International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 1373–1387
DOI: https://doi.org/10.2991/ijcis.d.210412.001; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Communication-Efficient Distributed SGD with Error-Feedback, Revisited

Tran Thi Phuong[1,2,3,*], Le Trieu Phong[3]

[1]*Faculty of Mathematics and Statistics, Ton Duc Thang University, No.19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam*
[2]*Meiji University, 1-1-1 Higashi-Mita, Tama-ku, Kawasaki-shi, Kanagawa, 214-8571, Japan*
[3]*National Institute of Information and Communications Technology (NICT) 4-2-1, Nukui-Kitamachi, Koganei, Tokyo, 184-8795, Japan*

ARTICLE INFO

Article History
Received 15 Jul 2020
Accepted 31 Mar 2021

Keywords
Optimizer
Distributed learning
SGD
Error-feedback
Deep neural networks

ABSTRACT

We show that the convergence proof of a recent algorithm called dist-EF-SGD for distributed stochastic gradient descent with communication efficiency using error-feedback of Zheng*et al.*, Communication-efficient distributed blockwise momentum SGD with error-feedback, in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), 2019, pp. 11446–11456, is problematic mathematically. Concretely, the original error bound for arbitrary sequences of learning rate is unfortunately incorrect, leading to an invalidated upper bound in the convergence theorem for the algorithm. As evidences, we explicitly provide several counter-examples, for both convex and nonconvex cases, to show the incorrectness of the error bound. We fix the issue by providing a new error bound and its corresponding proof, leading to a new convergence theorem for the dist-EF-SGD algorithm, and therefore recovering its mathematical analysis.

1. INTRODUCTION

1.1. Background

For training deep neural networks over large-scale and distributed datasets, distributed stochastic gradient descent (distributed SGD) is a vital method. In distributed SGD, a central server updates the model parameters using information transmitted from distributed workers, as illustrated in Figure 1.

Communication between the server and distributed workers can be a bottleneck in distributed SGD. Alleviating the bottleneck is a considerable concern of the community, so that variants of distributed SGD using gradient compression have been proposed to reduce the communication cost between workers and the server.

Recently, Zheng *et al.* [1] proposed an algorithm named dist-EF-SGD recalled in Algorithm 1, in which gradients are compressed before transmission, and errors between real and compressed gradients in one step of the algorithm are re-used in future steps.

1.2. Our Contributions

In this paper, we point out a flaw in the convergence proof of Algorithm 1 given in Zheng *et al.* [1]. We then fix the flaw by providing a new convergence theorem with a new proof for Algorithm 1.

Zheng *et al.* [1] stated the following theorem for any sequence of learning rate $\{\eta_t\}$.

Theorem A (Theorem 1 of [1], problematic). *Suppose that Assumptions 1-3 (given together with related notations in Section 2) hold. Assume that the learning rate* $0 < \eta_t < \frac{3}{2L}$ *for all* $t \geqslant 0$. *For sequence* x_t *generated from Algorithm 1, we have the following upper bound on the expected Euclidean norm of gradients:*

$$\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right] \leqslant \frac{4(f(x_0)-f^*)}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)} + \frac{2L\sigma^2}{M}\sum_{t=0}^{T-1}\frac{\eta_t^2}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)} + \frac{32L^2(1-\delta)G^2}{\delta^2}\left[1+\frac{16}{\delta^2}\right]\sum_{t=0}^{T-1}\frac{\eta_t\eta_{t-1}^2}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}$$

where $o \in \{0,\ldots,T-1\}$ *is an index such that the probability*

$$\Pr(o=k) = \frac{\eta_k\left(3-2L\eta_k\right)}{\sum_{t=0}^{T-1}\eta_t\left(3-2L\eta_t\right)}, \forall k = 0,\ldots,T-1.$$

Problem in Theorem A. Unfortunately, the proof of Theorem A as given in [1] becomes invalidated when the learning rate sequence $\{\eta_t\}$ is decreasing. In that proof, a lemma is employed to handle decreasing learning rate sequences. However, in Section 3 we

*Corresponding author. Email: tranthiphuong@tdtu.edu.vn

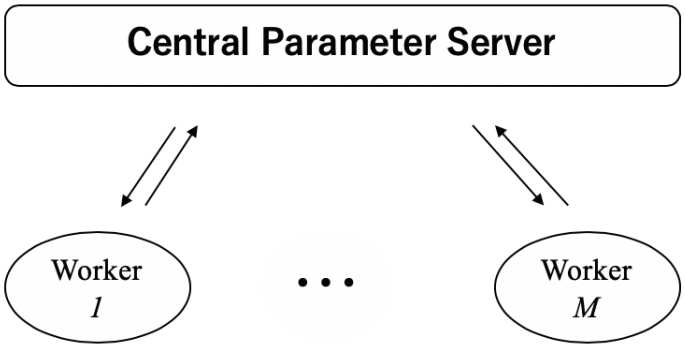

Figure 1 The computation model of distributed SGD. Multiple workers communicate with a central parameter server synchronously. Each worker, after local computations on its data, uploads selected results to the server. The server aggregates all uploaded results from the workers, and sends back the aggregated result from its computations to all workers. These are iterated for multiple rounds.

present several counter-examples showing that lemma does not hold. We move on to fix that lemma and finally obtain the following result as our correction for Theorem A.

Theorem 1. *(Our correction for Theorem A) With all notations and assumptions are identical to Theorem A, we have*

$$\mathbf{E}\left[\left\|\nabla f\left(x_o\right)\right\|^2\right] \leqslant \frac{4\left(f\left(x_0\right)-f^\star\right)}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)} + \frac{2L\sigma^2}{M}\sum_{t=0}^{T-1}\frac{\eta_t^2}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)} + \frac{8(1-\delta)(2-\delta)G^2L^2}{\delta\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\sum_{t=0}^{T-1}\eta_t\eta_{t-1}^2\sum_{k=0}^{t-1}\frac{\eta_{t-1-k}^2}{\eta_{t-1}^2}\alpha^k + \frac{16(1-\delta)(2-\delta)^3G^2L^2}{\delta^2\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\times \sum_{t=0}^{T-1}\eta_t\eta_{t-1}^2\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_{t-1}^2}\alpha^k,$$

where $\alpha = 1-\frac{\delta}{2}$ *and* $o \in \{0,\dots,T-1\}$ *is an index*

Algorithm 1 Distributed SGD with Error-Feedback (dist-EF-SGD) [1]

