Adaptive Top-K in SGD for Communication-Efficient Distributed Learning

Mengzhe Ruan[1,2]   Guangfeng Yan[1,2]   Yuanzhang Xiao[3]   Linqi Song[1,2]   Weitao Xu[1,2,*]

[1] City University of Hong Kong Shenzhen Research Institute, Shenzhen, China
[2] Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China
[3] Hawaii Advanced Wireless Technologies Institute, University of Hawaii at Manoa, Honolulu, HI, USA

Abstract—**Distributed stochastic gradient descent (SGD) with gradient compression has become a popular communication-efficient solution for accelerating distributed learning. One commonly used method for gradient compression is Top-K sparsification, which sparsifies the gradients by a fixed degree during model training. However, there has been a lack of an adaptive approach to adjust the sparsification degree to maximize the potential of the model's performance or training speed. This paper proposes a novel adaptive Top-K in SGD framework that enables an adaptive degree of sparsification for each gradient descent step to optimize the convergence performance by balancing the trade-off between communication cost and convergence error. Firstly, an upper bound of convergence error is derived for the adaptive sparsification scheme and the loss function. Secondly, an algorithm is designed to minimize the convergence error under the communication cost constraints. Finally, numerical results on the MNIST and CIFAR-10 datasets demonstrate that the proposed adaptive Top-K algorithm in SGD achieves a significantly better convergence rate compared to state-of-the-art methods, even after considering error compensation.**

I. INTRODUCTION

Nowadays, with extensive data collected in distributed networks, there is an increasing need for distributed learning algorithms that aggregate local gradients to learn a global models. Distributed stochastic gradient descent (SGD) is the core of most distributed learning algorithms [1]. In practical networks, however, the communication overhead of transmitting gradients often becomes the performance bottleneck due to the limited bandwidth. Gradient compression, which uses less information to represent the gradients, is an effective and efficient method to solve this problem. The compression methods, however, inevitably introduce compression noise which affects the convergence of the model. Therefore, how to choose the compression methods and the compression level efficiently to balance the trade-off between communication cost and convergence performance remains an open challenge.

Traditional compression methods often compress parameters with a *fixed* compression factor for all the training iterations, which may not be optimal. To further improve communication efficiency, an online learning method was proposed in [11] to adaptively adjust the degree of gradient sparsity when the total dataset is non-i.i.d distributed in the federated learning network. Unfortunately, there lacks a theoretical convergence analysis in their research. In [9], an adaptive quantization

method is proposed and its theoretical guarantee has also been proved. Nevertheless, the quantization method needs more computing resources than sparsification methods, which simply keep some components of the gradient and set others to zero. Therefore, we would like to investigate the adaptive sparsification methods in distributed SGD. We will improve upon Top-K, the most commonly-used biased sparsification method, which keeps only a few coordinates of the stochastic gradient with the largest magnitudes.

In this paper, we propose a novel adaptive Top-K SGD framework, named AdapTop-K, that aims to improve the convergence performance of Top-K while maintaining the same communication cost. Under the assumption of smoothness and Polyak-Lojasiewicz condition [10], we derive an upper bound on the gap between the loss function and the optimal loss to characterize the convergence error caused by limited iteration steps, sampling, and adaptive Top-K sparsification. Based on the theoretical analysis, we design an adaptive Top-K method by minimizing the convergence upper bound under the desired total communication cost. The proposed AdapTop-K algorithm adjusts the degree of sparsification by considering the desired model performance, the number of rounds, and the norm of gradients. We validate our theoretical analysis through experiments on image classification tasks on the MNIST and CIFAR-10 datasets. Numerical results show that AdapTop-K outperforms the baseline sparsification methods.

To summarize, our key contributions are as follows:

• We propose a novel framework to characterize the trade-off between the communication cost and the convergence rate by adaptively adjusting the gradient sparsification levels in distributed learning. We analyze the convergence error of the loss function under Top-K sparsification for gradients over different communication rounds. We isolate our bound on the convergence error to characterize the impact of adaptive sparsification on the convergence rate.

• We solve the optimization problem that minimizes the convergence error while keeping the same communication cost as Top-K. To achieve this, we propose a novel adaptive Top-K algorithm called AdapTop-K, which dynamically adjusts the degree of gradient sparsification during training to improve model performance.

• We validate the proposed AdapTop-K on the popular datasets and machine learning models, demonstrating that our proposed AdapTop-K outperforms state-of-the-art gradient

sparsification methods.

