - A}{S} \left( \tilde{g}^j + \eta U \nabla \mathcal{L}(\theta^t) \right) \right] \right\rangle + \left\langle \nabla \mathcal{L}(\theta^t), \mathbb{E}\left[ \frac{A_j}{S} \left( \tilde{g}^j + \eta U \nabla \mathcal{L}(\theta^t) \right) \right] \right\rangle$$

where $(a2)$ splits the selected workers at round $j$ into the ones selected by ages in set $\mathcal{A}_j$ with $|\mathcal{A}_j| = A_j = \min\{S, S^j\}$ and the ones selected by weights in set $\mathcal{M}_j \setminus \mathcal{A}_j$; $(b2)$
is because the selection by ages is essentially unweighted random selection in expectation and the selection by weights is equivalent to \( \tilde{g}^i \) in expectation. The two terms in (A2) are then bounded separately.

To bound the first term, we have:

\[
\left\langle \nabla \mathcal{L}(\theta^i), \mathbb{E} \left[ \frac{S - A_i}{S} \left( \tilde{g}^i + \eta U \nabla \mathcal{L}(\theta^i) \right) \right] \right\rangle \\
\overset{(a3)}{=} \frac{S - A_i}{S} \left\langle \nabla \mathcal{L}(\theta^i), \mathbb{E} \left[ -\frac{1}{B} \sum_{m=1}^{M} \sum_{u=0}^{U-1} \sum_{b=1}^{B} \eta p_m \nabla \mathcal{L}_m(\theta^{ja,m} \ast \eta, \ldots, \eta) + \eta U \sum_{m=1}^{M} \frac{p_m}{u} \nabla \mathcal{L}(\theta^i) \right] \right\rangle \\
\overset{(b3)}{=} \frac{S - A_i}{S} \left\langle \nabla \mathcal{L}(\theta^i), \mathbb{E} \left[ -\sum_{m=1}^{M} \sum_{u=0}^{U-1} \eta p_m \nabla \mathcal{L}_m(\theta^{ja,m}) + \eta U \sum_{m=1}^{M} \frac{p_m}{u} \nabla \mathcal{L}(\theta^i) \right] \right\rangle \\
\overset{(c3)}{=} \frac{S - A_i}{S} \left\langle \sqrt{\eta U} \nabla \mathcal{L}(\theta^i), -\sqrt{\eta U} \mathbb{E} \left[ \sum_{m=1}^{M} \sum_{u=0}^{U-1} \eta p_m \left( \nabla \mathcal{L}_m(\theta^{ja,m}) - \nabla \mathcal{L}(\theta^i) \right) \right] \right\rangle \\
\overset{(d3)}{\leq} \frac{S - A_i}{S} \left( \frac{\eta U}{2} \left\| \nabla \mathcal{L}(\theta^i) \right\|^2 + \frac{\eta}{2U} \mathbb{E} \left[ \left\| \frac{1}{M} \sum_{m=1}^{M} \sum_{u=0}^{U-1} \eta p_m \left( \nabla \mathcal{L}_m(\theta^{ja,m}) - \nabla \mathcal{L}(\theta^i) \right) \right\|^2 \right] \right) \\
\overset{(e3)}{\leq} \frac{S - A_i}{S} \left( \frac{\eta U}{2} \left\| \nabla \mathcal{L}(\theta^i) \right\|^2 + \frac{\eta}{2U} \sum_{m=1}^{M} \sum_{u=0}^{U-1} \mathbb{E} \left[ \left\| \nabla \mathcal{L}_m(\theta^{ja,m}) - \nabla \mathcal{L}(\theta^i) \right\|^2 \right] \right) \\
\overset{(f3)}{\leq} \frac{S - A_i}{S} \left( \frac{\eta U}{2} \left\| \nabla \mathcal{L}(\theta^i) \right\|^2 + \frac{\eta L^2}{2} \sum_{m=1}^{M} \sum_{u=0}^{U-1} \mathbb{E} \left[ \left\| \theta^{ja,m} - \theta^i \right\|^2 \right] \right) \\
\overset{(g3)}{\leq} \frac{S - A_i}{S} \left( \frac{\eta U}{2} \left\| \nabla \mathcal{L}(\theta^i) \right\|^2 + \frac{\eta L^2}{2} \sum_{m=1}^{M} \sum_{u=0}^{U-1} \mathbb{E} \left[ \left\| \frac{1}{2} + 15U^2 \eta^2 L^2 \right\| \nabla \mathcal{L}(\theta^i) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} \left( \sigma_L^2 + 6U \sigma_G^2 \right) \right) \right),
\end{equation}