Input: Loss function $\mathscr{L}$, learning rate $\{\eta_t\}$ with $\eta_{-1}=0$; number of workers M; compressor $\mathscr{C}(\cdot)$

Initialize: initial parameter $x_0 \in \mathbb{R}^d$; error $e_{0,i}=0\in\mathbb{R}^d$ on each worker i; error $\tilde{e}_0=0\in\mathbb{R}^d$ on server

for $t\in\{0,\dots,T-1\}$ **do**

• **on each worker** $1\leqslant i\leqslant M$:

pick data $\xi_{t,i}$ from the dataset

$g_{t,i}=\nabla\mathscr{L}\left(x_t,\xi_{t,i}\right)$ ▷ stochastic gradient

$p_{t,i}=g_{t,i}+\frac{\eta_{t-1}}{\eta_t}e_{t,i}$ ▷ gradient added with previous error

push $\Delta_{t,i}=\mathscr{C}\left(p_{t,i}\right)$ to server ▷ gradient compression at worker, and transmission

pull $\tilde{\Delta}_t$ from server

$x_{t+1}=x_t-\eta_t\tilde{\Delta}_t$ ▷ local weight update

$e_{t+1,i}=p_{t,i}-\Delta_{t,i}$ ▷ local error-feedback to next step

• **on central parameter server:**

pull $\Delta_{t,i}$ from each worker i

$\tilde{p}_t=\frac{1}{M}\sum_{i=1}^{M}\Delta_{t,i}+\frac{\eta_{t-1}}{\eta_t}\tilde{e}_t$ ▷ gradient average with error

push $\tilde{\Delta}_t=\mathscr{C}\left(\tilde{p}_t\right)$ to each worker ▷ gradient compression at server

$\tilde{e}_{t+1}=\tilde{p}_t-\tilde{\Delta}_t$ ▷ error on server

end for

such that the probability

$$\Pr(o=k)=\frac{\eta_k\left(3-2L\eta_k\right)}{\sum_{t=0}^{T-1}\eta_t\left(3-2L\eta_t\right)},\forall k=0,\dots,T-1.$$

In addition, we show that the upper bound in Theorem 1 becomes $O\left(\frac{1}{\sqrt{MT}}\right)$ for a proper choice of decreasing sequence $\{\eta_t\}$ in Corollary 2. Moreover the upper bound in Theorem 1 matches previous results given in Zheng *et al.* [1] when $\{\eta_t\}$ is nondecreasing (Corollary 1).

1.3. Paper Roadmap

We begin with notations and settings in Section 2. In Section 3, we provide counter-examples to justify the issue in [1] for both non-convex and convex cases. We then correct the issue in Section 4 and then present a proof for Theorem 1 in Section 5.

1.4. Related Works

The use of gradient compression for reducing the communication cost is widely considered in distributed machine learning recently. One line of research is to compress the gradient only on the worker

side before sending the result to the parameter server, namely one-side compression. The parameter server receives and aggregates these results and sends back the aggregated result to all workers. Some recent papers such as [2–5] are in this line of research.

Another line of research uses gradient compression on both workers and server, namely two-side compression. In these two-side compression methods, the workers send the compressed local gradients or some corrected forms of them to the parameter server, and the parameter server compresses the aggregated result before sending it back to all workers. Papers [1,6,7] use two-side compression with an identical method of gradient compression for both workers and the parameter server. Paper [8] considers two-side compression with flexible compression for both workers and the parameter server.

2. PRELIMINARIES

Let $\langle\cdot,\cdot\rangle$ be the inner product of vectors. The Cauchy–Schwarz inequality states that for all vectors u, v it holds that $|\langle u,v\rangle|^2 \leqslant \langle u,u\rangle \times \langle v,v\rangle$. The Young inequality $\gamma > 0$ (sometimes called the Peter–Paul inequality) states that $(a+b)^2 \leqslant (1+\gamma)a^2 + (1+1/\gamma)b^2$ $\forall a, b \in \mathbb{R}$. Let $\|\cdot\|$ be the Euclidean norm of a vector.

For completeness, we recall the algorithm of Zheng *et al.* [1] in Algorithm 1 and its explanation as follows. At iteration t, the scale $\frac{\eta_{t-1}}{\eta_t}$ of the local accumulated error vector $e_{t,i}$ is added to the gradient $g_{t,i}$ (line 7 of Algorithm 1) for the compression step. Each worker i stores these local accumulated error vector $e_{t,i}$ and local corrected gradient vector $p_{t,i}$ for the next iteration. The compressed $\Delta_{t,i}$ of $p_{t,i}$ are pushed to the parameter server. The parameter server aggregates these $\Delta_{t,i}$ and uses the aggregated result to update the global error-corrected vector $\tilde{p}_t$ (line 14 of Algorithm 1), which in turn is used to update the global accumulated error vector $\tilde{e}_{t+1}$ (line 16 of Algorithm 1). Each worker receives the compressed $\tilde{\Delta}_t$ of $\tilde{p}_t$ from the parameter server and uses it to update the parameter x_{t+1}.

In order to construct Algorithm 1, Zheng *et al.* [1] used the idea of Karimireddy *et al.* [9] that combined gradient compression with error correction. The innovative ideas of Zheng *et al.* [1] were to apply compression on the parameter server and to use the scale $\frac{\eta_{t-1}}{\eta_t}$ in line 7 of Algorithm 1. Unfortunately the scale $\frac{\eta_{t-1}}{\eta_t}$ caused an issue in the proof of convergence theorem of Algorithm 1. We examine this issue in details in Section 3.

2.1. Compressor and Assumptions

Following [5,9], an operator $\mathscr{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a δ-compressor for a number $\delta \in (0,1]$ if

$$\mathbf{E}_{\mathscr{C}} \parallel \mathscr{C}(x) - x \parallel^2 \leqslant (1-\delta) \parallel x \parallel^2 \tag{1}$$

where the expectation $\mathbf{E}_{\mathscr{C}}$ is taken over the randomness of $\mathscr{C}$.

Given a loss function $\mathcal{L}$, define $f(x) = \mathbf{E}_{\xi}[\mathcal{L}(x,\xi)]$ where $x \in \mathbb{R}^d$ is the (neural network) model parameters, and ξ is the data batch drawn from some unknown distribution. We consider the following assumptions on f, which are standard and have been used in previous works [1,9].

