Data-Aware Gradient Compression for FL in Communication-Constrained Mobile Computing

Rongwei Lu, *Student Member, IEEE*, Yutong Jiang, Yinan Mao, *Student Member, IEEE*, Chen Tang, Bin Chen, *Member, IEEE*, Laizhong Cui, *Senior Member, IEEE*, Zhi Wang, *Senior Member, IEEE*

***Abstract*—Federated Learning (FL) in mobile environments faces significant communication bottlenecks. Gradient compression has proven as an effective solution to this issue, offering substantial benefits in environments with limited bandwidth and metered data. Yet, it encounters severe performance drops in non-IID environments due to a one-size-fits-all compression approach, which does not account for the varying data volumes across workers. Assigning varying compression ratios to workers with distinct data distributions and volumes is therefore a promising solution. This work derives the convergence rate of distributed SGD with non-uniform compression, which reveals the intricate relationship between model convergence and the compression ratios applied to individual workers. Accordingly, we frame the relative compression ratio assignment as an n-variable chi-squared nonlinear optimization problem, constrained by a limited communication budget. We propose DAGC-R, which assigns conservative compression to workers handling larger data volumes. Recognizing the computational limitations of mobile devices, we propose the DAGC-A, which is computationally less demanding and enhances the robustness of compression in non-IID scenarios. Our experiments confirm that the DAGC-R and DAGC-A can speed up the training speed by up to 25.43% and 16.65% compared to the uniform compression respectively, when dealing with highly imbalanced data volume distribution and restricted communication.**

***Index Terms*—Federated Learning, Non-IID, Data-Aware Gradient Compression**

I. Introduction

With the widespread use of mobile devices and the progress in machine learning [1], there is a burgeoning interest in distributed machine learning (DML) using these portable platforms. FL is an increasingly important DML framework that addresses the critical need for data privacy in model training across multiple mobile devices. Despite its potential, this paradigm faces significant challenges due to communication bottlenecks, particularly as the number of devices scales up. Gradient compression has been identified as an effective solution to this challenge, by reducing the communication volume and offering a cost-effective option in bandwidth-limited and per-traffic billing mobile environments.

However, in non-IID scenarios, the performance of lossy gradient compression algorithms worsens, leading to poorer model convergence when contrasted with IID datasets [2], [3]. For instance, the same gradient compression algorithm [4] (with the hyperparameter set to 10%) reduces the accuracy by only 0.7% compared to the bulk synchronous parallel (BSP) [5] as a baseline in IID scenarios. However, this gap widens significantly to a 10.4% decrease in accuracy under non-IID conditions [3]. The reason for the drop in accuracy is that it uses the same aggressive gradient compression ratios for different workers, ignoring the fact that different workers usually have different data volumes and distributions [6]–[8]. In real-world non-IID scenarios, mobile devices are geographically distributed, and each worker collects its own dataset, resulting in skewed data distributions and volumes [3], [9]. To illustrate, within the Flickr-mammal dataset (denoted as Flickr in the following) [3], the worker with the largest data volume has 78% more samples than the worker with the second largest data volume (divided by subcontinent). Similarly, in the Google Landmark dataset v2 [10], the difference is even greater. The worker with the most data has at least 213% more images than its peers, with the division based on continent.

To facilitate our discussion, we introduce the terms *large workers* and *small workers* to refer to workers with large and small amounts of data, respectively. We use *worker size* to denote the quantity of local data samples. Current gradient compression algorithms often neglect the variation in *worker size*. Even in studies like SkewScout [3], which suggests adaptive methods to adjust gradient compression ratios based on data distribution differences, *large workers* are subjected to the same stringent compression strategy as their smaller counterparts. This leads to the loss of vital information that could otherwise help the model converge faster.

Our empirical analysis reveals important findings for developing a data-volume-aware gradient compression method. Primarily, we find that a *one-size-fits-all* compression strategy falls short in non-IID settings with communication constraints, as workers with diverse data volumes and distributions require tailored compression ratios. Furthermore, a compression strategy that assigns higher compression ratios[1] to *large workers* can reduce the number of training iterations needed to achieve

Rongwei Lu, Yutong Jiang and Zhi Wang are with Shenzhen International Graduate School, Tsinghua University, Shenzhen 518000, China (e-mail: {lurw24, jiang-yt24, wangzhi}@mails.tsinghua.edu.cn).

Yinan Mao is with the AI Cloud Business Group, Baidu Incorporated, Beijing 100085, China (e-mail: maoyinan01@baidu.com).

Chen Tang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: genprtung@gmail.com).

Bin Chen is with the Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: chenbin2021@hit.edu.cn).

Laizhong Cui is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: cuilz@szu.edu.cn).

[1]In this work, the compression ratio is defined as the ratio of the compressed data divided by the uncompressed data, referring to the gradient compression part of Sec. II in [11].

the same accuracy, in contrast to a uniform strategy. Leveraging these findings, we advocate for an adaptive algorithm that adjusts compression ratios based on the *worker size*, thereby optimizing the balance between compression efficiency and communication overhead.

The relative compressor is one of the most popular types of gradient sparse compressors, known for achieving efficient compression compared to other compressors like quantization [12] or low-rank [13]. In relative compressors, we can directly determine the compression ratio, and the compressor will transmit the corresponding number of elements. The main technical challenge in designing the adaptive relative compressor is: *given a fixed total communication budget, how to determine the compression ratio for each worker?* This budget constraint implies that, during each iteration, the total traffic sent from all clients to the server must not exceed a fixed limitation[2]. To tackle this issue, firstly we derive the convergence rate of distributed SGD with error feedback and different gradient compression ratios (denoted as non-uniform D-EF-SGD) with the relative compressors like Top-k [14], [16]. Our results demonstrate that in communication-constrained non-IID scenarios, there is a key term that directly affects the convergence rate of non-uniform D-EF-SGD with the relative compressors. Minimizing this term not only speeds up its convergence rate but also makes the algorithm robust to non-IID scenarios. To find a proper assignment of relative compression ratios, we formulate an n-variable chi-square nonlinear optimization problem with one constraint, which can be solved by the Lagrange multiplier method and finding the minimum of a one-dimensional function.

While the process of deriving the optimal compression ratios introduces negligible additional computational cost, the relative compressor has high computational overhead, making it less suitable for resource-constrained mobile environments. In contrast, the absolute compressor [17], which transmits elements with absolute values higher than a fixed threshold, does not allow for precise control over the compression ratio, and it is more efficient than the SOTA relative compressor Top-k. The compression cost of Top-k can be up to hundreds of times greater than that of the absolute compressor[3] for two main reasons: (1) the absolute compressor has a lower computational complexity of $\mathcal{O}(d)$ compared to the $\mathcal{O}(d \log k)$ complexity of Top-k selection, where d is the number of the model parameters; (2) Top-k selection performs poorly on accelerators such as GPUs. Therefore, we solve the technical challenge under the absolute compressor similarly. We derive the convergence rate under the absolute compressor and reveal the key factor in the rate. We then formulate the task identifying the optimal threshold as a symmetric optimization problem, solvable using the Lagrange multiplier method. By minimizing the key factor, we can obtain a faster convergence rate without introducing additional hyperparameters.

Based on these analyses, we propose DAGC (illustrated in Fig. 1), a low-cost data-aware adaptive gradient compression algorithm that strategically allocates different compression ratios depending on the *worker size*. DAGC is designed for the realistic non-IID scenario, where the local datasets in the mobile devices are collected based on the location and there is a significant difference in the *worker size*. DAGC is composed of DAGC-R for relative compressors and DAGC-A for absolute compressors. We define the number of workers as n, the compression ratio of the relative compressor, the threshold of the absolute compressor, and the training weight of the i-th worker as δ_i, λ_i and p_i respectively. We have $\frac{\delta_i}{\delta_j} \approx (\frac{p_i}{p_j})^{2/3}$ in DAGC-R and $\frac{\lambda_i}{\lambda_j} = (\frac{p_j}{p_i})^{2/3}$ in DAGC-A. Both align with the conclusion from the measurements, indicating that higher compression ratios should be given to workers with larger p_i. The time complexity of DAGC-R (as well as DAGC-A) to find the optimal δ_i (λ_i) is $\mathcal{O}(n)$ ($\mathcal{O}(1)$).

In a word, the contributions of our work are summarized as below:

- We experimentally reveal that setting higher compression ratios for *large workers* converges faster than the uniform compression under the fixed and limited communication volume.
- We generalize the convergence analysis of D-EF-SGD with both relative [19] and absolute compressors [17] to the context of non-uniform compression, where nodes are endowed with different compression ratios and training weights. Under communication-constrained non-IID scenarios, we show the key factor $\frac{\sum_{i=1}^{n} \frac{p_i}{\sqrt{\delta_i}}}{\sqrt{\delta_{min}}}$ for the relative compressor and $\sum_{i=1}^{n} p_i^2 \lambda_i^2$ for the absolute compressor.
- We propose two novel adaptive compression algorithms, DAGC-R and DAGC-A, designed for optimizing compression rate allocation in relative and absolute compressors, respectively. DAGC-R is developed by solving an n-variable chi-square nonlinear asymmetric optimization problem with a communication constraint. In DAGC-R, it has $\frac{\delta_i}{\delta_j} \approx (\frac{p_i}{p_j})^{2/3}$. Similarly, DAGC-A is developed by solving an n-variable chi-square nonlinear symmetrical optimization problem subject to the same constraint and has $\frac{\lambda_i}{\lambda_j} = (\frac{p_j}{p_i})^{2/3}$. DAGC-R and DAGC-A converge the fastest in communication-constrained non-IID scenarios.
- We employ DAGC-R and DAGC-A in both the real-world non-IID and artificially partitioned non-IID datasets. The experimental results confirm the correctness of our theory and show that our design can save up to 25.43% of iterations to converge to the same accuracy compared to the uniform compression.

II. Preliminaries

We specifically concentrate on FL using the D-EF-SGD algorithms alongside sparsification compressors. We aim to overcome the challenges posed by non-IID scenarios with constraint communication in FL. To provide a comprehensive understanding, we will briefly delve into the optimization

[2]We focus solely on compressing the communication from clients to the server, also known as upstream communication. This is because communication bottlenecks in the PS architecture typically occur in the upstream direction [14], [15]. The downstream communication optimization is beyond the scope of this work.

[3]According to Fig. 15(d) in the appendix of the work [18], the compression time of Top-k is hundreds of times that of SIDCo, and the absolute compressor is more effective than SIDCo.

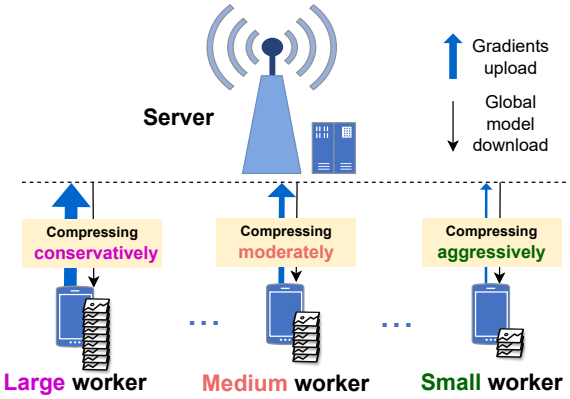


Fig. 1: High-level design of DAGC. DAGC sets different compression ratios to workers depending on the *worker size*. *Large workers* (*i.e.*, the workers with large data volumes and similarly to *small* and *medium workers*) are assigned conservative compression ratios, and *small workers* adopt aggressive compression ratios.

problem of FL, communication-constraint non-IID scenarios, and gradient compression, including the properties of relative and absolute compressors.

The optimization problem of FL: The distribution problem is taken into consideration in this project.

$$f^* := \min_{\mathbf{x}} \left[f(\mathbf{x}) := \sum_{i=1}^{n} p_i f_i(\mathbf{x}) \right],$$

In this research, we consider the objective function f, which is divided into n terms $f_i, i \in [n]$. p_i is the training weight of the worker i. Each p_i value should be greater than or equal to 0, and the summation of all p_i values equals 1. To facilitate explanation, we assign sequence numbers to the workers based on their training weights[4], with $p_i \geq p_{i+1}, \forall i \in [n-1]$.