where (a3) is derived from the definition of \( \tilde{g}^i \); (b3) and (c3) come from direct computation; (d3) uses the fact that \((x, y) \leq \frac{1}{2} \|x\|^2 + \|y\|^2\); (e3) is due to Jensen inequality and Cauchy-Schwarz inequality; (f3) follows from Assumption 1; and (g3) is due to the fact that \( \sum_{m=1}^{M} p_m = 1 \) and [40, Lemma 3], which proves that

\[
\mathbb{E} \left[ \left\| \frac{1}{M} \sum_{m=1}^{M} \sum_{u=0}^{U-1} \eta p_m \left( \nabla \mathcal{L}_m(\theta^{ja,m}) - \nabla \mathcal{L}(\theta^i) \right) \right\|^2 \right] \leq 5U^2 \eta^2 \left( \sigma_L^2 + 6U \sigma_G^2 \right) + 30U^2 \eta^2 \left\| \nabla \mathcal{L}(\theta^i) \right\|^2,
\end{equation}

under the condition that \( \eta \leq \frac{1}{8LU} \), where \( \sigma_L \) and \( \sigma_G \) are two constants defined in Assumption 2.

Likewise, the second term in (A2) can be bounded as below, with \( p_m \) replaced by \( \frac{1}{M} \):

\[
\left\langle \nabla \mathcal{L}(\theta^i), \mathbb{E} \left[ \frac{A_i}{S} \left( \tilde{g}^i + \eta U \nabla \mathcal{L}(\theta^i) \right) \right] \right\rangle \\
\leq \frac{A_i}{S} \left( \eta U \left( \frac{1}{2} + 15U^2 \eta^2 L^2 \right) \left\| \nabla \mathcal{L}(\theta^i) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} \left( \sigma_L^2 + 6U \sigma_G^2 \right) \right).
\]
Substituting (A3) and (A5) into (A2), we have

\[ T_1 \leq \eta U \left( \frac{1}{2} + 15U^2\eta^2L^2 \right) \| \nabla \mathcal{L}(\theta^t) \|^2 + \frac{5U^2\eta^3L^2}{2} \left( \sigma_d^2 + 6U\sigma_d^2 \right). \]  

(A6)

With \( \mathbb{I}\{ \cdot \} \) denoting the indicator function, and \( \mathcal{A} \setminus \mathcal{B} \) denotes the complementary set of set \( \mathcal{B} \) in set \( \mathcal{A} \), the term \( T_2 \) can be bounded as

\[ T_2 = \mathbb{E}[\| g^t \|^2] \]

\[ = \mathbb{E} \left[ \left\| \frac{1}{S} \sum_{m \in \mathcal{M}_U} g^t_m \right\|^2 \right] \]

\[ = \frac{1}{S^2} \mathbb{E} \left[ \left\| \sum_{m=1}^{M} \mathbb{I}\{ m \in \mathcal{M}_U \} g^t_m \right\|^2 \right] \]

\[ \leq \eta^2 S^2 \mathbb{E} \left[ \left\| \sum_{m=1}^{M} \mathbb{I}\{ m \in \mathcal{M}_U \} \sum_{u=0}^{U-1} \left( \nabla \mathcal{L}_m(\theta^t_{mu}, s_{mu}) - \nabla \mathcal{L}_m(\theta^t_{mu}) \right) \right\|^2 \right] \]

\[ + \eta^2 S^2 \mathbb{E} \left[ \left\| \sum_{m=1}^{M} \mathbb{I}\{ m \in \mathcal{M}_U \} \sum_{u=0}^{U-1} \nabla \mathcal{L}_m(\theta^t_{mu}) \right\|^2 \right] \]