Assumption 1. *f is lower-bounded, i.e., $f^* = \inf_{x\in\mathbb{R}^d} f(x) < \infty$, and L-smooth, i.e., f is differentiable and there exists a constant $L \geqslant 0$ such that*

$$\parallel \nabla f(x) - \nabla f(y) \parallel \leqslant L \parallel x - y \parallel, \quad \forall x, y \in \mathbb{R}^d. \tag{2}$$

By [10], the L-smooth condition in (2) implies that $\forall x, y \in \mathbb{R}^d$,

$$f(x) \leqslant f(y) + \langle \nabla f(y), x - y\rangle + \frac{L}{2}\|x-y\|^2. \tag{3}$$

Assumption 2. *Let $\mathbf{E}_t$ denote the expectation at iteration t. Then $\mathbf{E}_t[g_{t,i}] = \nabla f(x_t)$ and the stochastic gradient $g_{t,i}$ has bounded gradient, i.e.,*

$$\mathbf{E}_t\left[\left\|g_{t,i} - \nabla f(x_t)\right\|^2\right] \leqslant \sigma^2.$$

Assumption 3. *The full gradient ∇f is uniformly bounded, i.e., $\left\|\nabla f(x_t)\right\|^2 \leqslant \omega^2$.*

Under Assumptions 2 and 3, we have

$$\mathbf{E}_t\left[\|g_{t,i}\|^2\right] \leqslant G^2 = \sigma^2 + \omega^2, \tag{4}$$

because $\mathbf{E}_t\left[\left\|g_{t,i} - \nabla f(x_t)\right\|^2\right] \leqslant \sigma^2$, $\left\|\nabla f(x_t)\right\|^2 \leqslant \omega^2$, and the fact that $\mathbf{E}\left[\|X - \mathbf{E}[X]\|^2\right] + \|\mathbf{E}[X]\|^2 = \mathbf{E}\left[\|X\|^2\right]$.

2.2. Supporting Lemmas

We need a few supporting lemmas for proving Theorem 1.

Lemma 1. *Let $0 < M \in \mathbb{N}$ and $x_i \in \mathbb{R}^d$. Then*

$$\left\|\frac{1}{M}\sum_{i=1}^{M} x_i\right\|^2 \leqslant \frac{1}{M}\sum_{i=1}^{M}\|x_i\|^2.$$

Proof. Since $x_i \in \mathbb{R}^d$, x_i has the form $x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d}) \in \mathbb{R}^d$. We have

$$\begin{aligned}\left\|\frac{1}{M}\sum_{i=1}^{M} x_i\right\|^2 &= \frac{1}{M^2}\left\|\sum_{i=1}^{M} x_i\right\|^2 \\ &= \frac{1}{M^2}\left\|\left(\sum_{i=1}^{M} x_{i,1}, \sum_{i=1}^{M} x_{i,2}, \ldots, \sum_{i=1}^{M} x_{i,d}\right)\right\|^2 \\ &= \frac{1}{M^2}\sum_{j=1}^{d}\left(\sum_{i=1}^{M} x_{i,j}\right)^2.\end{aligned}$$

Applying the Cauchy–Schwarz inequality on $\left(\sum_{i=1}^{M} x_{i,j}\right)^2$ gives us

$$\left(\sum_{i=1}^{M} x_{i,j}\right)^2 \leqslant M\sum_{i=1}^{M} x_{i,j}^2.$$

Therefore

$$\begin{aligned}\left\|\frac{1}{M}\sum_{i=1}^{M}x_i\right\|^2 &\leqslant \frac{1}{M}\sum_{j=1}^{d}\left(\sum_{i=1}^{M}x_{i,j}^2\right)\\ &= \frac{1}{M}\sum_{i=1}^{M}\left(\sum_{j=1}^{d}x_{i,j}^2\right)\\ &= \frac{1}{M}\sum_{i=1}^{M}\|x_i\|^2\end{aligned}$$

which ends the proof. □

Lemma 2. *Let $\{a_t\}$, $\{\alpha_t\}$, $\{\beta_t\}$ be non-negative sequences in $\mathbb{R}$ such that $a_0 = 0$ and, for all $t \geqslant 0$,*

$$a_{t+1} \leqslant \alpha_t a_t + \beta_t. \tag{5}$$

Then

$$a_{t+1} \leqslant \beta_t + \sum_{j=1}^{t}\prod_{i=j}^{t}\alpha_i\beta_{j-1}.$$

In particular, if $\beta_t = \beta$ for all t, then

$$a_{t+1} \leqslant \beta\left(1 + \sum_{j=1}^{t}\prod_{i=j}^{t}\alpha_i\right).$$

Proof. By (5), we have

$$a_1 \leqslant \alpha_0 a_0 + \beta_0 = \beta_0.$$

Proving by induction, assume that then we have

$$a_t \leqslant \beta_{t-1} + \sum_{j=1}^{t-1}\prod_{i=j}^{t-1}\alpha_i\beta_{j-1}, \tag{6}$$

then we have

$$\begin{aligned}a_{t+1} &\leqslant \alpha_t a_t + \beta_t \quad \text{(by (5))}\\ &\leqslant \beta_t + \alpha_t\left(\beta_{t-1} + \sum_{j=1}^{t-1}\prod_{i=j}^{t-1}\alpha_i\beta_{j-1}\right) \quad \text{(by (6))}\\ &= \beta_t + \alpha_t\beta_{t-1} + \sum_{j=1}^{t-1}\prod_{i=j}^{t}\alpha_i\beta_{j-1}\\ &= \beta_t + \sum_{j=1}^{l}\prod_{i=j}^{l}\alpha_i\beta_{j-1},\end{aligned}$$

which ends the proof. □

3. THE ISSUE IN ZHENG *ET AL.* [1]

In order to prove the convergence theorem for Algorithm 1, Zheng *et al.* [1] have used the following lemma.

Lemma A (Lemma 2 of [1], incorrect). *For any $t \geqslant 0$, $\tilde{e}_t, e_{t,i}, \eta_t$ from Algorithm 1, compressor parameter δ at (1), and gradient bound G at (4),*

$$\mathbf{E}\left[\left\|\tilde{e}_t + \frac{1}{M}\sum_{i=1}^{M}e_{t,i}\right\|^2\right] \leqslant \frac{8(1-\delta)G^2}{\delta^2}\left(1 + \frac{16}{\delta^2}\right).$$

Intuitively, Lemma A can become incorrect because its right-hand side only depends on the gradient bound G and compressor parameter δ, and *does not* capture the scaling factor η_{t-1}/η_t of the errors $\tilde{e}_t$ and $e_{t,i}$ of Algorithm 1. More formally, the following claim states that Lemma A is invalidated when the learning rate sequence $\{\eta_t\}$ is decreasing.