Communication-constraint non-IID scenarios: In communication-constrained non-IID scenarios, on the data side, data on each node is isolated from others due to privacy and data security. This results in different data volumes and the classification of datasets stored by distributed workers. On the communication side, the bandwidth between mobile devices and the server is limited, often requiring communication across WAN, which incurs high costs. Therefore, achieving efficient training under these constraints can not only speed up the training process but also significantly reduce the expensive WAN communication costs. In this work, communication-constrained cases mean that the average compression ratio is less than or equal to 0.1%, and having the same communication budget means that the sum of the communication volume transmitted from all workers to the server per iteration is the same across different compression strategies.

Gradient compression: Unlike compressing the model [20], [21] for inference speedup, gradient compression focuses on compressing the gradient to reduce the traffic volume during the communication phase. Concerning compression tactics [22], compression approach can be categorized into (i) quantization [23]–[25], which converts high precision to low precision, consequently diminishing the number of transmitted bits; (ii) sparsification [17], [26], which retains solely certain elements of the gradient while assigning 0 to others; (iii) low-rank [13], which breaks down the gradient matrix to acquire multiple low-rank matrices.

According to whether the compression error (i.e., the norm of the difference between the parameter before and after compression) is independent of the compressed parameter $\mathbf{x}$, the compressors can be categorized as absolute or relative compressors [17]:

• *The relative compressor:* A relative compressor is one whose compression error is dependent on the compressed parameter. Classic relative compressors include Top-k and Random-k [11], [22]. The input parameter of a relative compressor is the compression ratio δ. The larger δ is, the more conservative the compression is. We represent the relative compressor as $\mathbf{C}_\delta$. By definition, $\mathbf{C}_\delta$ is a mapping that has the property $\mathbb{R}^d \to \mathbb{R}^d$:

$$\mathbb{E}_{\mathbf{C}_\delta} \|\mathbf{C}_\delta(\mathbf{x}) - \mathbf{x}\|^2 \leq (1-\delta)\|\mathbf{x}\|^2.$$

• *The absolute compressor:* The compressor error of the absolute compressor is independent of $\mathbf{x}$. The representative compressor is the hard-threshold algorithm [17]. The input parameter of the absolute compressor is the threshold λ. The higher λ indicates more aggressive compression. We denote the absolute compressor as $\mathbf{C}_\lambda$ and definitionally, $\mathbf{C}_\lambda$ is a mapping: $\mathbb{R}^d \to \mathbb{R}^d$, having the property:

$$\mathbb{E}_{\mathbf{C}_\lambda} \|\mathbf{C}_\lambda(\mathbf{x}) - \mathbf{x}\|^2 \leq d\lambda^2,$$

where d is the number of parameters of the model.

Distributed SGD with error-feedback and sparsification compressors (D-EF-SGD): D-EF-SGD [27] and Distributed Quantized SGD (D-QSGD) [15], [28] are two of the most dominant compressors in FL. In communication-constrained non-IID scenarios, D-EF-SGD has two advantages: First, it achieves lower compression ratios[5] (*i.e.*, transmits less information), which is favorable for resource-constrained scenarios; Second, D-EF-SGD has a lower dependence on the heterogeneity of the data, and thus is more suitable for highly skewed non-IID scenarios [30]. For these reasons, we focus on D-EF-SGD rather than D-QSGD.

III. Motivating Examples

In this section, we aim to validate the following two points through a series of motivating experiments: 1) In communication-constrained environments, a compression strategy with different compression ratios can achieve faster convergence compared to uniform compression. 2) A strategy that sets higher compression ratios for *large workers* usually achieves faster convergence than those for *small workers*,

[4]The order of the training weight remains the same in the following paper.

[5] $\frac{1}{32}$ is the tiniest compression ratio of D-QSGD [12], [29], nearly 3.4%, while less than 0.1% denotes the compression ratio of sparsification constrain [19]. Meanwhile, the compression ratio of D-QSGD is discrete. It emphasizes to adaptively adjust more complicated ratios. This issue is also discussed in the design of DC2 [11].

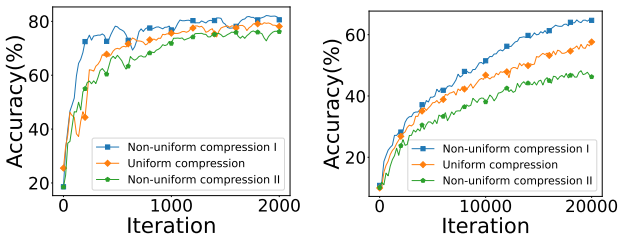


Fig. 2: The accuracy curves (Accuracy vs. Iterations) of Logistic@FMNIST (a) and LSTM@SCs (b) using different relative compression strategies. In scheme I (as well as scheme II), large workers are set lower (higher) compression ratios. The uniform compression is a one-size-fits-all strategy. Among these three strategies, non-uniform compression scheme I exhibits optimal performance.

reducing the number of iterations by up to 69.70%. To verify this conclusion, two non-uniform compression strategies are used in this paper: 1) *Non-uniform compression I give higher compression ratios to large workers* and lower compression ratios to small workers; 2) *Non-uniform compression II gives lower compression ratios to large workers* and higher compression ratios to small workers.

To provide empirical evidence for our research, we performed two tasks: Fashion-MNIST [31] (denoted as FMNIST) using Logistic and Speech Commands [32] (denoted as SCs) using LSTM. This ensures that our experimental findings are generalizable. To gain a better understanding of the results, we divide the workers into two categories: *large workers* and *small workers*. Our setup consists of 11 workers, with one *large worker* with datasets accounting for p_{large} of the global dataset, and 10 *small workers*, each with datasets accounting for $p_{small} = \frac{1-p_{large}}{10}$, $p_{large} = 50\%$. In these experiments, we impose communication constraints, specifically an average compression ratio of the aggressive ratio $\delta_{min} = 0.1\%$ as described in the appendix experiments [17], while maintaining consistent total communication volume (i.e., fixing $\sum_{i=1}^{n} \delta_i$). To simulate the non-IID scenario, the data points are generated based on the Dirichlet distribution, with the parameter set to 0.5.

Logistic on FMNIST: It can be observed from Fig. 2a that the initial stages witness faster progress in non-uniform compression I ($\delta_1 = 1\%$, and $\delta_i = 0.01\%$, $i \in [2, n]$) when compared to two alternative strategies. In order to achieve accuracies of 50%, 60%, 70%, and 80%, the iteration reduction for scheme I are 54.55%, 41.67%, 62.96%, and 62.50%, respectively, compared to the uniform compression ($\delta_i = 0.1\%$, $i \in [1, n]$). It is noteworthy that under such circumstances, the performance of scheme I is optimized, whereas non-uniform compression II ($\delta_1 = 0.01\%$, and $\delta_i = 0.11\%$, $i \in [2, n]$) experiences a relatively weaker performance.

LSTM on SCs: Fig. 2b demonstrates that, in the context of fixed communication volume, the model utilizing the non-uniform compression strategy I exhibits the fastest convergence. To attain accuracy levels of 50%, 55%, 60%, and 65%, the non-uniform compression strategy I diminishes the training time by 7.95%, 19.17%, 19.75%, and 27.06%, respectively, as compared to the uniform strategy with $p_{large} = 50\%$. The convergence rate of scheme II is significantly worse than the other two strategies.

In conclusion, the aforementioned empirical findings affirm the inadequacy of uniform compression as an optimal strategy when confronted with discrepancies in worker size. Conversely, the proposed approach, which assigns higher compression ratios to *large workers*, stands as an efficacious means of enhancing training speed.

TABLE I: Notation list.

Notation	Description
n	the number of workers
p_i	the training weight of the i-th node
δ_i/λ_i	the compression ratio/threshold of node i under the relative/absolute compressor
$\mathbf{C}_{\delta_i}/\mathbf{C}_{\lambda_i}$	the relative/absolute compressor and the hyper-parameter is δ_i/λ_i
$\mathbf{x}_t$	the global model in the t-th iteration
γ	learning rate
$\mathbf{g}^i(\mathbf{x}_t)$	the stochastic gradient of $\mathbf{x}_t$
$\mathbf{e}_t^i$	the local error term of node i in the t-th iteration
$\hat{\Delta}_t^i$	the compressed gradients transferred to the server from node i in the t-th iteration
$\hat{\Delta}_t^i$	the compressed gradients transferred to the server from node i in the t-th iteration
ζ_i	the distance of node i from the local distribution to the global distribution
σ^2	the upper variance bound of the noise of gradients

IV. Theoretical Analysis within the Relative Compressor

In this section, our primary focus is to address the central issue of this research: **Given the total amount of communication, how can we theoretically ascertain the best compression ratio for each worker using the relative compressor?**

Initially, we outline the convergence speed of non-uniform D-EF-SGD with the relative compressor, taking into account the non-convex, convex, and strongly convex scenarios. (Theorems in Sec. IV-B and the proof in Appendix)

Subsequently, we represent this challenge as an n-variable chi-square nonlinear asymmetrical optimization problem that comes with one constraint. (**Theorem 4** in Sec. IV-D)

Lastly, we propose our design DAGC-R in Sec. IV-E.

We tabulate the notations in Table I. Additionally, we showcase the pseudo-code of non-uniform D-EF-SGD in Algorithm 1.

Algorithm 1: Non-uniform D-EF-SGD

Input: n, γ, p_i, compressor $\mathbf{C}$, compression parameters $\delta_1, \ldots, \delta_n$ or $\lambda_1, \ldots, \lambda_n$, initial parameters $\mathbf{x}_0$, initial local error $\mathbf{e}_0^i = \mathbf{0}_d$

Output: $\mathbf{x}_T$

for $t = 0, \ldots, T-1$ **do**
  /* Worker does */
  for $i = 1, \ldots, n$ **do**
    Receive $\mathbf{x}_t$ from the server;
    $\mathbf{g}_t^i := \mathbf{g}^i(\mathbf{x}_t)$;
    if ***C*** *is the relative compressor* **then**
      $\hat{\Delta}_t^i := \mathbf{C}_{\delta_i}(\mathbf{e}_t^i + \mathbf{g}_t^i)$;
    else
      $\hat{\Delta}_t^i := \mathbf{C}_{\lambda_i}(\mathbf{e}_t^i + \mathbf{g}_t^i)$;
    end
    $\mathbf{e}_{t+1}^i := \mathbf{e}_t^i + \mathbf{g}_t^i - \hat{\Delta}_t^i$;
    Upload $\hat{\Delta}_t^i$;
  end
  /* Server does */
  Gather all $\hat{\Delta}_t^i$;
  $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \sum_{i=1}^n p_i \hat{\Delta}_t^i$;
  Send $\mathbf{x}_{t+1}$ to all workers
end
Return $\mathbf{x}_T$;

A. Assumptions

We assume functions are L-smooth with the gradient noise of SGD presumed to exhibit zero mean and a variance of σ^2. We gauge data heterogeneity using constants $\zeta_i^2 > 0$ and $Z^2 \geq 1$, which cap the variance among the n workers. In **Theorem 2, 3**, we posit that objective functions are μ-strongly convex. Detailed assumptions are shown below.

Assumption 1 (L-smoothness). We assume L-smoothness of $f_i, i \in [n]$, that is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \leq L\|\mathbf{y} - \mathbf{x}\|. \tag{1}$$

Assumption 2 (μ-strongly convexity). We assume μ-strong convexity of $f_i, i \in [n]$, that is, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f(\mathbf{x}) - (\mathbf{y}) \geq \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2. \tag{2}$$

Assumption 3 (Bounded gradient noise). We assume that we have access to stochastic gradient oracles $g^i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$ for each $f_i, i \in [n]$. For simplicity we only consider the instructive case of uniformly bounded noise for all $\mathbf{x} \in \mathbb{R}^d, i \in [n]$:

$$g^i(\mathbf{x}) = \nabla f_i(\mathbf{x}) + \boldsymbol{\xi}^i, \quad \mathbb{E}_{\boldsymbol{\xi}^i}\boldsymbol{\xi}^i = \mathbf{0}_d, \quad \mathbb{E}_{\boldsymbol{\xi}^i}\|\boldsymbol{\xi}^i\|^2 \leq \sigma^2. \tag{3}$$

Assumption 4 (Measurement of data heterogeneity). We measure data dissimilarity by constants $\zeta_i^2 \geq 0, Z^2 \geq 1$ that bound the variance across the n nodes. We have:

$$\|\nabla f_i(\mathbf{x})\|^2 \leq \zeta_i^2 + Z^2\|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d, i \in [n].$$

This is similar to the assumption in previous work [30], [33].