\[ \leq \eta^2 U \mathbb{E} \left[ \left\| \sum_{m=1}^{M} \mathbb{I}\{ m \in \mathcal{M}_U \} \sum_{u=0}^{U-1} \nabla \mathcal{L}_m(\theta^t_{mu}) \right\|^2 \right] \]

\[ \leq \frac{\eta^2 U}{SB} \sigma_d^2 + \frac{\eta^2}{S^2} \mathbb{E} \left[ \left\| \sum_{m \in \mathcal{A}'} \sum_{u=0}^{U-1} \nabla \mathcal{L}_m(\theta^t_{mu}) + \sum_{m \in \mathcal{M}_U \setminus \mathcal{A}'} \sum_{u=0}^{U-1} \nabla \mathcal{L}_m(\theta^t_{mu}) \right\|^2 \right] \]

\[ \leq \frac{\eta^2 U}{SB} \sigma_d^2 + \frac{2A^2\eta^2}{S^2} \mathbb{E} \left[ \left\| \sum_{u=0}^{U-1} \sum_{m \in \mathcal{A}'} \nabla \mathcal{L}_m(\theta^t_{mu}) \right\|^2 \right] \]

\[ + \frac{2(S-A^2)\eta^2}{S^2} \sum_{m \in \mathcal{M}_U \setminus \mathcal{A}'} \mathbb{E} \left[ \left\| \sum_{u=0}^{U-1} \nabla \mathcal{L}_m(\theta^t_{mu}) \right\|^2 \right], \]

(A7)
where \((a4)\) is due to the definition of \(g^i\); \((b4)\) comes directly from \((a3)\); \((c4)\) follows from the fact that \(\mathbb{E}[\|x\|^2] = \mathbb{E}[\|x - \mathbb{E}[x]\|^2 + \|\mathbb{E}[x]\|^2]\) and \(\mathbb{E}[\nabla L_m^m(\theta^{i, m, u}_j, \theta^{j, m, u}_j)] = \nabla L_m^m(\theta^{i, m, u}_j);\) \((d4)\) follows from Cauchy-Schwarz inequality; \((e4)\) is due to the fact that \(\mathbb{E}[\|x_1 + \ldots + x_n\|^2] = \mathbb{E}[\|x_1\|^2 + \ldots + \|x_n\|^2]\) if \(x_i^s\)’s are independent with zero mean; \((f4)\) is due to the definition that the subset of workers selected by ages at round \(j\) is \(A^j\) and \(|A^j| = A^j = \min\{S, S^j\}\); \((g4)\) is due to Cauchy-Schwarz inequality.

Next, we bound the term \(\mathbb{E} \left[ \left\| \sum_{u=0}^{U-1} \nabla L_m^m(\theta^{i, m, u}_j) \right\|^2 \right]\) in \((A7)\) and \((A3)\) as follows:

\[
\mathbb{E} \left[ \left\| \sum_{u=0}^{U-1} \nabla L_m^m(\theta^{i, m, u}_j) \right\|^2 \right] = \mathbb{E} \left[ \left\| \sum_{u=0}^{U-1} (\nabla L_m^m(\theta^{i, m, u}_j) - \nabla L_m^m(\theta^j) + \nabla L_m^m(\theta^j) - \nabla L^m(\theta^j) + \nabla L^m(\theta^j)) \right\|^2 \right] \tag{A8}
\]

\[
\leq 3U L^2 \sum_{u=0}^{U-1} \mathbb{E}[\|\theta^{i, m, u}_j - \theta^j\|^2] + 3U^2 \sigma^2_G + 3U^2 \|\nabla L(\theta^j)\|^2
\]

\[
\leq 15U^3 L^2 \eta^2 (\sigma^2_L + 6U \sigma^2_G) + (90U^4 L^2 \eta^2 + 3U^2)\|\nabla L(\theta^j)\|^2 + 3U^2 \sigma^2_G
\]

\[
= C_1 \|\nabla L(\theta^j)\|^2 + C_2,
\]

where \((a5)\) is due to Cauchy-Schwarz inequality and the bounded variance assumption; \((b5)\) follows from \((A4);\) \((c5)\) is due to the definition that \(C_1 = 90U^4 L^2 \eta^2 + 3U^2\) and \(C_2 = 15U^3 L^2 \eta^2 (\sigma^2_L + 6U \sigma^2_G) + 3U^2 \sigma^2_G.\)