Claim 1. *Lemma A (i.e., Lemma 2 of [1]) does not hold. More precisely, referring to Algorithm 1, there exist a sequence of loss functions $\mathcal{L}(x_t, \xi)$, a decreasing sequence $\{\eta_t\}_{t\geqslant -1}$, a number δ with respect to a compressor $\mathscr{C}$, and a step t such that*

$$\mathbf{E}\left[\left\|\tilde{e}_t + \frac{1}{M}\sum_{i=1}^{M}e_{t,i}\right\|^2\right] > \frac{8(1-\delta)G^2}{\delta^2}\left(1 + \frac{16}{\delta^2}\right). \tag{7}$$

Claim 1 is justified by the following counter-examples, in which we intentionally utilize the fact that the quotient η_{t-1}/η_t as in line 7 of Algorithm 1 can be large with *decreasing* learning rate sequences.

Counter-example 1. (Convex case) For $t \geqslant 0$ and $x_t, \xi \in \mathbb{R}$, we consider the sequence of loss functions

$$\mathcal{L}(x_t, \xi) = \varphi(x_t) = \frac{1}{4}x_t$$

in the constraint set $[-1, 1]$, the decreasing sequence of learning rate $\{\eta_t\}_{t\geqslant -1}$ with

$$\eta_{-1} = 0, \left\{\eta_t = \frac{1}{48t+2}\right\}_{t\geqslant 0},$$

the compressor $\mathscr{C} : \mathbb{R} \to \mathbb{R}$ such that $\forall x \in \mathbb{R}$

$$\mathscr{C}(x) = \frac{x}{0.77}.$$

Then at $t = 1$, Claim 1 holds true.

Proof. (Justification of Counter-example 1) It is trivial that the loss function $\mathcal{L}$ satisfies all the Assumptions 1, 2, and 3. The upper bound gradient of f is $G = \frac{1}{4}$ because we have $g_{t,i} = \frac{1}{4}\forall t, i$.

- The function $\mathscr{C}$ with $\mathscr{C}(x) = \frac{x}{0.77}$ is a compressor with respect to $\delta = 0.9$. Indeed, we have

$$\|\mathscr{C}(x) - x\|^2 \leqslant (1-\delta)\|x\|^2$$

$$\Leftrightarrow \left\|\tfrac{x}{0.77} - x\right\|^2 \leqslant (1-\delta)\|x\|^2$$

$$\Leftrightarrow \left\|\tfrac{1}{0.77} - 1\right\|^2 \leqslant 1-\delta.$$

The last inequality is equivalent to

$$\delta \leqslant 1 - \left\|\frac{1}{0.77} - 1\right\|^2 = 0.9107775341541.$$

Therefore $\delta = 0.9$ suffices.

To continue, let us consider the number of workers is $M = 2$. Initially $e_{0,i} = 0$ on each worker $i \in \{1,2\}$ and $\tilde{e}_0 = 0$ on server. Because the stochastic gradients are the same on each worker, the results of computations on each worker are the same. So it is sufficient to consider the computations on worker 1 in details.

- At $t = 0$ we have the computations on the workers and the server as follows.
  - On worker 1:

$$\begin{aligned} p_{0,1} &= g_{0,1} + \tfrac{\eta_{-1}}{\eta_0} e_{0,1} \\ &= g_{0,1} = \tfrac{1}{4}, \\ \Delta_{0,1} &= \mathscr{C}\left(p_{0,1}\right) \\ &= \tfrac{p_{0,1}}{0.77} = 0.3246753246753, \\ e_{1,1} &= p_{0,1} - \Delta_{0,1} = -0.07467532467532. \end{aligned}$$

  - On worker 2, $p_{0,2} = p_{0,1}$, $\Delta_{0,2} = \Delta_{0,1}$, and $e_{1,2} = e_{1,1}$.
  - On server:

$$\begin{aligned} \tilde{p}_0 &= \tfrac{1}{2}\left(\Delta_{0,1} + \Delta_{0,2}\right) + \tfrac{\eta_{-1}}{\eta_0}\tilde{e}_0 \\ &= \Delta_{0,1} = 0.3246753246753, \\ \tilde{\Delta}_0 &= \mathscr{C}\left(\tilde{p}_0\right) \\ &= \tfrac{\tilde{p}_0}{0.77} = 0.42165626581210, \\ \tilde{e}_1 &= \tilde{p}_0 - \tilde{\Delta}_0 = -0.09698094113678. \end{aligned}$$

- At $t = 1$ we have the computations on the workers and the server as follows:
  - On worker 1:

$$\begin{aligned} p_{1,1} &= g_{1,1} + \tfrac{\eta_0}{\eta_1} e_{1,1} \\ &= -1.6168831168831, \\ \Delta_{1,1} &= \mathscr{C}\left(p_{1,1}\right) \\ &= \tfrac{p_{1,1}}{0.77} = -2.0998482037443, \\ e_{2,1} &= p_{1,1} - \Delta_{1,1} = 0.48296508686119. \end{aligned}$$

  - On worker 2: $p_{1,2} = p_{1,1}$, $\Delta_{1,2} = \Delta_{1,1}$, and $e_{2,2} = e_{2,1}$.
  - On server:

$$\begin{aligned} \tilde{p}_1 &= \tfrac{1}{2}\left(\Delta_{1,1} + \Delta_{1,2}\right) + \tfrac{\eta_0}{\eta_1}\tilde{e}_1 \\ &= \Delta_{1,1} + \tfrac{\eta_0}{\eta_1}\tilde{e}_1, \\ &= -4.52437173216, \\ \tilde{\Delta}_1 &= \mathscr{C}\left(\tilde{p}_1\right) \\ &= \tfrac{\tilde{p}_1}{0.77} = -5.8758074443687, \\ \tilde{e}_2 &= \tilde{p}_1 - \tilde{\Delta}_1 = 1.3514357122048. \end{aligned}$$

Now we compute the left- and right-hand sides of (7) with $t = 2$.

$$\begin{aligned} &\left\|\tilde{e}_2 + \tfrac{1}{2}\left(e_{2,1} + e_{2,2}\right)\right\|^2 \\ &= \left\|\tilde{e}_2 + e_{2,1}\right\|^2 \\ &= 3.365026291613992 \end{aligned}$$

and

$$\begin{aligned} \tfrac{8(1-\delta)G^2}{\delta^2}\left(1 + \tfrac{16}{\delta^2}\right) &= \tfrac{8(0.1)\left(\frac{1}{4}\right)^2}{0.9^2}\left(1 + \tfrac{16}{0.9^2}\right) \\ &= 1.2810547172687086. \end{aligned}$$