B. Convergence rate of non-uniform D-EF-SGD with the relative compressor

Theorem 1 (Non-convex convergence rate of non-uniform D-EF-SGD with the relative compressor)**.** *Consider a function f, which maps from $\mathbb{R}^d$ to $\mathbb{R}$, and is L-consistent. We can find a learning rate γ such that $\gamma \leq \frac{1}{4LZ}\frac{\delta_{min}}{\sqrt{nC_Z}}$, where $C_Z = \sum_{i=1}^n \frac{\delta_{min}}{\delta_i} p_i^2$. This means that the number of*

$$\mathcal{O}\Big(\frac{\sigma^2 \sum_{i=1}^n p_i^2}{\epsilon^2} + \frac{\sqrt{n}(\zeta \sum_{i=1}^n \frac{p_i}{\sqrt{\delta_i}} + \sigma\sqrt{\sum_{i=1}^n p_i^2})}{\epsilon^{3/2}\sqrt{\delta_{min}}} + \frac{\sqrt{n}Z \sum_{i=1}^n \frac{p_i}{\sqrt{\delta_i}}}{\epsilon\sqrt{\delta_{min}}}\Big) \cdot LF_0 \tag{4}$$

iterations of non-uniform D-EF-SGD with the relative compressor ensures $\mathbb{E}f(\boldsymbol{x}_{final}) - f^* \leq \epsilon$, where F_0 is at least $f(\boldsymbol{x}_0) - f^*$, and $\boldsymbol{x}_{final} = \boldsymbol{x}_t$ refers to a version $\boldsymbol{x}_t$ from the set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{T-1}\}$, picked randomly.

Theorem 2 (Convex convergence rate of non-uniform D-E-F-SGD with the relative compressor, *i.e.*, $\mu = 0$)**.** *Consider a function f, which maps from $\mathbb{R}^d$ to $\mathbb{R}$, and is L-consistent and μ-convex. We can find a learning rate γ such that $\gamma \leq \frac{1}{14LZ}\frac{\delta_{min}}{\sqrt{nC_Z}}$, where $C_Z = \sum_{i=1}^n \frac{\delta_{min}}{\delta_i} p_i^2$. This means that the number of*

$$\mathcal{O}\Big(\frac{\sigma^2 \sum_{i=1}^n p_i^2}{\epsilon^2} + \frac{\sqrt{nL}(\zeta \sum_{i=1}^n \frac{p_i}{\sqrt{\delta_i}} + \sigma\sqrt{\sum_{i=1}^n p_i^2})}{\epsilon^{3/2}\sqrt{\delta_{min}}} + \frac{\sqrt{n}LZ \sum_{i=1}^n \frac{p_i}{\sqrt{\delta_i}}}{\epsilon\sqrt{\delta_{min}}}\Big) \cdot R_0^2 \tag{5}$$

iterations of non-uniform D-EF-SGD with the relative compressor ensures $\mathbb{E}f(\boldsymbol{x}_{final}) - f^* \leq \epsilon$, where R_0^2 is at least $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|$, and $\boldsymbol{x}_{final} = \boldsymbol{x}_t$ refers to a version $\boldsymbol{x}_t$ from the set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{T-1}\}$, picked randomly.

Theorem 3 (Strong convex convergence rate of non-uniform D-EF-SGD with the relative compressor, *i.e.*, $\mu > 0$)**.** *Consider a function f, which maps from $\mathbb{R}^d$ to $\mathbb{R}$, and is L-consistent and μ-convex. We can find a learning rate γ such that $\gamma \leq \frac{1}{14LZ}\frac{\delta_{min}}{\sqrt{nC_Z}}$, where $C_Z = \sum_{i=1}^n \frac{\delta_{min}}{\delta_i} p_i^2$. This means that the number of*

$$\tilde{\mathcal{O}}\Big(\frac{\sigma^2 \sum_{i=1}^n p_i^2}{\mu\epsilon} + \frac{\sqrt{nL}(\zeta \sum_{i=1}^n \frac{p_i}{\sqrt{\delta_i}} + \sigma\sqrt{\sum_{i=1}^n p_i^2})}{\mu\sqrt{\delta_{min}\epsilon}} + \frac{\sqrt{n}LZ \sum_{i=1}^n \frac{p_i}{\sqrt{\delta_i}}}{\mu\sqrt{\delta_{min}}}\Big) \tag{6}$$

iterations of non-uniform D-EF-SGD with the relative compressor ensures $\mathbb{E}f(\boldsymbol{x}_{final}) - f^* \leq \epsilon$, and $\boldsymbol{x}_{final} = \boldsymbol{x}_t$ refers to a version $\boldsymbol{x}_t$ from the sequence $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{T-1}\}$, selected probabilistically based on $(1 - \min\{\frac{\mu\gamma}{2}, \frac{\delta_{min}}{4}\})^{-t}$.

C. Analysis of the convergence rate in communication-constrained non-IID scenarios

To facilitate, we denote the order of magnitude of a as $OM(a)$, i.e., $OM(a) = \lfloor \log_{10} a \rfloor$. In this way, we have the following inequalities:

$$OM(\epsilon) \geq -4, \tag{7}$$

$$OM(n) \geq 1, \tag{8}$$

$$OM(\sigma) \leq OM(\zeta) + 1, \tag{9}$$

$$OM(\delta) \leq -3. \tag{10}$$

Eq. 7 and Eq. 8 are the default experimental settings. The Eq. 7 means that the model converges when $\epsilon = 10^{-4}$ according to the previous work [34]. n is typically taken from 10 to $1,000$, so we get the Eq. 8.

Eq. 9 and Eq. 10 are based on the communication-constrained non-IID scenario discussed in this paper. The Eq. 9 means that the bias due to the non-IID datasets is larger than the noise of the gradient, based on the accuracy degradation of the severe non-IID problem [3], [9]. In other words, if Eq. 9 does not hold, the negative effect of the non-IID problem is negligible, which is beyond the scope of this work. Due to $\delta \leq 0.1\%$ in the communication-constrained case, we get Eq. 10.

Based on the above inequalities, we have

$$\begin{aligned}
&OM(\zeta \sum_{i=1}^{n} \frac{p_i}{\sqrt{\delta_i}}) \\
&= OM(\zeta) + OM(\sum_{i=1}^{n} p_i) - \frac{1}{2} OM(\delta) \\
&\geq (OM(\sigma) - 1) + OM(\sqrt{\sum_{i=1}^{n} p_i^2}) + OM(\sqrt{n}) + \frac{3}{2} \\
&> OM(\sigma \sqrt{\sum_{i=1}^{n} p_i^2}),
\end{aligned}$$

so the convergence in the **Theorem 3** can be written into

$$\begin{aligned}
\tilde{\mathcal{O}}(&\frac{\sigma^2 \sum_{i=1}^{n} p_i^2}{\mu\epsilon} + \frac{\sqrt{nL}(\zeta \sum_{i=1}^{n} \frac{p_i}{\sqrt{\delta_i}})}{\mu\sqrt{\delta_{min}\epsilon}} \\
&+ \frac{\sqrt{n} L Z \sum_{i=1}^{n} \frac{p_i}{\sqrt{\delta_i}}}{\mu\sqrt{\delta_{min}}}).
\end{aligned}$$

For the sake of simplicity, we introduce a function of n variables, $\Phi(\delta_1, \ldots, \delta_n) = \frac{\sum_{i=1}^{n} \frac{p_i}{\sqrt{\delta_i}}}{\sqrt{\delta_{min}}}$. Then the convergence rate of **Theorem 3** can be further simplified as

$$\tilde{\mathcal{O}}(\frac{\sigma^2}{\epsilon} + \frac{\zeta\Phi}{\sqrt{\epsilon}} + \Phi).$$

Similarly, **Theorem 1, 2** can be both simplified as

$$\mathcal{O}(\frac{\sigma^2}{\epsilon^2} + \frac{\zeta\Phi}{\epsilon^{3/2}} + \frac{\Phi}{\epsilon}).$$

The difference between **Theorem 1** and **Theorem 2** is that there is a coefficient $\sqrt{L}$ on the second term and L on the third term in Theorem 2. It does not matter that they can be represented within the same paradigm.

We dissect the convergence speed under two circumstances:
• *Without gradient noise ($\sigma = 0$)*: non-uniform D-EF-SGD with the relative compressor shows sub-linear convergence at a pace of $\mathcal{O}(\frac{\zeta\Phi}{\sqrt{\epsilon}} + \Phi)$ (as well as $\mathcal{O}(\frac{\zeta\Phi}{\epsilon^{3/2}} + \frac{\Phi}{\epsilon})$ in the strongly convex cases (non-convex and convex cases). Minimizing Φ leads to the best convergence.
• *With gradient noise ($\sigma \neq 0$)*: Notably, the second term $\frac{\zeta\Phi}{\sqrt{\epsilon}}$ cannot be ignored, due to $OM(\frac{1}{\epsilon}) \leq OM(\frac{1}{\sqrt{\epsilon\delta}})$. Φ is still a key factor in the convergence rate. Reducing Φ can not only speed up the convergence rate but also mitigate the negative influence of non-IID scenarios.

Overall, minimizing Φ can (1) improve the convergence rate of non-uniform D-EF-SGD with the relative compressor; and (2) make the algorithm robust to non-IID scenarios, in communication-constrained non-IID scenarios regardless of the convexity.

D. The optimal compression ratios

Theorem 4 (Optimal δ_i)**.** *Under the premise that the overall communication traffic is fixed, i.e., $\sum_{i=1}^{n} \delta_i = n\bar{\delta}$, the following equation emerges. We set $\delta_j = \delta_{min} = \min\{\delta_1, \ldots, \delta_n\}$ and denote $P := \sum_{i=1}^{n} p_i^{2/3}$. The minimal value of $\Phi(\delta_1, \ldots, \delta_n)$ can be split into two distinct situations:*
• *j is not equal to n*

$$\Phi(\delta_1, \ldots, \delta_n) \geq \frac{1}{n\bar{\delta}}(p_j(1+Q_j) + p_n Q_j(1+Q_j)), \tag{11}$$

where $Q_j = \frac{P - p_j^{2/3}}{p_n^{2/3}}$. It takes the equal sign when $\delta_j = \frac{n\bar{\delta}}{Q_j+1}$ and $\delta_i = \frac{n\bar{\delta}}{Q_j+1} \frac{p_i^{2/3}}{p_n^{2/3}}, i \neq j$.
• *j is equal to n*

$$\Phi(\delta_1, \ldots, \delta_n) \geq \frac{1}{n\bar{\delta}}(p_j(1+Q_j) + p_{n-1} Q_j(1+Q_j)), \tag{12}$$

where $Q_j = \frac{P - p_j^{2/3}}{p_{n-1}^{2/3}}$. It takes the equal sign when $\delta_j = \frac{n\bar{\delta}}{Q_j+1}$ and $\delta_i = \frac{n\bar{\delta}}{Q_j+1} \frac{p_i^{2/3}}{p_{n-1}^{2/3}}, i \neq j$.

Remark 1. This theorem reduces the problem of finding the optimal δ_i, to get the minimal value of Φ, from a continuous space into a discrete sub-space with only n points. Since traversing the continuous space is impractical, this theorem does simplify the problem, making it solvable.

Remark 2. In IID scenarios, where $p_i = p_j$ and $\zeta_i = 0$ for all $i, j \in [n]$, it follows that $Q_i = n - 1$ for all $i \in [n]$, leading to $\delta_i = \bar{\delta}$ for all $i \in [n]$. This implies that the optimal compression strategy in IID scenarios is the uniform compression.