II. RELATED WORK

There are two main approaches to compress SGD to reduce communication cost: quantization and sparsification. Quantization compresses gradients by limiting the number of bits representing floating point numbers during communication. The gradient quantization was proposed in [6]. There are several variants of quantization, including error compensation [7], variance-reduced quantization [12], quantization to a ternary vector [13], and quantization of gradient difference [14]. Sparsification methods aim to reduce the number of non-zero entries in the stochastic gradients [8]. An aggressive sparsification method (Top-K) [5] is to keep very few coordinates of the stochastic gradient with the largest magnitudes. The methods can also be classified based on whether the compression is biased or unbiased. The unbiased methods could keep the expectation of compressed gradients as that of the true gradients [6] and [13]. In contrast, the biased methods introduce bias in the compression and more compression noise to the optimization process [5]. These methods can compress the gradient efficiently to speed up distributed training. However, they do not consider adaptively changing the degree of compression during training, which is the key difference between our method and existing methods.

III. SYSTEM MODEL

We consider a distributed learning system with a central server and M edge devices (workers). The workers collaborate to train a shared machine learning model by aggregating the gradient or its variant in cooperation with the central server.

The learning model is represented by the vector of its parameters $\mathbf{w} \in \mathbb{R}^d$, where d is the model size. The datasets are distributed over the M workers. We use $\mathcal{D}^i$ to denote the local dataset at worker i. The global loss function, denoted by $F : \mathbb{R}^d \to \mathbb{R}$, is defined as

$$F(\mathbf{w}) = \frac{1}{M} \sum_{i=1}^{M} f^i(\mathbf{w}),$$ (1)

$$\text{with } f^i(\mathbf{w}) = \mathbb{E}_{\xi^i \sim \mathcal{D}^i} \left[l^i(\mathbf{w}; \xi^i) \right],$$

where $l^i(\mathbf{w}; \xi^i)$ is the local loss function of the model parameters $\mathbf{w}$ at work i, given the mini-batch ξ^i randomly selected from worker i's local dataset $\mathcal{D}^i$.

The objective of the training is to find a model parameter $\mathbf{w}$ to minimize the global loss function in Eq. (1) :

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} F(\mathbf{w}).$$ (2)

The distributed SGD is the most popular method to solve this problem, where each worker i computes its local stochastic gradient $\mathbf{g}_t^i = \nabla l^i(\mathbf{w}_t; \xi^i)$ given parameters $\mathbf{w}_t$ at round t. Then the workers send the local gradient $\mathbf{g}_t^i$ to the central server. The server aggregates these gradients to update the model. To reduce the communication cost, we compress the local stochastic gradients before sending them to the server:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{M} \sum_{i=1}^{M} \mathcal{C}^i[\mathbf{g}_t^i],$$ (3)

where η_t is the learning rate at iteration t, and $\mathcal{C}^i[\cdot]$ is the compression operator. Without the gradient compressor, Eq. (3) reduces to the vanilla distributed SGD with $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_t}{M} \sum_{i=1}^{M} \mathbf{g}_t^i$. The same procedure is repeated until the convergence criterion or the maximum number of communication rounds is reached.

A commonly-used compression operator is Top-K, where each worker i keeps only k elements of the gradient $\mathbf{g}_t^i$ with the largest magnitudes and sets the other elements to zero [5]. In this work, we speed up the convergence of Top-K by adaptively choosing the sparsity of the gradient during the convergence process. Specifically, given a total of T rounds of gradient update, our goal is to find the optimal sparsity levels $k_0, \ldots, k_{T-1}$ in each round, so that the final model is as close to the optimal model as possible. It is natural to measure the gap from the optimal model by the difference between the expectation of the final global loss $F(\mathbf{w}_T)$ and the optimal loss $F^* = F(\mathbf{w}^*)$. Note that we need to take expectation of the final loss $F(\mathbf{w}_T)$ due to the stochastic gradient descent. Therefore, our design problem can be formulated as follows

$$\min_{k_0, \ldots, k_{T-1}} \quad \mathbb{E}\left[F(\mathbf{w}_T)\right] - F^*$$ (4)

$$\text{s.t.} \quad \sum_{t=0}^{T-1} k_t \leq K,$$

$$k_t \in \{0, 1, \ldots, d\}, \ t = 0, \ldots, T-1,$$

where K is the total budget for the communication overhead during the training. When comparing with other sparsification methods, we can set the communication budget K accordingly.