Thus

$$\left\|\tilde{e}_2 + \frac{1}{2}\left(e_{2,1} + e_{2,2}\right)\right\|^2 > \frac{8(1-\delta)G^2}{\delta^2}\left(1 + \frac{16}{\delta^2}\right),$$

and then Claim 1 follows. □

Counter-example 2. (Convex case) For $t \geqslant 0$ and $x_t, \xi \in \mathbb{R}$, we consider the sequence of loss functions

$$\mathcal{L}\left(x_t, \xi\right) = \varphi\left(x_t\right) = x_t^2$$

in the constraint set $[-1, 1]$, the decreasing sequence of learning rate $\{\eta_t\}_{t \geqslant -1}$ with

$$\eta_{-1} = 0, \left\{\eta_t = \frac{3}{4}\left(\frac{1}{26t+2}\right)\right\}_{t \geqslant 0},$$

and the following compressor $\mathscr{C} : \mathbb{R} \to \mathbb{R}$ with parameter $\delta = 0.9$ as in counter-example 1,

$$\mathscr{C}(x) = \frac{x}{0.77}.$$

Then at $t = 1$, Claim 1 holds true.

Proof. (Justification of Counter-example 2) It is trivial that the loss function $\mathcal{L}$ is 2-smooth and $\mathcal{L}$ satisfies all the Assumptions 1, 2, and 3. Since $\nabla f(x) = 2x$ and $g_{t,i} = 2x_t, \forall t, i$, we have the upper bound gradient of f in the constraint set $[-1, 1]$ as $G = 2$. Let us consider the number of workers $M = 2$. Initially $e_{0,i} = 0$ on each worker $i \in \{1, 2\}$ and $\tilde{e}_0 = 0$ server. Let us take $x_0 = 1$. Because the stochastic gradients are the same on each worker, the results of computations on each worker are the same. So it is sufficient to consider the computations on worker 1 in details.

- At $t = 0$ we have the computations on the workers and the server as follows:
  - On worker 1:

$$\begin{aligned} p_{0,1} &= g_{0,1} + \tfrac{\eta_{-1}}{\eta_0} e_{0,1} \\ &= g_{0,1} = 2, \\ \Delta_{0,1} &= \mathscr{C}\left(p_{0,1}\right) \\ &= \tfrac{p_{0,1}}{0.77} = 2.5974025974025974, \\ e_{1,1} &= pp_{0,1} - \Delta_{0,1} = -0.5974025974025974. \end{aligned}$$

  - On worker 2, $p_{0,2} = p_{0,1}$, $\Delta_{0,2} = \Delta_{0,1}$, and $e_{1,2} = e_{1,1}$.
  - On server:

$$\begin{aligned} \tilde{p}_0 &= \tfrac{1}{2}\left(\Delta_{0,1} + \Delta_{0,2}\right) + \tfrac{\eta_{-1}}{\eta_0}\tilde{e}_0 \\ &= \Delta_{0,1} = 2.5974025974025974, \\ \tilde{\Delta}_0 &= \mathscr{C}\left(\tilde{p}_0\right) \\ &= \tfrac{\tilde{p}_0}{0.77} = 3.3732501264968797, \\ x_1 &= x_0 - \eta_0 \tilde{\Delta}_0 \\ &= -0.26496879743632995, \\ \tilde{e}_1 &= \tilde{p}_0 - \tilde{\Delta}_0 = -0.7758475290942823. \end{aligned}$$

- At $t = 1$ we have the computations on the workers and the server as follows:
  - On worker 1:

$$\begin{aligned} g_{1,1} &= \nabla f\left(x_1\right) = -0.5299375948726599, \\ p_{1,1} &= g_{1,1} + \tfrac{\eta_0}{\eta_1} e_{1,1} \\ &= -8.893573958509023, \\ \Delta_{1,1} &= \mathscr{C}\left(p_{1,1}\right) \\ &= \tfrac{p_{1,1}}{0.77} = -11.550096050011717, \\ e_{2,1} &= p_{1,1} - \Delta_{1,1} = 2.656522091502694. \end{aligned}$$

  - On worker 2, $p_{1,2} = p_{1,1}$, $\Delta_{1,2} = \Delta_{1,1}$, and $e_{2,2} = e_{2,1}$.
  - On server:

$$\begin{aligned} \tilde{p}_1 &= \tfrac{1}{2}\left(\Delta_{1,1} + \Delta_{1,2}\right) + \tfrac{\eta_0}{\eta_1}\tilde{e}_1 \\ &= \Delta_{1,1} + \tfrac{\eta_0}{\eta_1}\tilde{e}_1 \\ &= -22.41196145733167, \\ \tilde{\Delta}_1 &= \mathscr{C}\left(\tilde{p}_1\right) \\ &= \tfrac{\tilde{p}_1}{0.77} = -29.106443451080093, \\ \tilde{e}_2 &= \tilde{p}_1 - \tilde{\Delta}_1 = 6.694481993748422. \end{aligned}$$

Now we compute the left- and right-hand sides of (7) with $t = 2$. We have

$$\begin{aligned} &\left\|\tilde{e}_2 + \tfrac{1}{2}\left(e_{2,1} + e_{2,2}\right)\right\|^2 \\ &= \left\|\tilde{e}_2 + e_{2,1}\right\|^2 \\ &= 87.44127740238307 \end{aligned}$$

and

$$\begin{aligned} \tfrac{8(1-\delta)G^2}{\delta^2}\left(1 + \tfrac{16}{\delta^2}\right) &= \tfrac{8(0.1)2^2}{0.9^2}\left(1 + \tfrac{16}{0.9^2}\right) \\ &= 81.98750190519735. \end{aligned}$$

Thus

$$\left\|\tilde{e}_2 + \frac{1}{2}\left(e_{2,1} + e_{2,2}\right)\right\|^2 > \frac{8(1-\delta)G^2}{\delta^2}\left(1 + \frac{16}{\delta^2}\right),$$

and then Claim 1 follows. □

Counter-example 3. (nonconvex case) For $t \geqslant 0, x_t, \xi \in \mathbb{R}$, we consider the sequence of loss functions

$$\mathscr{L}(x_t, \xi) = \varphi(x_t) = \frac{1}{1 + e^{-x_t}},$$

the decreasing sequence of learning rate $\{\eta_t\}_{t\geqslant -1}$ with

$$\eta_{-1} = 0, \left\{\eta_t = \frac{3}{2}\left(\frac{1}{48t + 2}\right)\right\}_{t\geqslant 0},$$

and the following compressor $\mathscr{C} : \mathbb{R} \to \mathbb{R}$ with parameter $\delta = 0.9$ as in counter-example 1,

$$\mathscr{C}(x) = \frac{x}{0.77}.$$

Then at $t = 1$, Claim 1 holds true.