E. DAGC-R and its implement in FL

The pseudo-code of DAGC under the relative compressor (denoted as DAGC-R) is presented in Algorithm 2. DAGC-R is structured in the following manner: 1) it uses **Theorem 4** to derive n local optimal solutions from a continuous space,

Algorithm 2: DAGC-R

Input: n, p_i, average compression ratio $\bar{\delta}$
Output: $\delta_1, \ldots, \delta_n$

```
/* The value of φ_j represents the
   minimal Φ(δ_1,...,δ_n) when δ_j is the
   least among all δ_i.                  */
```

Set $\phi_{min} = +\infty$;
for $j = n, n-1, \ldots, 1$ **do**
  if $j == n$ **then**
    Calculate ϕ_j using the right side of Eq. (12);
  else
    if $p_j == p_{j-1}$ **then**
      /* When weights are repeated, bypass this computation. */
      $\phi_j = \phi_{min}$;
    else
      Calculate ϕ_j using the right side of Eq. (11);
    end
  end
  if $\phi_j < \phi_{min}$ **then**
    $\phi_{min} = \phi_j$ and adjust the optimal values for $\delta_1, \ldots, \delta_n$;
  end
end
Return $\delta_1, \ldots, \delta_n$;

and 2) it gets the global optimal solution by traversing n local optimal solutions.

This procedure has negligible extra overhead compared to conventional gradient compression algorithms. It involves only one local computation at the server, with a time complexity of $\mathcal{O}(n)$ to derive n local optima, and a single communication step to send the optimal compression ratios to workers, both completed before training. DAGC-R does not require external nodes in a parameter server architecture, the server handles the calculations, while in decentralized training, any node can temporarily act as the server. A larger number of devices does not induce extra cost, so DAGC has good scalability.

V. Theoretical Analysis within the Absolute Compressor

In this section, we address the technical challenge encountered under the absolute compressor, *i.e.*, **Given the limited communication budget, what is the optimal λ_i?**

Initially, we outline the convergence speed of non-uniform D-EF-SGD with absolute compressors (Theorems in Sec. V-A). Then, we formulate the challenge as an n-variable chi-square nonlinear symmetrical optimization problem with the traffic budget. We propose DAGC-A based on **Theorem 8**. The details to prove theorems are shown in the Appendix.

A. Convergence rate of non-uniform D-EF-SGD with the absolute compressor

Theorem 5 (Non-convex convergence rate of non-uniform D-EF-SGD with the absolute compressor)**.** *Consider a function f, which maps from $\mathbb{R}^d$ to $\mathbb{R}$, and is L-consistent. We can find a learning rate γ such that $\gamma \leq \frac{1}{4L}$. Then there exists a stepsize $\gamma \leq \frac{1}{4L}$. This means that the number of*

$$\mathcal{O}\left(\frac{\sigma^2 \sum_{i=1}^n p_i^2}{\epsilon^2} + \frac{\sqrt{nd \sum_{i=1}^n p_i^2 \lambda_i^2}}{\epsilon^{\frac{3}{2}}} + \frac{1}{\epsilon}\right) \cdot L F_0 \tag{13}$$

iterations of non-uniform D-EF-SGD with the absolute compressor ensures $\mathbb{E}f(\boldsymbol{x}_{final}) - f^* \leq \epsilon$, where F_0 is at least $f(\boldsymbol{x}_0) - f^*$, and $\boldsymbol{x}_{final} = \boldsymbol{x}_t$ refers to a version $\boldsymbol{x}_t$ from the set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{T-1}\}$, picked randomly.

Theorem 6 (Convex convergence rate of non-uniform D-EF-SGD with the absolute compressor, *i.e.*, $\mu = 0$)**.** *Consider a function f, which is mapping from $\mathbb{R}^d$ to $\mathbb{R}$, L-consistent and μ-convex. We can find a learning rate γ such that $\gamma \leq \frac{1}{4L}$. This means that the number of*

$$\mathcal{O}\left(\frac{\sigma^2 \sum_{i=1}^n p_i^2}{\epsilon^2} + \frac{\sqrt{nLd \sum_{i=1}^n p_i^2 \lambda_i^2}}{\epsilon^{\frac{3}{2}}} + \frac{L}{\epsilon}\right) \cdot R_0^2 \tag{14}$$

iterations of non-uniform D-EF-SGD with the absolute compressor ensures $\mathbb{E}f(\boldsymbol{x}_{final}) - f^* \leq \epsilon$, where R_0^2 is at least $\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|$, and $\boldsymbol{x}_{final} = \boldsymbol{x}_t$ refers to a version $\boldsymbol{x}_t$ from the set $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{T-1}\}$, picked randomly.

Theorem 7 (Strong convex convergence rate of non-uniform D-EF-SGD with the absolute compressor, *i.e.*, $\mu > 0$)**.** *Consider a function f, which is mapping from $\mathbb{R}^d$ to $\mathbb{R}$, L-consistent and μ-convex. We can find a learning rate γ such that $\gamma \leq \frac{1}{4L}$. This means that the number of*

$$\tilde{\mathcal{O}}\left(\frac{\sigma^2 \sum_{i=1}^n p_i^2}{\mu\epsilon} + \frac{\sqrt{nLd \sum_{i=1}^n p_i^2 \lambda_i^2}}{\mu\sqrt{\epsilon}} + \frac{L}{\mu}\right) \tag{15}$$

iterations of non-uniform D-EF-SGD with the absolute compressor ensures $\mathbb{E}f(\boldsymbol{x}_{final}) - f^* \leq \epsilon$, and $\boldsymbol{x}_{final} = \boldsymbol{x}_t$ refers to a version $\boldsymbol{x}_t$ from the sequence $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{T-1}\}$, selected probabilistically based on $(1 - \frac{\mu\gamma}{2})^{-t}$.

We predominantly direct our attention to **Theorem 7**. Analogous to the analysis of non-uniform D-EF-SGD with the absolute compressor, our focus is constrained to scenarios wherein $\sigma = 0$ or during initial training phases. Under these circumstances, the non-uniform D-EF-SGD with the absolute compressor converges at a rate of $\mathcal{O}(\sqrt{\sum_{i=1}^n p_i^2 \lambda_i^2})$.

B. The optimal thresholds in communication-constrained non-IID scenarios

We demonstrate the theorem on the optimal λ_i, *i.e.*, **Theorem 8** and DAGC-A, We use the Lagrange multiplier method to prove this theorem and more details can be seen in Appendix.

Theorem 8 (Conversion from λ to δ and optimal λ_i)**.** *In the non-IID scenario where the total communication traffic is predetermined and constrained, we have*

Algorithm 3: DAGC-A

Input: n, p_i and the average threshold $\bar{\lambda}$
Output: thresholds $\lambda_1, \ldots, \lambda_n$

$P = \sum_{i=1}^{n} p_i^{\frac{2}{3}}$;
for $i = 1, \ldots, n$ **do**
  /* calculating λ_i according to **Theorem 8** */
  $\lambda_i = \frac{\bar{\lambda} P}{n} p_i^{-\frac{2}{3}}$
end
Return $\lambda_1, \ldots, \lambda_n$;

$$\lambda \propto \frac{1}{\delta} \tag{16}$$

and the optimal λ_i satisfying

$$\lambda_i = \frac{\bar{\lambda} P}{n} p_i^{-\frac{2}{3}}, \forall i \in [n], \tag{17}$$

where $P = \sum_{i=1}^{n} p_i^{\frac{2}{3}}$ and $\bar{\lambda} = \frac{n}{\sum_{i=1}^{n} \frac{1}{\lambda_i}}$.

Remark 3. In communication-constrained environments with non-IID datasets, the relationship between λ and δ is different from the previous work [17], which only presents a conversion formula applied to IID scenarios[6].

Remark 4. In IID scenarios, it follows that $\lambda_i = \bar{\lambda}$ for all i due to $p_i = p_j$. This implies that for the absolute compressor, the optimal compression strategy in IID scenarios is also the uniform compression.

C. DAGC-A and its implementation in FL

We demonstrate the formula for the optimal λ_i in **Theorem 8**, based on which we show the pseudo-code of DAGC-A in Algorithm 3. The difference between DAGC-A and DAGC-R is that the time complexity of computation in DAGC-A is only $\mathcal{O}(1)$, and the rest makes no difference in the implementation. The reason for this simplicity is that we make a one-step approximation to the computation of λ_i to δ_i, *i.e.,* $\frac{p_i}{p_j} = (\frac{\lambda_i}{\lambda_j})^{-\frac{3}{2}}$. This simplification takes into account the constrained arithmetic in edge computing. We sacrifice some performance to reduce the time complexity of DAGC from $\mathcal{O}(n)$ to $\mathcal{O}(1)$.

VI. Evaluation Experiments

The following questions are addressed by this evaluation:
• If the communication budget is limited and fixed, does DAGC surpass uniform compression in real-world datasets? (Fig. 3 in Sec. VI-B)
• As the size distribution becomes more imbalanced and the compression becomes more aggressive, will DAGC exhibit better performance? (Table III in Sec. VI-C and Table V in Sec. VI-E)

[6]The formula, demonstrated in end of the appendix of the work [17], is $\lambda \propto \frac{1}{\sqrt{\delta}}$. They derived this equation by taking $\zeta = 0$, equal to IID environments.

We showcase the superior performance of DAGC in both real-world non-IID and artificially partitioned non-IID datasets, particularly when confronted with highly imbalanced size distribution and constrained communication.

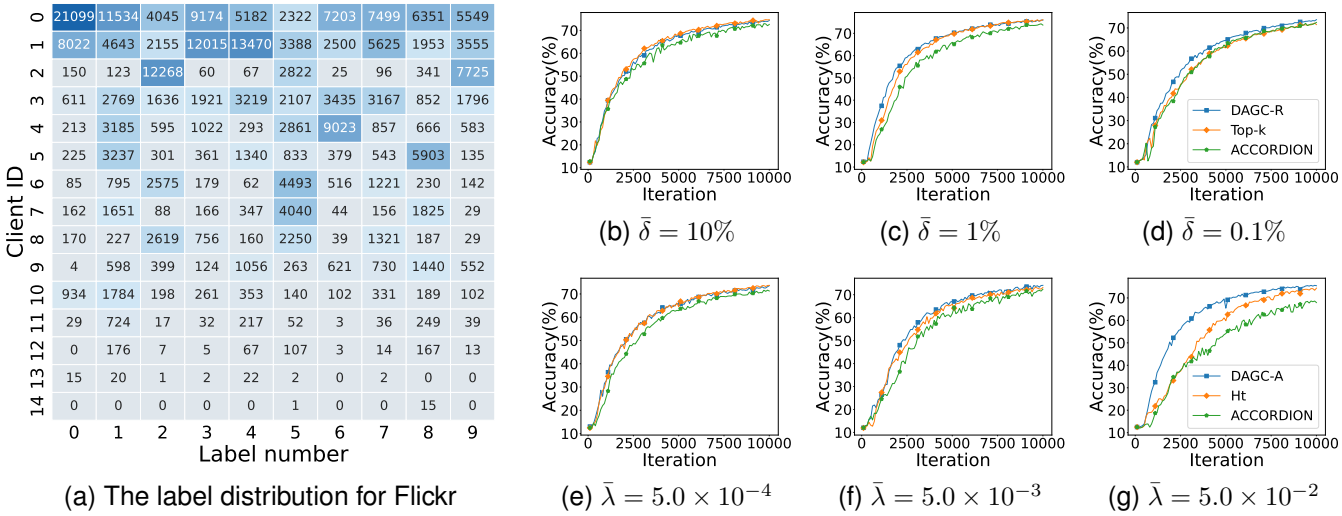
A. Experimental settings

Environment: The experiments are conducted on an Ubuntu 18.04.6 LTS server environment. The server is equipped with an Intel Xeon Silver 4210 CPU @2.20GHz and 4 Nvidia GeForce GTX 3090 GPUs, each with 24GB memory. Python version 3.8.12 is utilized, along with various libraries compatible with this Python version. For machine learning purposes, PyTorch 1.11.0 with CUDA 11.3 is employed as the primary toolkit.
Non-IID type: We run the experiments in two non-IID types:
• *Artificial non-IID data partition:* In order to simulate label imbalance, we assign a portion of samples from each label to individual workers based on the Dirichlet distribution. The concentration parameter is set to 0.5. This partitioning strategy is widely adopted for generating non-IID data [9], [37], [38].
• *Real-world datasets:* We utilize the Flickr [3] dataset as our real-world dataset. The dataset is downloaded from `https://doi.org/10.5281/zenodo.3676081` and the images are divided based on the subcontinent they belong to. After excluding damaged images and accounting for network limitations that prevented the download of certain images, we have a total of 15 workers. The data distribution is illustrated in Fig. 3a.
Experiment tasks: The experimental settings in this study encompass four different types, involving tasks related to image classification and speech recognition in Tab. II. The Convolutional Neural Network (CNN) employed consists of four layers, as mentioned earlier [6]. VGG11 [39] consists of 11 layers (including 8 convolutional layers and 3 fully-connected layers), using a continuous 3x3 small convolutional kernel. VGG11s is a simplified version of the VGG11 [40]. ResNet18 [41] is a deep convolutional neural network consisting of 18 convolutional layers and residual blocks, each of which mitigates the gradient vanishing problem by constant mapping. The LSTM model utilized has 2 hidden layers with a size of 128. In order to mitigate the accuracy loss caused by non-IID, all of these models exclude the batch normalization layer [3].