IV. PROPOSED ALGORITHM

In this section, we first provide convergence analysis of AdapTop-K given a sequence of sparsity levels $k_0, \ldots, k_{T-1}$. Based on the analysis, we then propose a practical algorithm for finding a sequence $k_0, \ldots, k_{T-1}$ that guarantees to outperform the standard Top-K method.

A. Convergence Analysis

For the convergence analysis, we make standard assumptions on the stochastic gradient and the loss function that are commonly used in the literature [9], [3], and [2].

Assumption 1: (Smoothness). There exists a non-negative constant L such that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$F(\mathbf{x}) - F(\mathbf{y}) - \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2,$$ (5)

where $\nabla F(\mathbf{y})$ is the gradient of the loss function $F(\cdot)$ at $\mathbf{y}$.

Assumption 2: (Polyak-Lojasiewicz Condition). There exists a constant $\mu \geq 0$ such that for any $\mathbf{w} \in \mathbb{R}^d$, we have

$$\|\nabla F(\mathbf{w})\|^2 \geq 2\mu(F(\mathbf{w}) - F^*).$$ (6)

Note that Assumption 2 is milder than the assumption of strong convexity [10].

Assumption 3: (Unbiasedness and Bounded Variance of Stochastic Gradient). The local stochastic gradients $\mathbf{g}^i$ are assumed to be independent and unbiased estimates of the local gradient $\nabla f^i(\mathbf{w}_t)$ with bounded variance:

$$\mathbb{E}_{\xi^i \sim \mathcal{D}^i} \left[\mathbf{g}_t^i \right] = \nabla f^i(\mathbf{w}_t), \qquad (7)$$
$$\mathbb{E}_{\xi^i \sim \mathcal{D}^i} \left[\|\mathbf{g}_t^i - \nabla f^i(\mathbf{w}_t)\|^2 \right] \leq \sigma^2.$$

As proven in [4], the gradient update in Eq. (3) can be rewritten as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathcal{C}[\mathbf{g}_t], \qquad (8)$$

where $\mathbf{g}_t$ is the stochastic gradient of the global loss function

$$\mathbf{g}_t = \nabla F(\mathbf{w}_t) + \mathbf{m}_t(\mathbf{w}_t), \qquad (9)$$

and $\mathcal{C}[\cdot]$ is the aggregate Top-K operator

$$\mathcal{C}[\mathbf{g}_t] = \nabla F(\mathbf{w}_t) + \mathbf{m}_t(\mathbf{w}_t) + \mathbf{b}_t(\mathbf{w}_t), \qquad (10)$$

where $\mathbf{m}_t(\mathbf{w}_t)$ is the noise in SGD and $\mathbf{b}_t(\mathbf{w}_t)$ is the bias introduced by sparsification.

By Assumption 3, the noise has zero mean and bounded variance, namely

$$\mathbb{E}[\mathbf{m}_t(\mathbf{w}_t)] = 0 \quad \text{and} \quad \mathbb{E}[\|\mathbf{m}_t(\mathbf{w}_t)\|^2] \leq \sigma^2. \qquad (11)$$

An upper bound of the variance of the bias is given in [5]. We summarize the result as a lemma here.

Lemma 1: (Bounded Variance of Stochastic Gradient with Top-K sparsification). The variance of the bias $\mathbf{b}_t(\mathbf{w}_t)$ is upper bounded by the mini-batch gradient $\mathbf{g}_t$ as follows: [5]

$$\|\mathbf{b}_t(\mathbf{w}_t)\|^2 \leq \left(1 - \frac{k}{d}\right) \|\mathbf{g}_t\|^2. \qquad (12)$$

With Lemma 1, we prove an upper bound of the optimality gap under the adaptive sparsity levels of $k_0, \ldots, k_{T-1}$.

Theorem 1: (Upper Bound for Convergence Error). Under Assumptions 1–3, given the initial parameter $\mathbf{w}_0$ and constant stepsize $\eta_t = \eta \leq \frac{1}{L}$, the optimality gap of the adaptive Top-K method is upper bounded by

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \underbrace{\left(1 - \frac{\eta\mu}{d}k\right)^T [F(\mathbf{w}_0) - F^*]}_{\mathbf{M}(k)} \qquad (13)$$
$$+ \underbrace{\frac{d\sigma^2}{2k\mu}\left(1 - \frac{k}{d} + \eta L\right)\left[1 - \left(1 - \frac{\eta\mu}{d}k\right)^T\right]}_{\mathbf{N}(k)}$$
$$- \underbrace{\sum_{t=0}^{T-1}\left[\left(\frac{\eta n_t}{2d}\|\mathbf{g}_t\|^2\right)\left(1 - \frac{\eta\mu}{d}k\right)^{T-1-t}\right]}_{\text{the only term that depends on } n_t},$$

where $k = \frac{K}{T}$ is the average sparsity level and $n_t = k_t - k$ is the deviation from the average sparsity level at round t.