Proof. (Justification of Counter-example 3) First, we check that Assumptions 1-3 are satisfied.

- The function ϕ is lower-bounded because $0 \leqslant \phi(x) \leqslant 1, \forall x \in \mathbb{R}$. The upper bound of $\nabla\phi(x)$ is $G = \frac{1}{4}$, since

$$\begin{aligned} \nabla\varphi(x) \leqslant \tfrac{1}{4} &\Leftrightarrow \varphi(x)(1-\varphi(x)) \leqslant \tfrac{1}{4} \\ &\Leftrightarrow -\varphi(x)^2 + \varphi(x) - \tfrac{1}{4} \leqslant 0 \\ &\Leftrightarrow -\left(\varphi(x) - \tfrac{1}{2}\right)^2 \leqslant 0, \end{aligned} \tag{8}$$

which holds true for all $x \in \mathbb{R}$.

- The function ϕ is L-smooth, with $L = 1$. Indeed, for all $x, y \in \mathbb{R}$, we have

$$\begin{aligned} &|\nabla\varphi(x) - \nabla\varphi(y)| \\ &= |\varphi(x)(1-\varphi(x)) - \varphi(y)(1-\varphi(y))| \\ &= |\varphi(x) - \varphi(y) + (\varphi(y) - \varphi(x))(\varphi(y) + \varphi(x))| \\ &= |[\varphi(x) - \varphi(y)][1 - (\varphi(x) + \varphi(y))]|. \end{aligned}$$

Since $0 \leqslant \varphi(\xi) \leqslant 1\,(\forall\xi)$, we have $0 \leqslant \varphi(x) + \varphi(y) \leqslant 2$. Therefore $-1 \leqslant 1 - (\varphi(x) + \varphi(y)) \leqslant 1$, and hence

$$|[\varphi(x) - \varphi(y)][1 - (\varphi(x) + \varphi(y))]| \leqslant |\varphi(x) - \varphi(y)|.$$

This means that in order to prove $|\nabla\varphi(x) - \nabla\varphi(y)| \leqslant |x - y|$, it is sufficient to prove

$$|\varphi(x) - \varphi(y)| \leqslant |x - y|. \tag{9}$$

If $x \geqslant y$, we obtain $\phi(x) \geqslant \phi(y)$. Therefore

$$\begin{aligned} (9) &\Leftrightarrow \varphi(x) - \varphi(y) \leqslant x - y \\ &\Leftrightarrow \varphi(x) - x \leqslant \varphi(y) - y \end{aligned}$$

Let $\phi(x) = \phi(x) - x$. Because $\nabla\phi(x) = \nabla\varphi(x) - 1 \leqslant \frac{1}{4} - 1 < 0$ by (8), we have $\phi(x)$ is a decreasing function. Therefore the inequality $\varphi(x) - x \leqslant \varphi(y) - y$ holds true, and hence (9) is proven. By the same technique, we obtain (9) for the case $x < y$.

To continue, let us consider the number of workers $M = 2$. We initialize $e_{0;i} = 0$ on each worker $i \in \{1, 2\}$ and $\tilde{e}_0 = 0$ on server. Let us take $x_0 = 0$. Because the stochastic gradients are the same on each worker, the results of computations on each worker are the same. So it is sufficient to consider the computations on worker 1 in details. We have

$$g_{0,1} = g_{0,2} = \nabla\varphi\left(x_0\right) = 0.25.$$

- At $t = 0$ we have the computations on the workers and the server as follows:
  - On worker 1:

$$\begin{aligned} p_{0,1} &= g_{0,1} + \tfrac{\eta_{-1}}{\eta_0} e_{0,1} \\ &= g_{0,1} = 0.25, \\ \Delta_{0,1} &= \mathscr{C}\left(p_{0,1}\right) \\ &= \tfrac{p_{0,1}}{0.77} = 0.3246753246753247, \\ e_{1,1} &= p_{0,1} - \Delta_{0,1} = -0.07467532467532467. \end{aligned}$$

  - On worker 2, $p_{0,2} = p_{0,1}$, $\Delta_{0,2} = \Delta_{0,1}$, and $e_{1,2} = e_{1,1}$.
  - On server:

$$\begin{aligned} \tilde{p}_0 &= \tfrac{1}{2}\left(\Delta_{0,1} + \Delta_{0,2}\right) + \tfrac{\eta_{-1}}{\eta_0}\tilde{e}_0 \\ &= \Delta_{0,1} = 0.3246753246753247, \\ \tilde{\Delta}_0 &= \mathscr{C}\left(\tilde{p}_0\right) \\ &= \tfrac{\tilde{p}_0}{0.77} = 0.42165626581210996, \\ x_1 &= x_0 - \eta_0\tilde{\Delta}_0 \\ &= -0.3162421993590825, \\ \tilde{e}_1 &= \tilde{p}_0 - \tilde{\Delta}_0 = -0.09698094113678529. \end{aligned}$$

- At $t = 1$ we have the computations on the workers and the server as follows.
  - On worker 1:

$$\begin{aligned} g_{1,1} &= \nabla\varphi\left(x_1\right) = 0.243852158038919 \\ p_{1,1} &= g_{1,1} + \tfrac{\eta_0}{\eta_1} e_{1,1} \\ &= -1.6230309588441978, \\ \Delta_{1,1} &= \mathscr{C}\left(p_{1,1}\right) \\ &= \tfrac{p_{1,1}}{0.77} = -2.1078324140833735, \\ e_{2,1} &= p_{1,1} - \Delta_{1,1} = 0.4848014552391757. \end{aligned}$$