By substituting the upper bounds \((A6), (A7)\) of the terms \(T_1\) and \(T_2\) in \((A1)\), we readily have:
\[ \mathbb{E}[\mathcal{L}(\theta^{t+1})] \]
\[
\leq \mathcal{L}(\theta^t) - \eta U \left( \frac{1}{2} - 15U^2 \eta^2 L^2 \right) \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + \frac{L\eta^2 A^i}{S^2} \sum_{m \in A^i} \mathbb{E} \left[ \left\| \mathcal{L}_m(\theta_m^{i,u}) \right\|^2 \right] + \frac{5U^2 \eta^3 L^2}{2} (\sigma^2_L + 6U\sigma^2_G) + \frac{\eta^2 UL}{2SB} \sigma^2_L \\
+ \frac{L\eta^2 (S - A^i)}{S^2} \sum_{m \in M_t \setminus A^i} \mathbb{E} \left[ \left\| \mathcal{L}_m(\theta_m^{i,u}) \right\|^2 \right]
\]
\[
\leq \mathcal{L}(\theta^t) - \eta U \left( \frac{1}{2} - 15U^2 \eta^2 L^2 \right) \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} (\sigma^2_L + 6U\sigma^2_G) + \frac{L\eta^2 (A^i)^2 + (S - A^i)^2}{S^2} (C_1 \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + C_2) + \frac{\eta^2 UL}{2SB} \sigma^2_L \\
\leq \mathcal{L}(\theta^t) - \eta U \left( \frac{1}{2} - 15U^2 \eta^2 L^2 - \frac{L\eta(2(A^i)^2 - 2SA^i + S^2)}{S^2 U} C_1 \right) \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} (\sigma^2_L + 6U\sigma^2_G) + \frac{L\eta^2 (2(A^i)^2 - 2SA^i + S^2)}{S^2} C_2 \\
\leq \mathcal{L}(\theta^t) - \eta U \left( \frac{1}{2} - 15U^2 \eta^2 L^2 - \frac{L\eta}{U} C_1 \right) \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} (\sigma^2_L + 6U\sigma^2_G) + \frac{L\eta^2 (2(A^i)^2 - 2SA^i + S^2)}{S^2} C_2 \\
\leq \mathcal{L}(\theta^t) - \eta U \left( \frac{1}{2} - 15U^2 \eta^2 L^2 - \frac{L\eta}{U} C_1 \right) \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} (\sigma^2_L + 6U\sigma^2_G) + \frac{L\eta^2 (2(A^i)^2 - 2SA^i + S^2)}{S^2} C_2 \\
\leq \mathcal{L}(\theta^t) - \eta U \left( \frac{1}{2} - 15U^2 \eta^2 L^2 - \frac{L\eta}{U} C_1 \right) \left\| \nabla \mathcal{L}(\theta^t) \right\|^2 + \frac{5U^2 \eta^3 L^2}{2} (\sigma^2_L + 6U\sigma^2_G) + \frac{\eta^2 UL}{2SB} \sigma^2_L \\
+ \frac{2L\eta^2 (A^i)^2}{S^2} C_2 - \frac{2L\eta^2 A^i}{S} C_2,
\]

(A9)

where (a6) comes from direct substitution; (b6) uses the result in (A8); (c6) follows from direct computation; (d6) uses the fact that 0 \leq A^i \leq S; and (e6) follows from the fact that there exists a constant c such that 0 < c < \frac{1}{2} - 15U^2 \eta^2 L^2 - \eta(90U^3 L^2 \eta^2 + 3U). The proof of Lemma 1 is then complete.