Proof: See the appendix. The proofs of all the other results can be found in our technical report [15]. ∎

The upper bound in (13) has two parts. The first part is the sum of the first two terms $\mathbf{M}(k) + \mathbf{N}(k)$, which depends only on the average sparsity level k. The second part is the third term, which is the only term that depends on $n_0, \ldots, n_{T-1}$. When $n_t = 0$ for all t, the upper bound reduces to $\mathbf{M}(k) + \mathbf{N}(k)$, namely the bound for the vanilla Top-K method.

B. The Proposed AdapTop-K Algorithm

We aim to minimize the upper bound of the optimality gap in (13) by choosing $n_0, \ldots, n_{T-1}$. Since only the third term depends on the adjustments $n_0, \ldots, n_{T-1}$, the optimization problem can be formulated as

$$\max_{n_0, \ldots, n_{T-1}} \quad \sum_{t=0}^{T-1}\left[\left(\frac{\eta n_t}{2d}\|\mathbf{g}_t\|^2\right)\left(1 - \frac{\eta\mu}{d}k\right)^{T-1-t}\right] \qquad (14)$$
$$\text{s.t.} \quad \sum_{t=0}^{T-1} n_t \leq 0,$$
$$n_t \in \{-k, \ldots, d-k\}, \ t = 0, \ldots, T-1,$$

where the first constraint comes from the constraint on the communication overhead in (4) and the second constraint comes from the fact that $k_t \in \{0, \ldots, d\}$.

Since the objective function is linear in n_t, the optimal solution should assign the largest possible values to the n_t's with the largest coefficients

$$\left(\frac{\eta}{2d}\|\mathbf{g}_t\|^2\right)\left(1 - \frac{\eta\mu}{d}k\right)^{T-1-t}, \qquad (15)$$

subject to the upper bound $d - k$ and the budget of total communication overhead. However, the major challenge is that the coefficients in (15) depend on the gradients $\mathbf{g}_t$, which are stochastic due to the randomly selected mini-batches and are dependent on our choice of sparsity levels $n_0, \ldots, n_{t-1}$ up to round t. Therefore, we cannot solve the optimization problem (14) directly. Instead, we choose to maximize an upper bound of the objective function, which is obtained by bounding the norm of the stochastic gradients $\mathbf{g}_t$.

Lemma 2: (Upper Bound for Stochastic Gradient). Under Assumptions 1–3, given the initial parameter $\mathbf{w}_0$ and constant stepsize $\eta_t = \eta \leq \frac{1}{L}$, the stochastic gradient in Eq. (9) can be upper bounded by

$$\mathbb{E}[\|\mathbf{g}_t\|^2] \leq \frac{2d}{k\eta} \cdot \frac{F(\mathbf{w}_0)}{t} + \frac{d\sigma^2}{k}(\eta L + 1) \triangleq \frac{\alpha}{t} + \beta \qquad (16)$$

where $\alpha \triangleq \frac{2d}{k\eta}F(\mathbf{w}_0)$ and $\beta \triangleq \frac{d\sigma^2}{k}(\eta L + 1)$.

Based on Lemma 2, we obtain the following upper bound of the objective function in (14)

$$\frac{\eta}{2d}\sum_{t=0}^{T-1}\left[\left(\frac{\alpha}{t} + \beta\right)\left(1 - \frac{\eta\mu}{d}k\right)^{T-1-t}\right] \cdot n_t$$
$$\triangleq \frac{\eta}{2d}\sum_{t=0}^{T-1}(A_t B_t) \cdot n_t, \qquad (17)$$

where $A_t \triangleq \frac{\alpha}{t} + \beta$ and $B_t \triangleq (1 - \frac{\eta\mu}{d}k)^{T-1-t}$.

Finally, the optimization problem to solve is

$$\max_{n_0, \ldots, n_{T-1}} \quad \frac{\eta}{2d}\sum_{t=0}^{T-1}(A_t B_t) \cdot n_t \qquad (18)$$
$$\text{s.t.} \quad \sum_{t=0}^{T-1} n_t \leq 0,$$
$$n_t \in \{-k, \ldots, d-k\}, \ t = 0, \ldots, T-1.$$

The objective function in (18) is linear in n_t with coefficient $A_t B_t$. We can prove the following monotonicity results.