We denote A@B as the task, which uses the B datasets to train A model. For tasks CNN@CIFAR-10, VGG11s@Flickr, VGG11@CIFAR-100, and ResNet18@CIFAR-10, the learning rates are 0.01, 0.1, 0.01, and 0.01, respectively, and the batch sizes are 32 for all tasks. For the Speech Recognition task, we utilize the Speech Commands dataset [32] (referred to as SCs). From SCs, we select the 10 categories with the highest number of samples. Specifically, we extract $4,000$ samples from each of these categories, with $3,000$ samples allocated for training purposes and the remaining $1,000$ samples reserved for testing. The batch size is 8 and the step size is 0.1 in LSTM@SCs.
Baselines: In our comparative analysis, DAGC is evaluated against established uniform compression strategies. For

TABLE II: Summary of the experiment settings used in this work.

Task	Model	Dataset	non-IID type	Quality metric	Training iterations	Experiment Section
Image Classification	Logistic	FMNIST [31]	Artificially	Top-1 Accuracy	$5,000$	Sec. III, VI
	CNN	CIFAR-10 [35]	Artificially		$10,000$	Sec. VI
	VGG11s	Flickr [3]	Real-world		$10,000$	Sec. VI
	VGG11	CIFAR-100 [36]	Artificially		$50,000$	Sec. VI
	ResNet18	CIFAR-10	Artificially		$10,000$	Sec. VI
Speech Recognition	LSTM	SCs [32]	Artificially	Top-1 Accuracy	$10,000$	Sec. III, VI



(a) The label distribution for Flickr (b) $\bar{\delta} = 10\%$ (c) $\bar{\delta} = 1\%$ (d) $\bar{\delta} = 0.1\%$ (e) $\bar{\lambda} = 5.0 \times 10^{-4}$ (f) $\bar{\lambda} = 5.0 \times 10^{-3}$ (g) $\bar{\lambda} = 5.0 \times 10^{-2}$

Fig. 3: The label distribution for Flickr (a) and training curves (Accuracy vs. Iterations) for VGG11s@Flickr under the relative compression ((b)-(d)) and the absolute compression ((e)-(g)) on different compression levels (left to right). DAGC outperforms other uniform compression strategies facing limited communication under the fixed budget.

DAGC-R, Top-k and ACCORDION [42] serve as the baselines. In the case of DAGC-A, the hard-threshold (denoted as Ht) and ACCORDION are the selected baselines. The compression ratio of Top-k is set to $\bar{\delta}$ and the threshold of Ht is $\bar{\lambda}$. ACCORDION is the state-of-the-art sparsified adaptive gradient compression algorithm, which compresses the gradient using aggressive compression in the critical regime and conservative compression if not. Specifically, within the relative compressor, the aggressive (conservative) compression ratio of ACCORDION is set to $\delta_{\min}$ ($\delta_{\max}$). Conversely, under the absolute compression, the aggressive compression threshold is set to $\lambda_{\max}$.

The number of workers and the worker size: This experimental setting is specifically for generating artificial non-IID data partitions. We set the number of workers equal to 10. The worker size does not undergo a dichotomous division (as used in Sec. III for simplicity). Instead, p_i is an arithmetic series. To increase the randomness of the series and to better match real-world datasets, we incorporate a Dirichlet distribution (with a concentration parameter of 0.5) into p_i, $i \in [2, n-1]$. This results in an approximate arithmetic series that remains descending order. This approach allows us to generate artificially non-IID datasets with different worker sizes in Federated Learning [9], [43]. To measure the imbalance of datasets, we define the skew ratio (abbreviated as SR) as p_1/p_n.

B. DAGC-R in real-world non-IID scenarios

The experimental results indicate that the performance of DAGC-R surpasses Top-k with uniform compression and ACCORDION in real-world scenarios with fixed communication volumes.

Fig. 3a displays the data distribution of Flickr [3] (sliced by subcontinent), which has a skew ratio of $4,997$ ($\approx \frac{79958}{16}$). We find out the 10 categories with the highest number of labels in all the images and select these 10 categories of images from 15 workers as the training dataset.

From Fig. 3b, it can be seen that at an average compression rate $\bar{\delta} = 10\%$, DAGC-R converges almost as fast as Top-k and both are slightly faster than ACCORDION.

Fig. 3c shows that DAGC-R converges faster than Top-k in the early stages, and then gradually equalizes with Top-k later on. DAGC-R has always been faster than ACCORDION.

The training accuracy curve in Fig. 3d shows that at $\bar{\delta} = 0.1\%$, DAGC-R consistently has a superior performance compared to Top-k and ACCORDION. DAGC-R achieves the same accuracy (70%) with 16.65% and 13.46% fewer iterations (from $8,410$ iterations and $8,100$ iterations to $7,010$ iterations) relative to Top-k and ACCORDION, respectively.

Overall, DAGC-R outperforms Top-k with uniform compression and ACCORDION in the normal compression interval with fixed communication volumes.

TABLE III: Accuracy of different relative gradient compression algorithms under different SR and average relative compression ratios $\bar{\delta}$. Increasing SR indicates a greater imbalance among the size distribution of datasets. $\bar{\delta}$ quantifies the extent of communication limitations. Numbers demonstrate that DAGC-R surpasses the performance of the uniform compression on all tasks. The superiority is particularly notable in environments where worker size distribution is highly uneven and communication bandwidth is constrained.

Model @Dataset	SR	$\bar{\delta}$	**DAGC-R**	Top-k	ACCORDION
CNN @CIFAR-10	10	10%	70.02%	**70.14%**	69.77%
		1%	69.66%	**69.68%**	69.60%
		0.1%	**68.85%**	68.72%	68.77%
	100	10%	68.68%	**68.79%**	68.59%
		1%	**69.25%**	68.87%	68.46%
		0.1%	**69.29%**	67.36%	67.93%
	1,000	10%	68.00%	**68.12%**	67.89%
		1%	**68.06%**	67.28%	66.49%
		0.1%	**68.25%**	67.35%	67.75%
LSTM @SCs	10	10%	78.40%	**78.53%**	77.93%
		1%	**77.10%**	76.37%	76.87%
		0.1%	**77.07%**	76.23%	76.93%
	100	10%	75.17%	**75.50%**	74.70%
		1%	**74.60%**	73.53%	73.87%
		0.1%	**73.33%**	71.53%	70.70%
	1,000	10%	**73.03%**	72.87%	71.43%
		1%	**72.67%**	71.17%	72.03%
		0.1%	**71.80%**	70.27%	71.13%
Logistic @FMNIST	10	10%	**83.44%**	83.38%	83.41%
		1%	**83.36%**	83.21%	83.32%
		0.1%	**83.18%**	83.09%	83.15%
	100	10%	83.35%	83.38%	**83.41%**
		1%	**83.26%**	83.23%	83.18%
		0.1%	**83.13%**	82.92%	83.10%
	1,000	10%	**83.17%**	**83.17%**	82.86%
		1%	**83.24%**	83.05%	83.16%
		0.1%	**83.06%**	82.95%	82.90%

TABLE IV: Accuracy of different relative compressors under different numbers of workers n in CNN@CIFAR-10 under SR=100 and $\bar{\delta} = 0.1\%$.

n	**DAGC-R**	Top-k	ACCORDION
5	**67.68%**	67.02%	67.03%
10	**69.29%**	68.87%	68.46%
20	**68.94%**	68.55%	68.70%
50	**69.33%**	68.53%	68.60%
100	**69.77%**	69.26%	69.19%
200	**69.82%**	69.24%	69.79%

C. DAGC-R in artificially partitioned non-IID scenarios

The detailed experimental results presented in illustrate the outperformance of DAGC-R compared to Top-k and ACCORDION in the case of highly skewed datasets.

Comparison of different skew ratios: The superiority of DAGC-R in both CNN@CIFAR-10 and LSTM@SCs tasks becomes more obvious as the skew ratio increases. DAGC-R, Top-k and ACCORDION perform similarly when SR $= 10$. However, when SR increases to $1,000$, the accuracy of DAGC-R is higher than Top-k and ACCORDION.

Comparison of different $\bar{\delta}$: DAGC-R is always the best in both tasks regardless of the skew ratio, when $\bar{\delta} = 0.1\%$. This suggests that DAGC-R is suitable for situations where communications are extremely limited.

Comparison of different numbers of workers: In Table IV, we compare the performance of several compression algorithms under different n. Even under $n = 200$, DAGC-R still shows better performance than other algorithms, which indicates the scalability of DAGC-R.

D. DAGC-A in real-world non-IID scenarios

The experimental results show that DAGC-A outperforms hard-threshold and ACCORDION in the real-world non-IID dataset. The advantage over hard-threshold and ACCORDION becomes more pronounced as compression becomes more aggressive.

Fig. 3e and Fig. 3f illustrate that in the conservative case of compression, DAGC-A and the hard-threshold perform similarly but both outperform ACCORDION.

Fig. 3g shows that under very aggressive compression, DAGC-A converges much faster than hard-threshold and ACCORDION, especially in the first half of the training stage exhibiting a very clear advantage. DAGC-A ultimately saves 25.43% of iterations (from $6,960$ iterations to $5,190$ iterations) over hard-threshold to reach 70% accuracy, while ACCORDION does not reach the same accuracy until the end of training.

E. DAGC-A in artificially partitioned non-IID scenarios

The experimental results show that DAGC-A performs better than Ht and ACCORDION in extremely communication-constrained situations and datasets with high skew ratios. The detailed experimental results are in Table V.

Comparison of different skew ratios: DAGC-A behaves more prominently when the skew ratio is relatively large. In the CNN@CIFAR-10 task, fixing $\bar{\lambda} = 0.05$, the training accuracy of DAGC-A is 1.35%, 6.54%, and 7.72% higher than that of Ht for the skew ratio of $10, 100, 1,000$, respectively. In the LSTM@SCs task, DAGC-A always performs best when SR $= 1,000$, while it only performs best when SR $= 10$ with average $\bar{\lambda} = 0.0005$.

Comparison of different $\bar{\lambda}$: When the average lambda is larger, which means that in more aggressive compression, DAGC-A always achieves the highest accuracy after the same number of iterations. Keeping the skew ratio constant, the advantage of DAGC-A in both CNN@CIFAR-10 and LSTM@SCs tasks becomes more obvious as $\bar{\lambda}$ increases.

Comparison of different numbers of workers: As n increases, the accuracy improvement decreases. This is because the larger worker size increases the number of input samples per iteration, thus shortening the training iterations. However, DAGC-A is still superior to other compression strategies under varying n.

F. Comparsion of real time cost

In Fig. 4, we show the training curves that illustrate accuracy over time for DAGC-R compared to other compression strategies and the baseline without compression. The experiments were conducted in the dynamic network, with an average bandwidth of 8.4 Mb/s [44].

Fig. 4a shows that at $\bar{\delta} = 0.1\%$, DAGC-R significantly reduces the time needed to reach 60% (as well as 70%) accuracy.

TABLE V: Accuracy of different absolute gradient compression algorithms under different SR and average absolute compression thresholds $\bar{\lambda}$. A larger $\bar{\lambda}$ represnets a more aggressive compression. DAGC-A also surpasses the performance of the uniform absolute compressors in communication-constrained non-IID scenarios.