**Appendix B: Proof of Theorem 1**

With Lemma 1, we have
\[
\mathbb{E}[\mathcal{L}(\theta^{t+1})] \leq \mathcal{L}(\theta^t) - c\eta U \mathbb{V} \mathcal{L}(\theta^t) + \frac{5U^2\eta^3L^2}{2} \left(\sigma_L^2 + 6U\sigma_G^2\right)
\]
\[
+ \frac{L^2(2A^0 L + S^2)}{S^2} - \frac{2L\eta^2 A}{S} C_2
\]
\[
\leq \mathcal{L}(\theta^t) - c\eta U \mathbb{V} \mathcal{L}(\theta^t) + \frac{5U^2\eta^3L^2}{2} \left(\sigma_L^2 + 6U\sigma_G^2\right)
\]
\[
+ \frac{\eta^2UL}{2SB} \sigma_L^2 + 3L\eta^2 C_2 - \frac{2L\eta^2 A}{S} C_2,
\]
where (a8) is from (A9); and (b8) uses the fact that \(0 \leq A^t \leq S\).

With the age-based mechanism, we denote the minimum number of rounds to traverse all the workers in \(\mathcal{M}\) as \(R\), i.e., we have \(A^0 + \ldots + A^{R-1} \geq M\). By rearranging the terms in (B10) and summing from \(j = 0, \ldots, R - 1\), we can have:
\[
\frac{1}{R} \mathbb{E} \left[ \sum_{j=0}^{R-1} \mathbb{V} \mathcal{L}(\theta^j) \right] \leq \frac{\mathcal{L}(\theta^0) - \mathcal{L}(\theta^R)}{c\eta UR} + V_1,
\]
where \(V_1 = \frac{1}{c\eta U} \left(\frac{\eta^2UL}{2SB} \sigma_L^2 + \frac{5U^2\eta^3L^2}{2} \left(\sigma_L^2 + 6U\sigma_G^2\right) + (3L\eta^2 - \frac{2L\eta^2 M}{SR}) C_2\right)\).

Thus, when \(J\) is a multiple of \(R\), we can then readily write
\[
\frac{1}{J} \mathbb{E} \left[ \sum_{j=0}^{J-1} \mathbb{V} \mathcal{L}(\theta^j) \right] \leq \frac{\mathcal{L}(\theta^0) - \mathcal{L}^*}{c\eta UJ} + V,
\]
where we used Assumption 1 and \(V = \frac{1}{c\eta U} \left(\frac{\eta L}{2SB} \sigma_L^2 + \frac{5U^2\eta^3L^2}{2} \left(\sigma_L^2 + 6U\sigma_G^2\right) + (3\eta L - \frac{2L\eta M}{SR}) (15U^2L^2\eta^2(\sigma_L^2 + 6U\sigma_G^2) + 3U\sigma_G^2)\right)\). The proof is then complete.

**Appendix C: Additional experiments**

In this section, we provide more simulation results, including testing the performance of the AgeSel, FedAvg, OCS and RR on CIFAR-10 dataset and testing the performance of AgeSel considering different selections of \(S\) with \(S = 3\) and \(S = 15\).

Figures 5 and 6 present the performance of the four schemes mentioned in the main text in terms of communication cost and training rounds. It can be clearly observed that under datasets much larger than the MNIST, i.e., CIFAR-10, AgeSel still achieves the best performance regarding the two metrics.

Figures 7, 8, 9 and 10 test the performances of the four schemes under \(S = 15\) and \(S = 3\). Horizontally speaking, when \(S = 15\), although the difference between the schemes is reduced, the proposed AgeSel still achieves the best performance in terms of communication cost and training rounds. When \(S = 3\), all the schemes experience more fluctuations, but AgeSel also needs the lowest number of training rounds and communication cost. Vertically speaking, When \(S = 15\), AgeSel needs more communication cost to reach the same accuracy as \(S = 5\) due to the increase of number of workers selected. When \(S = 3\), although AgeSel requires...
Fig. 5  Comparison of FedAvg, OCS, RR and AgeSel in terms of training rounds with CIFAR-10

Fig. 6  Comparison of FedAvg, OCS, RR and AgeSel in terms of communication cost with CIFAR-10

Fig. 7  Comparison of FedAvg, OCS, RR and AgeSel in terms of training rounds with $S = 15$
Fig. 8  Comparison of FedAvg, OCS, RR and AgeSel in terms of communication cost with $S = 15$

Fig. 9  Comparison of FedAvg, OCS, RR and AgeSel in terms of training rounds with $S = 3$

Fig. 10  Comparison of FedAvg, OCS, RR and AgeSel in terms of communication cost with $S = 3$
less communication cost than $S = 5$, the variance is relatively large and the performance is more unstable. To summarize, AgeSel is consistently the best algorithm regardless of the choice of $S$. With regard to the selection of $S$, if $S$ is too large, then the communication cost would be high; if $S$ is too small, then the performance is unstable. Therefore, a value in the middle, in this case $S = 5$, is a decent selection of $S$ to reach a balance between communication cost and performance stability.