Lemma 3: The coefficient $A_t B_t$ first decreases with t and then increases with t. Specifically, we have

$$A_{t+1}B_{t+1} < A_t B_t, \text{ for } t < \hat{t} \triangleq \left\lfloor \frac{-\alpha + \sqrt{\Delta}}{2\beta} \right\rfloor, \text{ and}$$

$$A_{t+1}B_{t+1} \geq A_t B_t, \text{ for } t \geq \hat{t}, \quad (19)$$

where $\Delta \triangleq \alpha^2 - \frac{4\alpha\beta}{\ln B}$, $B \triangleq 1 - \frac{\eta\mu}{d}k$, and $\lfloor \cdot \rfloor$ is the floor function.

Given the monotonicity result in Lemma 3, we design the following adaptive sparsity levels

$$\begin{cases} n_t = +\gamma k \Rightarrow k_t = (1+\gamma)k, & t \in [0, \frac{\hat{t}}{2}) \cup [\frac{\hat{t}+T}{2}, T-1] \\ n_t = -\gamma k \Rightarrow k_t = (1-\gamma)k, & t \in [\frac{\hat{t}}{2}, \frac{\hat{t}+T}{2}), \end{cases} \quad (20)$$

where γ is the scaling factor (i.e., a hyperparameter). In the above scheme, n_t takes the negative value half the training time and the positive value the other half, which satisfies the communication budget constraint. To maximize the objective function, we set n_t to be positive when $A_t B_t$ is larger.

We can prove that the above adaptive sparsity levels result in a lower convergence error compared to the vanilla Top-K.

Corollary 1: (Convergence Error Bound using AdapTop-K in distributed SGD). Under the adaptive sparsity levels in Eq. (20), the optimality gap is upper bounded by

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \mathbf{M}(k) + \mathbf{N}(k)$$

$$+ \frac{\eta\gamma k}{2d} \underbrace{\left(\sum_{t=\frac{\hat{t}}{2}}^{\frac{\hat{t}+T-1}{2}} A_t B_t - \sum_{t=0}^{\frac{\hat{t}}{2}} A_t B_t - \sum_{t=\frac{\hat{t}+T-1}{2}}^{T-1} A_t B_t \right)}_{\text{always less than 0 because of (19)}}$$

$$< \underbrace{\mathbf{M}(k) + \mathbf{N}(k)}_{\text{upper bound for SGD with vanilla Top-K}}. \quad (21)$$

The pseudo-code of distributed SGD with the proposed AdapTop-K method is provided in Algorithm 1.

V. EVALUATION

In this section, we conduct experiments on two widely used datasets, namely MNIST and CIFAR-10, to validate the effectiveness of our proposed AdapTop-K method. We conduct experiments for $M = 8$ workers and use canonical networks to evaluate the performance on the image classification task using different algorithms: fully-connected network on the MNIST dataset, and Resnet18 on the CIFAR-10 dataset. The above datasets are the database commonly used for training various image processing systems. Other parameters information is shown in Table I. We use test accuracy to measure the learning performance. We compare our proposed AdapTop-K in SGD with the vanilla Top-K.

Fig. 1 shows the comparison results of the classic Top-K algorithm and our proposed AdapTop-K on the MNIST dataset. Fig. 1a and Fig. 1b show the test accuracy curves and the training loss curves on the MNIST dataset. It shows

---

Algorithm 1 AdapTop-K in Distributed SGD

Input: Maximum iterations number T, learning rate η, initial point $\mathbf{w}_0 \in \mathbb{R}^d$, fixed k value, adjusted scale factor γ, hyper-parameters $\hat{t}$

Output: $\mathbf{w}_t$
1: **for** $t = 0, 1, ...T - 1$ **do**
2:    **On each worker** $i = 1, ..., M$:
3:    Compute stochastic local gradient $\mathbf{g}_t^i$
4:    **if** $t \in [\frac{\hat{t}}{2}, \frac{\hat{t}+T}{2})$ **then**
5:      Set k_t to $k - \gamma k$
6:    **else**
7:      Set k_t to $k + \gamma k$
8:    **end if**
9:    Compress gradient $\mathbf{g}_t^i$ to $\mathcal{C}_{k_t}[\mathbf{g}_t^i]$
10:   Send $\mathcal{C}_{k_t}[\mathbf{g}_t^i]$ to server
11:   Receive $\mathbf{w}_{t+1}$ from server
12:   **On server**:
13:   Collect M compressed gradients $\mathcal{C}_{k_t}[\mathbf{g}_t^i]$ from workers
14:   Aggregation: $\mathcal{C}_{k_t}[\mathbf{g}_t] = \sum_{i=1}^{M} \mathcal{C}_{k_t}[\mathbf{g}_t^i]$
15:   Update global parameters: $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta}{M}\mathcal{C}_{k_t}[\mathbf{g}_t]$
16:   Send $\mathbf{w}_{t+1}$ back to all workers
17: **end for**