Model @Dataset	SR	$\bar{\lambda}$	**DAGC-A**	Ht	ACCORDION
CNN @CIFAR-10	10	0.0005	69.79%	69.73%	**69.89%**
		0.005	**68.52%**	68.06%	68.39%
		0.05	**63.69%**	62.34%	61.80%
	100	0.0005	69.12%	**69.31%**	68.37%
		0.005	**68.25%**	67.59%	67.65%
		0.05	**63.47%**	56.93%	60.33%
	1,000	0.0005	68.04%	**68.12%**	67.28%
		0.005	**67.27%**	66.58%	66.87%
		0.05	**64.17%**	56.45%	63.54%
LSTM @SCs	10	0.0005	77.63%	**78.03%**	76.60%
		0.005	76.60%	**76.90%**	76.73%
		0.05	**75.27%**	73.97%	73.43%
	100	0.0005	75.10%	74.83%	**75.40%**
		0.005	**74.13%**	73.93%	72.60%
		0.05	**73.37%**	70.90%	72.13%
	1,000	0.0005	**72.13%**	72.10%	71.10%
		0.005	**71.97%**	70.57%	70.70%
		0.05	**70.17%**	67.53%	68.70%
Logistic @FMNIST	10	0.0005	83.31%	**83.35%**	83.15%
		0.005	83.26%	83.19%	**83.30%**
		0.05	**82.99%**	82.65%	82.97%
	100	0.0005	**83.29%**	83.16%	83.20%
		0.005	**83.24%**	83.20%	83.17%
		0.05	**83.04%**	82.50%	83.04%
	1,000	0.0005	**83.19%**	**83.19%**	82.93%
		0.005	**83.18%**	83.12%	82.82%
		0.05	**83.14%**	82.10%	83.05%

TABLE VI: Accuracy of different absolute compression algorithms with varying numbers of workers n in CNN@CIFAR-10 under SR=100 and $\bar{\lambda} = 0.05$.

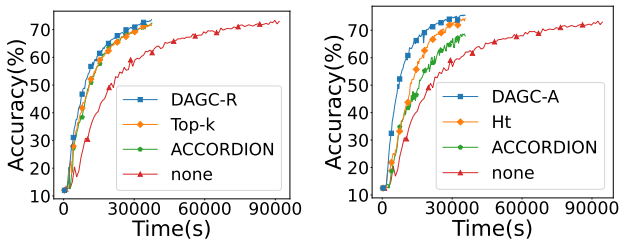
n	**DAGC-A**	Ht	ACCORDION
5	**61.68%**	56.93%	59.12%
10	**63.47%**	56.93%	60.33%
20	**64.21%**	60.30%	64.09%
50	**65.96%**	63.52%	65.10%
100	**66.76%**	63.34%	66.39%
200	**68.15%**	65.14%	66.41%

Specifically, it saves 16.78%, 17.56% and 57.16% (12.75%, 11.12% and 56.14%) compared to Top-k, ACCORDION, and training without compression, respectively. Similarly, Fig. 4b reveals that, at $\bar{\lambda} = 0.05$, DAGC-A saves 36.19%, 50.64% and 68.28% (as well as 29.08%, 49.48% and 71.47%) of the time required to achieve 60% (70%) accuracy compared to Ht, ACCORDION and training without compression.

Overall, DAGC-R demonstrates more efficient convergence, achieving high accuracy within a shorter time frame compared to both the baseline and other compression strategies.

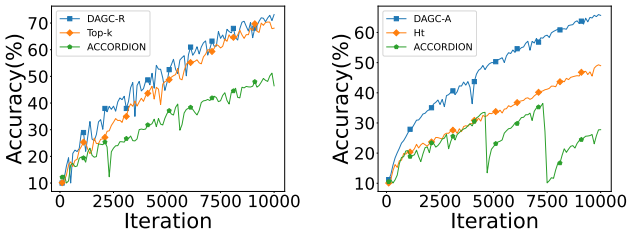
G. Scalability for models and datasets

We expand the experiments to ResNet18 and VGG11 models, as well as the CIFAR-100 dataset, with the results shown in Fig. 5. The experiments uses SR $= 100$, $\bar{\delta} = 0.1\%$ for the relative compression and $\bar{\lambda} = 0.05$ for the absolute compression. In all sub-figures, DAGC outperforms the uniform compression, demonstrating that DAGC has excellent scalability for different models and datasets.

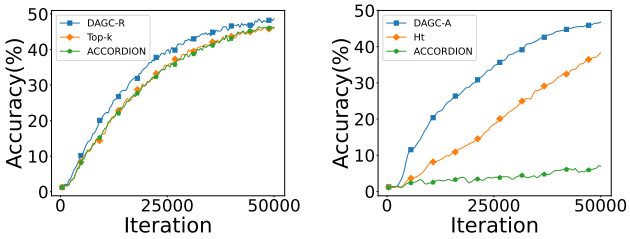


(a) DAGC-R (b) DAGC-A

Fig. 4: The training curves (Accuracy vs. Time) for VGG11s@Flickr under the relative compression (a) and the absolute compression (b). DAGC outperforms other compression strategies, and the training is without compression.



(a) DAGC-R ResNet18@CIFAR-10 (b) DAGC-A ResNet18@CIFAR-10



(c) DAGC-R VGG11@CIFAR-100 (d) DAGC-A VGG11@CIFAR-100

Fig. 5: The training curves (Accuracy vs. Iterations) for ResNet18 @CIFAR-10 and VGG11@CIFAR-100 under the relative compression (a, c) and the absolute compression (b, d). DAGC performs better in all cases.

In Fig. 5a (as well as Fig. 5b), DAGC-R (DAGC-A) can save up to 9.96% (52.74%) iterations compared to Top-k (Ht). Similarly, in Fig. 5c (as well as Fig. 5d), DAGC-R (DAGC-A) can save up to 16.67% (59.5%) iterations compared to Top-k (Ht). It should be noted that Ht and ACCORDION using Ht have severe accuracy degradation, showing its bad scalability for large datasets and models.

VII. Related Works

Communication optimization in communication-constraint non-IID scenarios aims to solve the communication bottleneck, while eliminating the accuracy degradation due to non-IID datasets. The work [45] proposes the Delayed Gradient Averaging algorithm, which enhances the DML efficiency by

delaying the gradient averaging step, allowing simultaneous local computation and communication. The work [46] proposes a completely parallelizable FL algorithm, P-FedAvg, which extends the traditional FedAvg by allowing multiple parameter servers to collaboratively train a model, ensuring efficient convergence and scalability in a Parallel Federated Learning architecture. A Resource-efficient FL system is proposed in the work [47] to tackle issues of resource heterogeneity in FL. There are also some works that propose to use a hierarchical architecture for DML in mobile environments, which matches the structure of LAN-WAN. The work [48] first proposes a hierarchical architecture in communication-constrained non-IID scenarios. The work [49] verifies the convergence of the hierarchical PS architecture in such scenarios. This work [50] proposes a training scheduling strategy for DML in the LAN-WAN architecture. However, these methods neither mitigate the accuracy degradation brought about by non-IID scenarios nor are they friendly to mobile communication environments that charge by data usage. Compared to these methods, gradient compression algorithms save more communication costs savings in a per-traffic billing mobile network environment [51], [52].

Adaptive gradient compression provides the ability to fine-tune compression parameters, a feature not often available in traditional algorithms [53], [54]. This enhances their robustness, particularly in varying scenarios that involve dynamic network environments and data heterogeneity. DC2 [11], a control setup based on network latency for handling compression, was proposed. This innovative system ensures timely completion of model training, even amidst fluctuating network conditions. This patent [55] introduces the system design of a statistical-based gradient compression method for a distributed training system that is based on the work [18]. The work [33] proposes a systematic examination of the trade-off between compression and model accuracy in Federated Learning, introducing an adaptation framework to optimize the compression rate in each iteration for improved model accuracy while reducing network traffic. The work proposes a transmission strategy, called FedLC [56], which combines model compression, forward error correction, and retransmission to improve the network utilization in FL with lossy communication. The study introduces SkewScout [3], aiming to enhance the robustness of algorithms particularly in non-IID scenarios. SkewScout achieves this by dynamically modulating the compression ratio according to the disparity in loss among workers, which is a parameter notoriously challenging to evaluate. As a result, the implementation of SkewScout becomes a complex task.

Theoretically optimizing gradient compression: Some works [28] attempt to improve the convergence rate, *i.e.*, tighter upper bounds and lower time complexity, to optimize existing compression algorithms for mobile scenarios with communication constraints and heterogeneous data. The work [57], [58] utilizes smoothness matrices to boost existing compressors in both theory and practice. The study takes a comparison between D-QSGD and D-EF-SGD in non-IID conditions, integrating bias correction [30] to enhance their data dependency. The paper also provides a theoretical analysis of the robustness of the hard threshold sparsification algorithm which transmits solely the absolute gradient values exceeding a fixed hard threshold [17]. The findings indicate that compared to Top-k, this algorithm is more resilient when dealing with non-IID challenges.

Data-aware methods: There have been studies that put forward the idea of data-aware node selection in DML. They introduce a technique [59] that uses data volumes as a criterion for worker selection within Federated Learning. This stands in contrast to the traditional gradient-based methods [60]. These experiments confirm that the data-volume-oriented node selection approach is superior to the uniform selection tactic in non-IID situations. This indicates that allowing *large workers* to transmit more data could potentially be advantageous. Yet there don't seem to be any existing studies introducing data-aware algorithms in the gradient compression domain. In our study, we introduce a data-aware gradient compression algorithm, accompanied by an in-depth theoretical analysis.

VIII. Conclusion

In the study, we introduce an innovative gradient compression algorithm, drawing from a fresh perspective by considering the unevenness in data volume sizes, which enhances its robustness with non-IID datasets. As an initial step, we present empirical evidence supporting the idea that assigning higher compression ratios to workers dealing with larger data volumes can accelerate convergence. Following this, we establish the convergence rate for non-uniform D-EF-SGD when applied in conjunction with either relative or absolute compressors. We derive the key factors, which greatly affect the model convergence in the communication-constrained non-IID environment. By minimizing this factor, we propose DAGC-R, which sets $\frac{\delta_i}{\delta_j} \approx (\frac{p_i}{p_j})^{2/3}$, and DAGC-A, where $(\frac{\lambda_i}{\lambda_j})^{2/3} = \frac{p_j}{p_i}$. We assess the effectiveness of DAGC by conducting tests on both real-world datasets and artificially partitioned non-IID datasets. The results of these evaluation experiments showcase that DAGC has the capability to reduce the number of iterations by as much as 25.43% on the real-world non-IID datasets. Furthermore, on artificially separated non-IID datasets, it enhances the accuracy by a substantial 3.14%.

References

[1] S. Bubeck, V. Chandrasekaran, R. Eldan, *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 1273–1282.

[3] K. Hsieh, A. Phanishayee, O. Mutlu, and P. Gibbons, "The non-iid data quagmire of decentralized machine learning," in *International Conference on Machine Learning*, PMLR, 2020, pp. 4387–4398.

[4] K. Hsieh, A. Harlap, N. Vijaykumar, *et al.*, "Gaia:{geo-distributed} machine learning approaching {lan} speeds," in *14th USENIX Symposium on Networked Systems Design and Implementation*, 2017, pp. 629–647.

[5] L. G. Valiant, "A bridging model for parallel computation," *Communications of the ACM*, vol. 33, no. 8, pp. 103–111, 1990.

[6] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[7] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.

[8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.

[9] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *IEEE International Conference on Data Engineering*, 2022.

[10] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2 - a large-scale benchmark for instance-level recognition and retrieval," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[11] A. M. Abdelmoniem and M. Canini, "Dc2: Delay-aware compression control for distributed machine learning," in *IEEE Conference on Computer Communications 2021*, 2021, pp. 1–10. DOI: 10.1109/INFOCOM42981.2021.9488810.

[12] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "Qsgd: Communication-efficient sgd via gradient quantization and encoding," *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., 2017.

[13] T. Vogels, S. P. Karimireddy, and M. Jaggi, "Powersgd: Practical low-rank gradient compression for distributed optimization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] Y. Lin, S. Han, H. Mao, Y. Wang, and B. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *ICLR*, 2018.

[15] L. Cui, X. Su, Y. Zhou, and Y. Pan, "Slashing communication traffic in federated learning by transmitting clustered model updates," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2572–2589, 2021.

[16] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified sgd with memory," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[17] A. Sahu, A. Dutta, A. M Abdelmoniem, T. Banerjee, M. Canini, and P. Kalnis, "Rethinking gradient sparsification as total error minimization," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[18] A. M Abdelmoniem, A. Elzanaty, M.-S. Alouini, and M. Canini, "An efficient statistical-based gradient compression technique for distributed training systems," *Proceedings of Machine Learning and Systems*, vol. 3, pp. 297–322, 2021.