Author Contributions FZ and JZ came up with the idea; FZ wrote the manuscript; JZ and XW revised and polished the manuscript.

Funding This work was supported by the National Natural Science Foundation of China Grants No. 62101134 and No. 62071126, and the Innovation Program of Shanghai Municipal Science and Technology Commission Grant 20JC1416400.

Data Availability The EMNIST dataset could be downloaded from https://www.nist.gov/itl/products-and-services/emnist-dataset.

Declarations

Conflict of interest This work was supported by the National Natural Science Foundation of China Grants No. 62101134 and No. 62071126, and the Innovation Program of Shanghai Municipal Science and Technology Commission Grant 20JC1416400.

References

1. Li M, Andersen DG, Park JW, Smola AJ, Su BY (2014) Scaling distributed machine learning with the parameter server. In: Proceedings of the USENIX OSDI, pp 583–598
2. Lian X, Zhang C, Zhang H, Hsieh CJ, Zhang W, Liu J (2017) Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. In: Proceedings of the NeurIPS, vol 30, pp 5336–5346
3. Garcia Lopez P, Montresor A, Epema D, Datta A, Higashino T, Ianniti A et al (2015) Edge-centric computing: vision and challenges. ACM SIGCOMM Comput Commun Rev 45(5):37–42
4. Hong K, Lilletun D, Ramachandran U, Ottenwälder B, Koldehofe B (2013) Mobile fog: a programming model for large-scale applications on the internet of things. In: Proceedings of the Second ACM SIGCOMM Workshop on Mobile Cloud Computing, pp 15–20
5. Bonomi F, Milito R, Zhu J, Addepalli S (2012) Fog computing and its role in the internet of things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, pp 13–16
6. Deshpande A, Guestrin C, Madden SR, Hellerstein JM, Hong W (2005) Model-based approximate querying in sensor networks. VLDB J 14:417–443
7. Dean J, Corrado GS, Monga R, Kai C, Ng AY (2012) Large scale distributed deep networks. In: Proceedings of the NeurIPS, vol 25, pp 1223–1231
8. Yan N, Wang K, Pan C, Chai KK (2022) Performance analysis for channel-weighted federated learning in OMA wireless networks. IEEE Signal Process Lett 29:772–776
9. Uddin MP, Xiang Y, Yearwood J, Gao L (2021) Robust federated averaging via outlier pruning. IEEE Signal Process Lett 29:409–413
10. Han S-S, Kim Y-K, Jeon Y-B, Park J, Park D-S, Hwang D, Jeong C-S (2020) Distributed deep learning platform for pedestrian detection on the convergence environment. J Supercomput 76(7):5460–5485
11. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings of the COMPSTAT, pp 177–186
12. Mcmahan HB, Moore E, Ramage D, Hampson S, Arcas B (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the Artificial intelligence and statistics, pp 1273–1282
13. Chen T, Guo Z, Sun Y, Yin W (2021) CADA: communication-adaptive distributed Adam. In: Proceedings of the ICAIS, pp 613–621
14. Chen W, Horvath S, Richtarik P (2020) Optimal client sampling for federated learning. arXiv preprint arXiv:2010.13723
15. Goetz J, Malik K, Bui D, Moon S, Liu H, Kumar A (2019) Active federated learning. preprint arXiv:1909.12641
16. Zinkevich M, Weimer M, Smola AJ, Li L (2010) Parallelized stochastic gradient descent. Proc Neural Inf Process Syst 23:2595–2603
17. Arjeyani Y, Shamir O (2015) Communication complexity of distributed convex learning and optimization. In: Proceedings of Advances in Neural Information Processing Systems, vol 28, pp 1756–1764
18. Zhou F, Cong G (2018) On the convergence properties of a K-step averaging stochastic gradient descent algorithm for nonconvex optimization. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp 3219–3227
19. Wang J, Joshi G (2021) Cooperative SGD: a unified framework for the design and analysis of local-update sgd algorithms. J Mach Learn Res 22(1):9709–9758