Dataset	MNIST	CIFAR-10
Networks	fully-connected network	ResNet18
Model Size	$d = 785$	$d = 1 \times 10^7$
Learning Rate	0.1	0.1
Batch Size	32	32
Workers	8	8
Iterations	3,000	7,000
Compression Ratio	128/256/512	128/256/512
γ	0.5	0.5

TABLE I: Experimental Setting.

how the model performance changes with iterations for several different values of the sparsification factor (128 or 256). The accuracy of the original distributed SGD reaches 98.02%. In Fig. 1a, the AdapTop-K achieves 97.03% accuracy which is better than 96.64% from Top-K. In Fig. 1b, the AdapTop-K achieves 96.21% accuracy which is higher than 95.41% from Top-K. The curve corresponding to the AdapTop-K achieves better performance than fixed Top-K compression when the compression ratios ($\frac{d}{k}$) are 128 and 256, respectively.

Similarly, Fig. 2 shows the comparison results of the fixed Top-K and our proposed AdapTop-K on CIFAR-10 dataset. Fig. 2a and Fig. 2b show the test accuracy curves and the training loss curves. It shows how the model performance changes with iterations for several different values of the sparsification factor (128 or 256). The accuracy of the original distributed SGD reaches 90.92%. In Fig. 2a, the AdapTop-K achieves 82.11% accuracy which is better than 81.36% from Top-K. In Fig. 2b, the AdapTop-K achieves 80.31% accuracy which is higher than 79.30% from Top-K. The curve corresponding to the AdapTop-K achieves better performance than fixed Top-K compression when the compression ratios ($\frac{d}{k}$) are 128 and 256, respectively. We keep the communication
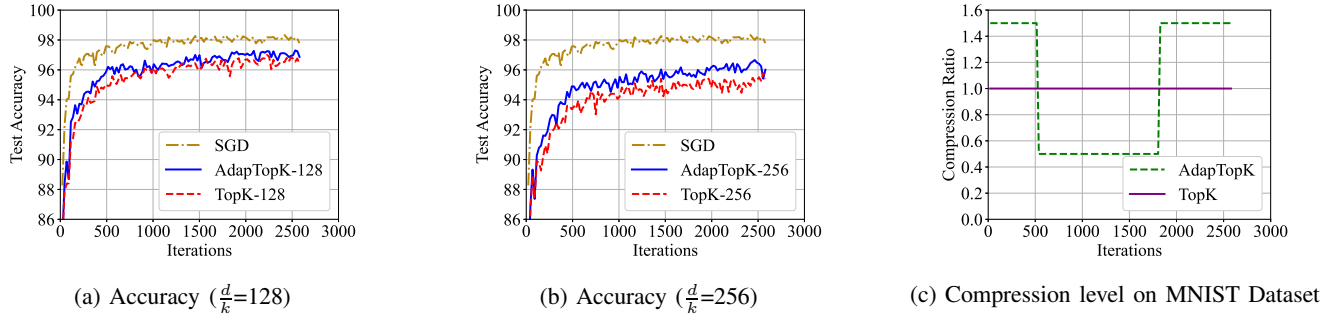
(a) Accuracy ($\frac{d}{k}$=128) (b) Accuracy ($\frac{d}{k}$=256) (c) Compression level on MNIST Dataset

Fig. 1: Evaluation results of different methods on MNIST Dataset.

(a) Accuracy ($\frac{d}{k}$=128) (b) Accuracy ($\frac{d}{k}$=256) (c) Compression level on CIFAR-10 Dataset

Fig. 2: Evaluation results of different methods on CIFAR-10.

(a) Accuracy with ec ($\frac{d}{k}$=256) (b) Accuracy with ec ($\frac{d}{k}$=512)

Fig. 3: Evaluation with error compensation on MNIST.

(a) Accuracy with ec ($\frac{d}{k}$=256) (b) Accuracy with ec ($\frac{d}{k}$=512)

Fig. 4: Evaluation with error compensation on CIFAR-10.