  - On worker 2, $p_{1,2} = p_{1,1}$, $\Delta_{1,2} = \Delta_{1,1}$, and $e_{2,2} = e_{2,1}$.
  - On server:

$$\begin{aligned}\tilde{p}_1 &= \tfrac{1}{2}\left(\Delta_{1,1}+\Delta_{1,2}\right)+\tfrac{\eta_0}{\eta_1}\tilde{e}_1\\ &= \Delta_{1,1}+\tfrac{\eta_0}{\eta_1}\tilde{e}_1\\ &= -4.532355942503006,\\ \tilde{\Delta}_1 &= \mathscr{C}\left(\tilde{p}_1\right)\\ &= \tfrac{\tilde{p}_1}{0.77} = -5.886176548705203,\\ \tilde{e}_2 &= \tilde{p}_1-\tilde{\Delta}_1 = 1.3538206062021967.\end{aligned}$$

Now we compute the left- and right-hand sides of (7) with $t=2$. We have

$$\begin{aligned}\left\|\tilde{e}_2+\tfrac{1}{2}\left(e_{2,1}+e_{2,2}\right)\right\|^2 &= \left\|\tilde{e}_2+e_{2,1}\right\|^2\\ &= 3.3805310848189216\end{aligned}$$

and, with $\delta=0.9$,

$$\begin{aligned}\tfrac{8(1-\delta)G^2}{\delta^2}\left(1+\tfrac{16}{\delta^2}\right) &= \tfrac{8(0.1)\left(\frac{1}{4}\right)^2}{0.9^2}\left(1+\tfrac{16}{0.9^2}\right)\\ &= 1.2810547172687086.\end{aligned}$$

Thus

$$\left\|\tilde{e}_2+\frac{1}{2}\left(e_{2,1}+e_{2,2}\right)\right\|^2 > \frac{8(1-\delta)G^2}{\delta^2}\left(1+\frac{16}{\delta^2}\right),$$

and hence Claim 1 follows. □

4. CORRECTING THE ERROR BOUND OF ZHENG *ET AL.* [1]

In general, the error $\mathbf{E}\left[\left\|\tilde{e}_{t+1}+\frac{1}{M}\sum_{i=1}^{M}e_{t+1,i}\right\|^2\right]$ is bounded as follows:

Theorem 2. *(Fix for Lemma 2 of [1]) With $\tilde{e}_t, e_{t,i}, \eta_t$, and δ from Algorithm 1, for arbitrary $\{\eta_t\}$, we have*

$$\begin{aligned}&\mathbf{E}\left[\left\|\tilde{e}_{t+1}+\frac{1}{M}\sum_{i=1}^{M}e_{t+1,i}\right\|^2\right]\\ &\leqslant \frac{2(1-\delta)(2-\delta)G^2}{\delta}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k\\ &\quad+\frac{4(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{j=0}^{t}\alpha^{t-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_t^2}\alpha^k,\end{aligned}$$

where $\alpha=1-\frac{\delta}{2}$ and gradient bound G is at (4).

Remark 1. [Sanity check of the new upper bound] The right-hand side of Theorem 2 can become large together with decreasing leaning rate sequences. Therefore, the error bounds of the sequences in counter-examples 1-3 do satisfy Theorem 2. Indeed, the upper bound on the error in Theorem 2 at $t=1$ is

$$\begin{aligned}U = &\tfrac{2(2-\delta)(1-\delta)G^2}{\delta}\left(1+\tfrac{\eta_0^2}{\eta_1^2}\left(1-\tfrac{\delta}{2}\right)\right)\\ &+\tfrac{4(1-\delta)(2-\delta)^3G^2}{\delta^2}\left(1+2\tfrac{\eta_0^2}{\eta_1^2}\left(1-\tfrac{\delta}{2}\right)\right).\end{aligned}$$

Concretely, at sanity check,

• in counter-example 1:

$$U = 33.550763888888895$$

which is indeed larger than

$$\left\|\tilde{e}_2+\frac{1}{M}\sum_{i=1}^{M}e_{2,i}\right\|^2 = 3.365026291613992.$$

• in counter-example 2:

$$U = 675.8530370370372$$

which is indeed larger than

$$\left\|\tilde{e}_2+\frac{1}{M}\sum_{i=1}^{M}e_{2,i}\right\|^2 = 87.44127740238307.$$

• in counter-example 3:

$$U = 33.550763888888895$$

which is indeed larger than

$$\left\|\tilde{e}_2+\frac{1}{M}\sum_{i=1}^{M}e_{2,i}\right\|^2 = 3.3805310848189216.$$

Proof. [**Proof of Theorem 2**] We have

$$\begin{aligned}&\mathbf{E}\left[\left\|\tilde{e}_{t+1}+\tfrac{1}{M}\sum_{i=1}^{M}e_{t+1,i}\right\|^2\right]\\ &\leqslant 2\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right]+2\mathbf{E}\left[\left\|\tfrac{1}{M}\sum_{i=1}^{M}e_{t+1,i}\right\|^2\right] \qquad (10)\\ &\leqslant 2\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right]+\tfrac{2}{M}\sum_{i}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right],\end{aligned}$$

where the first inequality is by the fact that $(a+b)^2\leqslant 2a^2+2b^2, \forall a,b$, and the second inequality is by Lemma 1. We will separately bound the two terms of (10). Firstly, we consider $\frac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right]$. We have

$$\frac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right] = \frac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|\mathscr{C}\left(p_{t,i}\right)-p_{t,i}\right\|^2 \qquad (11)\right.$$

$$\leqslant \frac{1-\delta}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right] \tag{12}$$

$$= \frac{1-\delta}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|g_{t,i}+\frac{\eta_{t-1}}{\eta_t}e_{t,i}\right\|^2\right] \tag{13}$$

$$\begin{aligned}\leqslant\ & \tfrac{(1-\delta)(1+\gamma)}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|\frac{\eta_{t-1}}{\eta_t}e_{t,i}\right\|^2\right]\\ &+\tfrac{(1-\delta)(1+1/\gamma)}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|g_{t,i}\right\|^2\right]\end{aligned} \tag{14}$$

$$\begin{aligned}\leqslant\ & (1-\delta)(1+\gamma)\tfrac{\eta_{t-1}^2}{\eta_t^2}\left(\tfrac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t,i}\right\|^2\right]\right)\\ &+(1-\delta)(1+1/\gamma)G^2,\end{aligned} \tag{15}$$

where (11) and (13) is by the setting of $e_{t+1,i}$ and $p_{t,i}$ in Algorithm 1, (12) is by the definition of compressor $\mathscr{C}$, (14) is by Young inequality with any $\gamma > 0$, and (15) is by (4). Now, for all $t \geqslant 0$, applying Lemma 2 to the inequality (15) with

$$a_{t+1} = \tfrac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right],$$
$$\alpha_t = (1-\delta)(1+\gamma)\tfrac{\eta_{t-1}^2}{\eta_t^2},$$
$$\beta = (1-\delta)(1+1/\gamma)G^2,$$

we have

$$\frac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right] \leqslant \beta\left(1+\sum_{j=1}^{t}\prod_{i=j}^{t}\alpha_i\right)$$