20. Woodworth B, Wang J, Mcmahan B, Srebro N (2018) Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. In: Proceedings of the Neural Information Processing Systems, vol 31, pp 8505–8515
21. Li X, Huang K, Yang W, Wang S, Zhang Z (2019) On the convergence of FedAvg on non-iid data. arXiv preprint arXiv:1907.02189
22. Haddadpour F, Kamani MM, Mahdavi M, Cadambe VR (2019) Local SGD with periodic averaging: tighter analysis and adaptive synchronization. In: Proceedings of the Neural Information Processing Systems, vol 32, pp 11082–11094
23. Wang S, Tuor T, Salonidis T, Leung KK, Makaya C, He T, Chan K (2019) Adaptive federated learning in resource constrained edge computing systems. IEEE J Sel Areas Commun 37(6):1205–1221
24. Yu H, Yang S, Zhu S (2019) Parallel restarted SGD with faster convergence and less communication: demystifying why model averaging works for deep learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 5693–5700
25. Khaleed A, Mishchenko K, Richtárik P (2019) First analysis of local GD on heterogeneous data. arXiv preprint arXiv:1907.04715
26. Malinovskiy G, Kovalev D, Gasanov E, Condat L, Richtarik P (2020) From local SGD to local fixed-point methods for federated learning. In: International Conference on Machine Learning, pp 6692–6701
27. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2020) Federated optimization in heterogeneous networks. In: Proceedings of Machine Learning and Systems, vol 2, pp 429–450
28. Yang H, Fang M, Liu J (2021) Achieving linear speedup with partial worker participation in non-iid federated learning. arXiv preprint arXiv:2101.11203
29. Ye Q, Zhou Y, Shi M, Lv J (2022) FLASGD: free local SGD with parallel synchronization. J Supercomput 78(10):12410–12433
30. Ribero M, Vishalo H (2020) Communication-efficient federated learning via optimal client sampling. arXiv preprint arXiv:2007.15197
31. Cho YJ, Gupta S, Joshi G, Yağan O (2020) Bandit-based communication-efficient client selection strategies for federated learning. In: Proceedings of the Asilomar Conference on Signals, Systems, and Computers, pp 1066–1069
32. Shah V, Wu X, Sanghavi S (2020) Choosing the sample with lowest loss makes SGD robust. In: International Conference on Artificial Intelligence and Statistics, pp 2120–2130
33. Cho YJ, Wang J, Joshi G (2022) Towards understanding biased client selection in federated learning. In: International Conference on Artificial Intelligence and Statistics, pp 10351–10375
34. Yin D, Pananjady A, Lam M, Papailiopoulos D, Ramchandran K, Bartlett P (2018) Gradient diversity: a key ingredient for scalable distributed learning. In: Proceedings of the ICAIS, pp 1998–2007
35. Kaul S, Yates R, Gruteser M (2012) Real-time status: How often should one update? In: Proceedings of the IEEE INFOCOM, pp 2731–2735
36. Ozfatwa E, Buyukates B, Gündüz D, Ulukus S (2020) Age-based coded computation for bias reduction in distributed learning. In: Proceedings of the IEEE GLOBECOM, pp 1–6
37. Yang HH, Arafa A, Quek TQS, Vincent Poor H (2020) Age-based scheduling policy for federated learning in mobile edge networks. In: Proceedings of the IEEE ICASSP, pp 8743–8747
38. Yang HH, Liu Z, Quek TQ, Poor HV (2019) Scheduling policies for federated learning in wireless networks. IEEE Trans Commun 68(1):317–333
39. Buyukates B, Ulukus S (2021) Timely communication in federated learning. In: Proceedings of the IEEE INFOCOM, pp 1–6
40. Reddi S, Charles Z, Zaheer M, Garrett Z, Rush K, Konečný J et al (2020) Adaptive federated optimization. arXiv preprint arXiv:2003.00295

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.