[19] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," *arXiv preprint arXiv:1704.05021*, 2017.

[20] C. Tang, K. Ouyang, Z. Wang, *et al.*, "Mixed-precision neural network quantization via learned layer-wise importance," in *European Conference on Computer Vision*, Springer, 2022, pp. 259–275.

[21] C. Tang, L. L. Zhang, H. Jiang, *et al.*, "Elasticvit: Conflict-aware supernet training for deploying fast vision transformer on diverse mobile devices," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[22] H. Xu, C.-Y. Ho, A. M. Abdelmoniem, *et al.*, "Grace: A compressed communication framework for distributed machine learning," in *IEEE International Conference on Distributed Computing Systems*, 2021.

[23] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "Signsgd: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*, PMLR, 2018, pp. 560–569.

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

[25] D. Basu, D. Data, C. Karakus, and S. Diggavi, "Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[27] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Robust and communication-efficient federated learning from non-iid data," *IEEE Transactions on Neural Networks and Learning systems*, vol. 31, no. 9, pp. 3400–3413, 2019.

[28] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated Learning With Quantized Global Model Updates," *arXiv e-prints*, arXiv:2006.10672, arXiv:2006.10672, Jun. 2020. arXiv: 2006.10672 [cs.IT].

[29] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "SignSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Machine Learning Research, vol. 80, PMLR, 2018, pp. 560–569.

[30] S. U. Stich, "On communication compression for distributed optimization on heterogeneous data," *arXiv preprint arXiv:2009.02388*, 2020.

[31] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[32] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[33] L. Cui, X. Su, Y. Zhou, and J. Liu, "Optimal rate adaption in federated learning with compressed communications," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, IEEE, 2022, pp. 1459–1468.

[34] A. Koloskova, S. U. Stich, and M. Jaggi, "Sharper convergence guarantees for asynchronous sgd for distributed and federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 202–17 215, 2022.

[35] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[36] A. Krizhevsky *et al.*, "Learning multiple layers of features from tiny images," 2009.

[37] Q. Li, B. He, and D. Song, "Practical one-shot federated learning for cross-silo setting," *arXiv preprint arXiv:2010.01017*, 2020.

[38] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," *arXiv preprint arXiv:2002.06440*, 2020.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[40] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," in *International Joint Conference on Neural Networks*, IEEE, 2019, pp. 1–8.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] S. Agarwal, H. Wang, K. Lee, S. Venkataraman, and D. Papailiopoulos, "Adaptive gradient communication via critical learning regime identification," *Machine Learning and Systems*, vol. 3, pp. 55–80, 2021.

[43] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7611–7623, 2020.

[44] M. Zhang, F. Wang, Y. Zhu, J. Liu, and B. Li, "Serverless empowered video analytics for ubiquitous networked cameras," *IEEE Network*, vol. 35, no. 6, pp. 186–193, 2021.

[45] L. Zhu, H. Lin, Y. Lu, Y. Lin, and S. Han, "Delayed gradient averaging: Tolerate the communication latency for federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 995–30 007, 2021.

[46] X. Liu, Z. Zhong, Y. Zhou, *et al.*, "Accelerating federated learning via parallel servers: A theoretically guaranteed approach," *IEEE/ACM Transactions on Networking*, vol. 30, no. 5, pp. 2201–2215, 2022.

[47] A. M. Abdelmoniem, A. N. Sahu, M. Canini, and S. A. Fahmy, "Refl: Resource-efficient federated learning," in *Proceedings of the Eighteenth European Conference on Computer Systems*, 2023, pp. 215–232.

[48] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, IEEE, 2020, pp. 1–6.

[49] J. Wang, S. Wang, R.-R. Chen, and M. Ji, "Demystifying why local aggregation helps: Convergence analysis of hierarchical sgd," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 8548–8556.

[50] J. Yuan, M. Xu, X. Ma, A. Zhou, X. Liu, and S. Wang, "Hierarchical federated learning through lan-wan orchestration," *arXiv preprint arXiv:2010.11612*, 2020.

[51] A. Malekijoo, M. J. Fadaeieslam, H. Malekijou, M. Homayounfar, F. Alizadeh-Shabdiz, and R. Rawassizadeh, "Fedzip: A compression

framework for communication-efficient federated learning," *arXiv preprint arXiv:2102.01593*, 2021.
[52] F. Haddadpour, M. M. Kamani, A. Mokhtari, and M. Mahdavi, "Federated learning with compression: Unified analysis and sharp guarantees," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 2350–2358.
[53] J. Guo, W. Liu, W. Wang, *et al.*, "Accelerating distributed deep learning by adaptive gradient quantization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2020, pp. 1603–1607.
[54] J. Wang and G. Joshi, "Adaptive communication strategies to achieve the best error-runtime trade-off in local-update sgd," *Machine Learning and Systems*, vol. 1, pp. 212–229, 2019.
[55] A. M. Abdelmoniem, A. Elzanaty, M. Canini, and M.-S. Alouini, *Statistical-based gradient compression method for distributed training system*, 2022.
[56] X. Su, Y. Zhou, L. Cui, and J. Liu, "On model transmission strategies in federated learning with lossy communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 34, no. 4, pp. 1173–1185, 2023. DOI: 10.1109/TPDS.2023.3240883.
[57] M. Safaryan, F. Hanzely, and P. Richtárik, "Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 25 688–25 702, 2021.
[58] B. Wang, M. Safaryan, and P. Richtárik, "Theoretically better and numerically faster distributed optimization with smoothness-aware quantization techniques," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9841–9852, 2022.
[59] C. Dupuy, T. G. Roosta, L. Long, C. Chung, R. Gupta, and S. Avestimehr, "Learnings from federated learning in the real world," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022. [Online]. Available: https://www.amazon.science/publications/learnings-from-federated-learning-in-the-real-world.
[60] H. Wu and P. Wang, "Node selection toward faster convergence for federated learning on non-iid data," *IEEE Transactions on Network Science and Engineering*, 2022.

IX. Appendix

A. Proof of ***Theorem 1, 2, 3***

We define a virtual sequence that aids in our derivation, referring to [30]:

$$\tilde{\mathbf{x}}_0 = \mathbf{x}_0, \quad \tilde{\mathbf{x}}_{t+1} := \tilde{\mathbf{x}}_t - \gamma \sum_{i=1}^{n} p_i g_t^i.$$

The error term that illustrates the gap between the virtual sequence and the actual sequence is represented as

$$\tilde{\mathbf{x}}_t - \mathbf{x}_t = \gamma \sum_{i=1}^{n} p_i \mathbf{e}_t^i.$$

For further discussion, let's define $G_t := \mathbf{E}\|\nabla f(x_t)\|^2$, $E_t = \sum_{i=1}^n p_i^2 \mathbb{E}\|\mathbf{e}_t^i\|^2$, $\tilde{F}_t := \mathbf{E} f(\tilde{\mathbf{x}}_t) - f^*$ and $F_t := \mathbf{E} f(\mathbf{x}_t) - f^*$.

Lemma 1. *Considering a function f, which is L-smooth. If the learning rate γ is less than or equal to $\frac{1}{4L}$, the following is true for the iterations of non-uniform D-EF-SGD with relative compression:*

$$\tilde{F}_{t+1} \leq \tilde{F}_t - \frac{\gamma}{4} G_t + \gamma^2 \frac{L \sum_{i=1}^n p_i^2 \sigma^2}{2} + \gamma^3 \frac{nL^2}{2} E_t. \quad (18)$$

If f is exhibits μ-convexity, the following is observed

$$X_{t+1} \leq (1 - \frac{\gamma\mu}{2}) X_t - \frac{\gamma}{2} F_t + \gamma^2 \sum_{i=1}^{n} p_i^2 \sigma^2 + 3\gamma^3 nL E_t. \quad (19)$$

Proof. Similar to the analysis in [30], we conclude

$$\begin{aligned} \tilde{F}_{t+1} &\leq \tilde{F}_t - \frac{\gamma}{4} G_t + \gamma^2 \frac{L}{2} \mathbb{E}\|\sum_{i=1}^{n} p_i \xi_t^i\|^2 \\ &+ \gamma^3 \frac{L^2}{2} \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2, \\ X_{t+1} &\leq (1 - \frac{\gamma\mu}{2}) X_t - \frac{\gamma}{2} F_t + \gamma^2 \mathbb{E}\|\sum_{i=1}^{n} p_i \xi_t^i\|^2 \\ &+ 3\gamma^3 L \mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2. \end{aligned}$$

With the independent ξ_t^i and **Assumption 3** in [30], the following equation emerges:

$$\mathbb{E}_{\xi_t}\|\sum_{i=1}^{n} p_i \xi_t^i\|^2 = \sum_{i=1}^{n} p_i^2 \mathbb{E}_{\xi_t}\|\xi_t^i\|^2 \leq \sum_{i=1}^{n} p_i^2 \sigma^2.$$

Additionally, we derive:

$$\mathbb{E}\|\tilde{\mathbf{x}}_t - \mathbf{x}_t\|^2 = \mathbb{E}\|\sum_{i=1}^{n} p_i e_t^i\|^2 \leq {}^7 n \sum_{i=1}^{n} p_i^2 \mathbb{E}\|e_t^i\|^2 = nE_t.$$

Consequently, our targeted outcomes are achieved.

Lemma 2. *It holds*

$$\begin{aligned} E_{t+1} &\leq (1 - \frac{\delta_{min}}{2}) E_t + \sum_{i=1}^{n} p_i^2 \sigma^2 \\ &+ \frac{2}{\delta_{min}} (C_\zeta \zeta^2 + C_Z Z^2 G_t), \end{aligned} \quad (20)$$

where $C_\zeta = C_Z = \sum_{i=1}^n \frac{\delta_{min}}{\delta_i} p_i^2$.

Proof. With the analysis in [30], it follows

$$\begin{aligned} &\mathbb{E}_{\xi_t^i, C_\delta}\|\mathbf{e}_{t+1}^i\|^2 \\ &\leq (1 - \frac{\delta}{2})\|\mathbf{e}_t^i\|^2 + \frac{2}{\delta}\|\nabla f_i(\mathbf{x}_t)\|^2 + (1-\delta)\sigma^2 \\ &\leq (1 - \frac{\delta}{2})\|\mathbf{e}_t^i\|^2 + \frac{2}{\delta}(\zeta_i^2 + Z^2\|\nabla f(\mathbf{x}_t)\|^2) + \sigma^2. \end{aligned} \quad (21)$$

The final inequality is based on **Assumption 4** as cited in [30]. Then we incorporate the distinct compression ratios from various workers into Eq. 21 and aggregate the outcomes:

$$\begin{aligned} E_{t+1} &\leq (1 - \frac{\delta_{min}}{2}) \sum_{i=1}^{n} p_i^2 \|\mathbf{e}_t^i\|^2 \\ &+ \frac{2}{\delta_{min}} (\sum_{i=1}^{n} \frac{\delta_{min}}{\delta_i} p_i^2)(\zeta^2 + Z^2 G_t) + \sum_{i=1}^{n} p_i^2 \sigma^2, \end{aligned}$$

where $\delta_{min} = \min\{\delta_1, \ldots, \delta_n\}, \zeta = \max\{\zeta_1, \ldots, \zeta_n\}$.