We can see that AdapTop-K significantly increases the bits assigned at the early stage and the late stage of training and improves the gradient accuracy as the training goes on.

After that, we add the error compensation [7] (abbreviated as ec) in Fig. 3 and Fig. 4 in our experiments, because it is a popular technique to improve the performance of distributed SGD with gradient compression. It shows how the model performance changes with iterations for several different values of the sparsification factor (256 or 512) when we add the error compensation. In these experiments, we use the bigger compression ratios (e.g., 256 and 512) because error compensation may reduce optimization errors in the training process to improve the total performance. Fig. 3 and Fig. 4 show the comparison results of the classic Top-K algorithm and our proposed AdapTop-K (all with error compensation) on MNIST and CIFAR-20 datasets. In Fig. 3a, the AdapTop-K achieves 97.50% accuracy which is higher than 96.71% from Top-K. In Fig. 3b, the AdapTop-K achieves 97.10% accuracy which is better than 96.24% from Top-K. In Fig. 4a, the AdapTop-K achieves 89.18% accuracy which is better than 88.66% from Top-K. In Fig. 4b, the AdapTop-K achieves 88.68% accuracy which is higher than 87.64% from Top-K. The curve corresponding to the AdapTop-K achieves better performance than fixed Top-K compression when the compression ratios ($\frac{d}{k}$) are 256 and 512, respectively. The results show that the AdapTop-K algorithm with error compensation achieves better performance under stable communication cost. Overall, the evaluation results demonstrate that the AdapTop-K outperforms the baselines.

cost of the AdapTop-K stable compared with the classic Top-K in the total training process. It can be seen that our adaptive sparsification strategy can effectively improve the convergence rate and model performance with the pure Top-K algorithm. Fig. 1c and Fig. 2c both show the gradient sparsification level in the training process of AdapTop-K on different datasets.

VI. Conclusion

This paper proposes AdapTop-K, a novel adaptive gradient sparsification strategy for distributed SGD. The proposed method adjusts the sparsification levels adaptively by considering the gradient and the current iteration step. The experimental results for image classification show that AdapTop-K is superior to the state-of-the-art gradient compression methods in reducing the communication cost.

Acknowledgment

VII. Appendix

A. Proof for Theorem 1

Using Eq. (8) and Assumption 1, we get:

$$\mathbb{E}[F(\mathbf{w}_{t+1})] \leq F(\mathbf{w}_t) - \eta\langle\nabla F(\mathbf{w}_t),\mathcal{C}(\mathbf{g}_t)\rangle + \frac{\eta^2 L}{2}\mathbb{E}\|\mathcal{C}(\mathbf{g}_t)\|^2$$

(use $\mathbb{E}\|\mathcal{C}(\mathbf{g}_t)\|^2 = \mathbb{E}\|\mathcal{C}(\mathbf{g}_t) - [\mathcal{C}(\mathbf{g}_t) - (\mathbf{g}_t - \nabla F(\mathbf{w}_t))]\|^2 +$

$\mathbb{E}\|\mathbb{E}[\mathcal{C}(\mathbf{g}_t) - (\mathbf{g}_t - \nabla F(\mathbf{w}_t))]\|^2$ and Assumption 3)

$$\leq F(\mathbf{w}_t) - \eta\langle\nabla F(\mathbf{w}_t), \mathbb{E}[\mathcal{C}(\mathbf{g}_t) - (\mathbf{g}_t - \nabla F(\mathbf{w}_t))]\rangle$$
$$+ \frac{\eta^2 L}{2}(\sigma^2 + \mathbb{E}\|\mathcal{C}(\mathbf{g}_t) - (\mathbf{g}_t - \nabla F(\mathbf{w}_t))\|^2)$$

$$\leq F(\mathbf{w}_t) + \frac{\eta}{2}(\mathbb{E}\|\mathcal{C}(\mathbf{g}_t) - (\mathbf{g}_t - \nabla F(\mathbf{w}_t))\|^2)$$

$$- 2\langle\nabla F(\mathbf{w}_t), \mathbb{E}[\mathcal{C}(\mathbf{g}_t) - (\mathbf{g}_t - \nabla F(\mathbf{w}_t))]\rangle) + \frac{\eta^2 L}{2}\sigma^2 \ (\eta \leq \frac{1}{L})$$

(from $\mathbb{E}\|\nabla F(\mathbf{w}_t) + \mathcal{C}(\mathbf{g}_t) - \mathbf{g}_t\|^2 = \mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2 +$