Lemma 3 (Lyapunov function)**.** *Considering a function f, which is L-smooth. If the learning rate γ is less than or equal to $\frac{\delta_{min}}{4LZ\sqrt{nC_Z}}$. Then it holds*

$$\begin{aligned} \Xi_{t+1} &\leq \Xi_t - \frac{\gamma}{8} G_t + \gamma^2 \frac{L \sum_{i=1}^n p_i^2 \delta^2}{2} + \\ &\gamma^3 \left(\frac{L^2 n}{\delta_{min}}\right) \left(\frac{2C_\zeta \zeta^2}{\delta_{min}} + \sum_{i=1}^{n} p_i^2 \sigma^2\right), \end{aligned} \quad (22)$$

[7]The inequality follows from the fact that $\|\sum_{i=1}^k a_i\|^2 \leq k \sum_{i=1}^k \|a_i\|^2$.

where $\Xi_t := \tilde{F}_t + bE_t, b = \dfrac{\gamma^3 L^2 n}{\delta_{min}}$. *Additionally, assuming* f *is both* L*-smooth and* μ*-convex and* γ *is less or equal to* $\dfrac{\delta_{min}}{14LZ\sqrt{nC_Z}}$, *the following holds:*

$$\begin{aligned}\Psi_{t+1} \quad \leq \quad & \left(1 - \min\left\{\frac{\gamma\mu}{2}, \frac{\delta}{4}\right\}\right)\Psi_t - \frac{1}{8L}G_t + \gamma^2\sum_{i=1}^{n} p_i^2\sigma^2 \\ & +\gamma^3\left(\frac{12Ln}{\delta_{min}}\right)\left(\frac{2C_\zeta\zeta^2}{\delta_{min}} + \sum_{i=1}^{n} p_i^2\sigma^2\right), \end{aligned} \tag{23}$$

where $\Psi_t := X_t + aE_t$ *with* $a = \dfrac{12\gamma^3 nL}{\delta_{min}}$.

Proof. For smooth functions, we incorporate Eq. 18 and 20 into the right side of the expression $\Xi_{t+1} := \tilde{F}_{t+1} + bE_{t+1}$.

In the case of convex functions, we introduce Eq. 19 and 20 into the right side of the expression $\Psi_{t+1} := X_{t+1} + aE_{t+1}$. This concludes the entire proof.

For the function that is non-convex, by integrating Eq. 22 with Appendix F's Lemma 27 from [30], we successfully validate **Theorem 1**. When addressing a convex function where $\mu = 0$, by employing Eq. 23 in conjunction with Lemma 27 from [30] found in Appendix F, we confirm **Theorem 2**. For the function showcasing strong convexity characterized by $\mu > 0$, using Eq. 23 and referring to Lemma 25 in Appendix F of [30], we establish the truth of **Theorem 3**.

B. Proof of ***Theorem 4***

It's essential to recognize that the primary challenge can be converted to discerning the local optimal solution for the function $\Phi(\delta_1, \ldots, \delta_n) = \frac{\frac{p_1}{\sqrt{\delta_1}} + \ldots + \frac{p_n}{\sqrt{\delta_n}}}{\sqrt{\delta_{min}}}$ taking into account the constraint $\sum_{i=1}^{n}\delta_i = n\bar{\delta}$ and the condition $\delta_i > 0$ for all values of i ranging from 1 to n. The demonstration of **Theorem 4** is broken down into two phases. In the initial phase, the problem with n variables and a single constraint is recast into an optimization problem with only one variable, as depicted by Eq. 4. In the subsequent phase, the minimum for this single-variable optimization issue is determined.

Lemma 4. *Suppose that* $a_i, b_i > 0, \forall i \in \{1, \ldots, n\}$ *with* $\sum_{i=1}^{n} a_i = A$ *(*A *is a constant),* b_i *are constants, we have*

$$\sum_{i=1}^{n}\frac{b_i}{\sqrt{a_i}} \geq A^{-\frac{1}{2}}(\sum_{i=1}^{n} b_i^{\frac{2}{3}})^{\frac{3}{2}}. \tag{24}$$

The inequality takes equal if $a_i = Ab_i^{\frac{2}{3}}(\sum_{i=1}^{n} b_i^{\frac{2}{3}})^{-1}$.

Proof. With the equality constrain on a_i, we define a Lagrangian function as follows:

$$\mathbb{L} = \sum_{i=1}^{n}\frac{b_i}{\sqrt{a_i}} + \sigma(\sum_{i=1}^{n} a_i - A).$$

Based on the condition for optimality, we can deduce:

$$\begin{cases} \frac{\partial \mathbb{L}}{\partial \sigma} & = \sum_{i=1}^{n} a_i - A = 0 \\ \frac{\partial \mathbb{L}}{\partial a_i} & = -\frac{1}{2}b_i a_i^{-\frac{3}{2}} + \sigma = 0, \forall i \in \{1, \ldots, n\} \end{cases}.$$

From the system of equations provided, we can deduce the sought-after result.

With **Lemma 4**, the function $\Phi(\delta_1, \ldots, \delta_n)$ is transformed into a one-dimensional function. Supposing $\delta_{min} = \delta_j \leq \min\{\delta_i\}, i \in \{1, \ldots, n\} \setminus \{j\}$, we define $b_i = p_i$ if $i \in [1, j-1]$, and for others $b_i = p_{i+1}$. We also set $a_i = \delta_i$ and $A = (n\bar{\delta} - \delta_j)$.

$$\Phi(\delta_1, \ldots, \delta_n) \geq \frac{p_j}{\delta_j} + \frac{(P - p_j^{\frac{2}{3}})^{\frac{3}{2}}}{\sqrt{(n\bar{\delta} - \delta_j)\delta_j}}, \text{ where } P = \sum_{i=1}^{n} p_i^{\frac{2}{3}}. \tag{25}$$

Eq. 25 is achieved when $\delta_i = (n\bar{\delta} - \delta_j)p_i^{\frac{2}{3}}(P - p_j^{\frac{2}{3}})^{-1}, i \neq j$. Given that p_i is sorted in descending sequence, $\min\{\delta_i\}$ is δ_n if $j \in \{1, \ldots, n-1\}$, or else $\min\{\delta_i\}$ is δ_{n-1}.

Taking into account that the minimum of the right side of Eq. 25 depends on the range of δ_j, we'll evaluate the scenarios for $j \in [1, n-1]$ and $j = n$ separately.

• If $j \in [1, n-1]$, we have

$$\delta_{min} = \delta_j = \frac{(n\bar{\delta} - \delta_j)p_n^{\frac{2}{3}}}{P - p_j^{\frac{2}{3}}}.$$

By setting $Q_j = \frac{P - p_j^{2/3}}{p_n^{2/3}}, j \in [1, n-1]$ and using $\delta_j \leq \min\{\delta_i\}$. We deduce the range for $\delta_j \in (0, \frac{n\bar{\delta}}{Q_j + 1}]$. By defining $H(\delta_j) = \frac{p_j}{\delta_j} + \frac{(P - p_j^{\frac{2}{3}})^{\frac{3}{2}}}{\sqrt{(n\bar{\delta} - \delta_j)\delta_j}}$, we can compute the derivative of $H(\delta_j)$:

$$\begin{aligned} H'(\delta_j) = & -p_j\delta_j^{-2} \\ & -\frac{1}{2}(P - p_j^{\frac{2}{3}})^{\frac{3}{2}}[(n\bar{\delta} - \delta_j)\delta_j]^{-\frac{3}{2}}(n\bar{\delta} - 2\delta_j) < 0. \end{aligned}$$

Thus we get the minimum of $H(\delta_j)$ at $\delta_j = \frac{n\bar{\delta}}{Q_j + 1}$:

$$H(\delta_j) \geq H(\frac{n\bar{\delta}}{Q_j + 1}) = \frac{1}{n\bar{\delta}}(p_n Q_j(1 + Q_j) + p_j(1 + Q_j)). \tag{26}$$

We combine Eq. 25 and 26 and complete the first case ($j \neq n$) in the proof.

• If $j = n$, we set $Q_n = \frac{P - p_n^{2/3}}{p_{n-1}^{2/3}}$ and have

$$\min\{\delta_i\} = \delta_{n-1} = \frac{n\bar{\delta} - \delta_n}{Q_n}.$$

We get the range of δ_n is $(0, \dfrac{n\bar{\delta}}{Q_n + 1}]$. In this range, $H'(\delta_j) < 0$ (the proof process is the same as $j \neq n$). We have

$$\begin{aligned} H(\delta_j) & \geq H(\frac{n\bar{\delta}}{Q_n + 1}) \\ & = \frac{1}{n\bar{\delta}}(p_n(1 + Q_n) + p_{n-1}Q_n(1 + Q_n)). \end{aligned} \tag{27}$$

We combine Eq. 25 and 27 and complete the second case ($j = n$) in the proof.

C. Proof of **Theorem 5, 6, 7**

Lemma 5. *Let f be L-smooth. If the stepsize $\gamma \leq \frac{1}{4L}$, then it holds for the iterates of non-uniform D-EF-SGD with the absolute compressor with the absolute compressor:*

$$\tilde{F}_{t+1} \leq \tilde{F}_t - \frac{\gamma}{4} G_t + \gamma^2 \frac{L \sum_{i=1}^n p_i^2 \sigma^2}{2} + \gamma^3 \frac{nL^2 d}{2} \sum_{i=1}^n p_i^2 \lambda_i^2. \tag{28}$$

If f is in addition μ-convex, we have

$$\begin{aligned} X_{t+1} \leq (1 - \frac{\gamma\mu}{2}) X_t - \frac{\gamma}{2} F_t \\ + \gamma^2 \sum_{i=1}^n p_i^2 \sigma^2 + 3\gamma^3 nLd \sum_{i=1}^n p_i^2 \lambda_i^2. \end{aligned} \tag{29}$$

Proof. According to the property of the absolute compressor, we have

$$E_t \leq d \sum_{i=1}^n p_i \lambda_i^2. \tag{30}$$

We combine the property and Lemma 1 and complete the proof.

For functions that is non-convexity, by incorporating Eq. 28 with Lemma 27 from Appendix F of [30], we can validate **Theorem 5**. When dealing with convex functions where $\mu = 0$, integrating Eq. 29 with Lemma 27 from Appendix F of [30] allows us to substantiate **Theorem 6**. For functions exhibiting pronounced convexity (where$\mu > 0$), by merging Eq. 29 with Lemma 25 from Appendix F of [30], we prove **Theorem 7**.

D. Proof of **Theorem 8**

Note that the condition is that the total communication traffic is constrained. Previous work [17] demonstrated a conversion formula from the threshold λ_i to the relative compression ratio δ_i in IID scenarios, *i,e,*. $\lambda_i = \frac{D}{\sqrt{\delta_i}}$. This formula can not apply to this work, for that we focus on communication-constrained non-IID scenarios.

To get the optimal λ_i, we assume that $\frac{\delta_i}{\delta_j} = (\frac{\lambda_i}{\lambda_j})^\Gamma$. Here, $\Gamma < 0$ since δ_i is negatively correlated to γ_i. Then we have $\sum_{i=1}^n \lambda_i^\Gamma = n\bar{\lambda}^\Gamma$ according to $\sum_{i=1}^n \delta_i = n\bar{\delta}$ in Sec. IX-B.

We use the Lagrange multiplier method and define a Lagrangian function as follows:

$$\mathbb{L} = \sum_{i=1}^n p_i^2 \lambda_i^2 + \sigma(\sum_{i=1}^n \lambda_i^\Gamma - n\bar{\lambda}^\Gamma).$$

By the optimality condition, we have

$$\begin{cases} \frac{\partial \mathbb{L}}{\partial \lambda_i} & = 2p_i^2 \lambda_i + \sigma\Gamma\lambda_i^{\Gamma-1} = 0, \forall i \in \{1, \ldots, n\} \\ \frac{\partial \mathbb{L}}{\partial \sigma} & = \sum_{i=1}^n \lambda_i^\Gamma - n\bar{\lambda}^\Gamma = 0 \end{cases}.$$

According to $\frac{\partial \mathbb{L}}{\partial \lambda_i} = 0$, we have $\frac{p_i}{p_j} = (\frac{\lambda_i}{\lambda_j})^{\frac{\Gamma}{2}-1}$. We combine this result with the property of DAGC-R, *i.e.*, $\frac{\delta_i}{\delta_j} \approx (\frac{p_i}{p_j})^{2/3}$, resulting in $\Gamma \approx -1$. Which means $\lambda_i \propto \frac{1}{\delta_i}$ and $\frac{p_i}{p_j} \approx (\frac{\lambda_i}{\lambda_j})^{-\frac{3}{2}}$.

For ease of calculation, we next use the equation, *i.e.*, $\frac{p_i}{p_j} = (\frac{\lambda_i}{\lambda_j})^{-\frac{3}{2}}$, rather than approximately equal. In this way, we have

$$\begin{cases} 2p_i^2 \lambda_i - \sigma\lambda_i^{-2} & = 0, \forall i \in \{1, \ldots, n\} \\ \sum_{i=1}^n \lambda_i^{-1} - n\bar{\lambda}^{-1} & = 0 \end{cases}.$$

By solving system of equations above, the proof completes.