$\mathbb{E}\|\mathcal{C}(\mathbf{g}_t) - \mathbf{g}_t\|^2 + 2\mathbb{E}\langle\nabla F(\mathbf{w}_t),\mathcal{C}(\mathbf{g}_t) - \mathbf{g}_t\rangle)$

$$\leq F(\mathbf{w}_t) + \frac{\eta}{2}(\mathbb{E}\|\mathcal{C}(\mathbf{g}_t) - \mathbf{g}_t\|^2 - \mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2) + \frac{\eta^2 L}{2}\sigma^2$$

(from Eq. (12) and assume that $k_t = k + n_t$, we have:

$$\mathbb{E}\|b_t(\mathbf{w})\|^2 = \mathbb{E}\|\mathbf{g}_t - \mathcal{C}(\mathbf{g}_t)\|^2 \leq \mathbb{E}[(1 - \frac{k_t}{d})\|\mathbf{g}_t\|^2]$$

$$\leq \mathbb{E}[(1 - \frac{k}{d})\|\nabla F(\mathbf{w}_t)\|^2 + (1 - \frac{k}{d})\sigma^2 - \frac{n_t}{d}\|\mathbf{g}_t\|^2],$$

then put this equation back to our above derivation)

$$\leq F(\mathbf{w}_t) - \frac{\eta k}{2d}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{\eta}{2}(1 - \frac{k}{d} + \eta L)\sigma^2 - \frac{\eta n_t}{2d}\|\mathbf{g}_t\|^2.$$

Therefore, we use Assumption 2 and get convergence rate as

$$\mathbb{E}[F(\mathbf{w}_{t+1})] - F^* \leq (1 - \frac{\eta k\mu}{d})(\mathbb{E}(F(\mathbf{w}_t) - F^*)$$
$$+ \frac{\eta}{2}(1 - \frac{k}{d} + \eta L)\sigma^2 - \frac{\eta n_t}{2d}\|\mathbf{g}_t\|^2.$$

After recursion and simplification, we get:

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq (1 - \frac{\eta\mu}{d}k)^T[\mathbb{E}[F(\mathbf{w}_0)] - F^*]$$
$$+ \frac{d}{2k\mu}(1 - \frac{k}{d} + \eta L)\sigma^2[1 - (1 - \frac{\eta\mu}{d}k)^T]$$
$$- \sum_{t=0}^{T-1}[(\frac{\eta n_t}{2d}\|\mathbf{g}_t\|^2)(1 - \frac{\eta\mu}{d}k)^{T-1-t}].$$

References

[1] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang *et al.*, "Large scale distributed deep networks," in *in Proceedings of Conference in Neural Information Processing Systems (NeurIPS)*, 2012.

[2] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 40, no. 1, pp. 342–358, 2021.

[3] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[4] A. Ajalloeian and S. U. Stich, "On the convergence of SGD with biased gradients," in *Proceedings of Workshop in International Conference on Machine Learning (ICML)*, 2020.

[5] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[6] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[7] J. Wu, W. Huang, J. Huang, and T. Zhang, "Error compensated quantized SGD and its applications to large-scale distributed optimization," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 5325–5333.

[8] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," in *Proceedings of Conference in Neural Information Processing Systems (NeurIPS)*, 2018.

[9] G. Yan, T. Li, S.-L. Huang, T. Lan, and L. Song, "AC-SGD: Adaptively Compressed SGD for Communication-Efficient Distributed Learning," *IEEE Journal on Selected Areas in Communications (JSAC)*, vol. 40, no. 9, pp. 2678–2693, 2022.

[10] H. Karimi, J. Nutini, and M. Schmidt, "Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Lojasiewicz Condition," in *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2016, pp. 795–811.

[11] P. Han, S. Wang, and K. K. Leung, "Adaptive gradient sparsification for efficient federated learning: An online learning approach," in *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2020, pp. 300–310.

[12] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang, "ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 4035–4043.

[13] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Proceedings of Conference in Neural Information Processing Systems (NeurIPS)*, 2017.

[14] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik, "Distributed learning with compressed gradient differences," *arXiv preprint arXiv:1901.09269*, 2019.

[15] M. Ruan, G. Yan, Y. Xiao, L. Song, and W. Xu, "Adaptive Top-K in SGD for Communication-Efficient Distributed Learning," *arXiv preprint arXiv:2210.13532*, 2022.

This figure "fig1.png" is available in "png" format from:

http://arxiv.org/ps/2210.13532v2