Moreover, since

$$\begin{aligned}1+&\sum_{j=1}^{t}\prod_{i=j}^{t}\alpha_i\\ =\ & 1+\sum_{j=1}^{t}\prod_{i=j}^{t}(1-\delta)(1+\gamma)\frac{\eta_{i-1}^2}{\eta_i^2}\\ =\ & 1+\sum_{j=1}^{t}\frac{\eta_{j-1}^2}{\eta_t^2}[(1-\delta)(1+\gamma)]^{t-(j-1)}\\ =\ & \sum_{j=1}^{t+1}\frac{\eta_{j-1}^2}{\eta_t^2}[(1-\delta)(1+\gamma)]^{t-(j-1)}\\ =\ & \sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}[(1-\delta)(1+\gamma)]^{k},\end{aligned}$$

we obtain

$$\begin{aligned}&\tfrac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right]\\ \leqslant\ & (1-\delta)(1+1/\gamma)G^2\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}[(1-\delta)(1+\gamma)]^{k}.\end{aligned}$$

By choosing $\gamma = \frac{\delta}{2(1-\delta)}$, we have

$$(1-\delta)(1+1/\gamma) = \frac{(1-\delta)(2-\delta)}{\delta}$$

and

$$(1-\delta)(1+\gamma) = 1-\frac{\delta}{2}.$$

Therefore

$$\begin{aligned}&\tfrac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|e_{t+1,i}\right\|^2\right]\\ &\leqslant \tfrac{(2-\delta)(1-\delta)G^2}{\delta}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k,\end{aligned} \tag{16}$$

where $\alpha = 1-\frac{\delta}{2}$. Next, we consider the term $\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right]$ of (10). By the setting of $e_{t+1,i}$ and $p_{t,i}$ in Algorithm 1 and the definition of compressor $\mathscr{C}$, we have

$$\begin{aligned}&\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right]\\ =\ & \mathbf{E}\left[\left\|\mathscr{C}\left(\tilde{p}_t\right)-\tilde{p}_t\right\|^2\right]\\ \leqslant\ & (1-\delta)\mathbf{E}\left[\left\|\tilde{p}_t\right\|^2\right]\\ =\ & (1-\delta)\mathbf{E}\left[\left\|\tfrac{1}{M}\sum_{i=1}^{M}\mathscr{C}\left(p_{t,i}\right)+\frac{\eta_{t-1}}{\eta_t}\tilde{e}_t\right\|^2\right]\\ \leqslant\ & (1-\delta)(1+\gamma)\tfrac{\eta_{t-1}^2}{\eta_t^2}\mathbf{E}\left[\left\|\tilde{e}_t\right\|^2\right]\\ &+(1-\delta)(1+1/\gamma)\mathbf{E}\left[\left\|\tfrac{1}{M}\sum_{i=1}^{M}\mathscr{C}\left(p_{t,i}\right)\right\|^2\right],\end{aligned}$$

where the last inequality is by Young inequality for any $\gamma > 0$. Looking at $\mathbf{E}\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\mathscr{C}\left(p_{t,i}\right)\right\|^2\right]$, we have

$$\begin{aligned}&\mathbf{E}\left[\left\|\tfrac{1}{M}\sum_{i=1}^{M}\mathscr{C}\left(p_{t,i}\right)\right\|^2\right]\\ &\leqslant \tfrac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|\mathscr{C}\left(p_{t,i}\right)\right\|^2\right]\end{aligned} \tag{17}$$

$$\leqslant \frac{1}{M}\sum_{i=1}^{M}\left(2\mathbf{E}\left[\left\|\mathscr{C}\left(p_{t,i}\right)-p_{t,i}\right\|^2\right]+2\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right]\right) \tag{18}$$

$$\begin{aligned}&\leqslant \tfrac{1}{M}\sum_{i=1}^{M}\left(2(1-\delta)\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right]+2\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right]\right)\\ &= 2(2-\delta)\tfrac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right],\end{aligned} \tag{19}$$

where (17) is by Lemma 1, (18) is by the fact that $(a+b)^2 \leqslant 2a^2 + 2b^2, \forall a,b$, (19) is by the definition of compressor $\mathscr{C}$. Therefore Moreover, (12) and (16) yield. Therefore

$$\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right] \leqslant (1-\delta)(1+\gamma)\frac{\eta_{t-1}^2}{\eta_t^2}\mathbf{E}\left[\left\|\tilde{e}_t\right\|^2\right] + 2(2-\delta)(1-\delta)(1+1/\gamma)\frac{1}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right].$$

Moreover, (12) and (16) yield

$$\frac{1-\delta}{M}\sum_{i=1}^{M}\mathbf{E}\left[\left\|p_{t,i}\right\|^2\right] \leqslant \frac{(2-\delta)(1-\delta)G^2}{\delta}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k.$$

Therefore

$$\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right] \leqslant (1-\delta)(1+\gamma)\frac{\eta_{t-1}^2}{\eta_t^2}\mathbf{E}\left[\left\|\tilde{e}_t\right\|^2\right] + \frac{2(2-\delta)^2(1-\delta)(1+1/\gamma)G^2}{\delta}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k.$$

With $\gamma = \frac{\delta}{2(1-\delta)}$, since $(1-\delta)(1+1/\gamma) = \frac{(1-\delta)(2-\delta)}{\delta}$ and $(1-\delta)(1+\gamma) = 1-\frac{\delta}{2} = \alpha$, we have

$$\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right] \leqslant \alpha\frac{\eta_{t-1}^2}{\eta_t^2}\mathbf{E}\left[\left\|\tilde{e}_t\right\|^2\right] + \frac{2(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k.$$

By applying Lemma 2 with

$$a_t = \mathbf{E}\left[\left\|\tilde{e}_i\right\|^2\right],$$
$$\alpha_t = \alpha\frac{\eta_{t-1}^2}{\eta_t^2},$$
$$\beta_t = \frac{2(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k,$$

we obtain

$$\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right] \leqslant \beta_t + \sum_{j=1}^{t}\left(\prod_{i=j}^{t}\alpha_i\beta_{j-1}\right).$$

Since

$$\prod_{i=j}^{t}\alpha_i\beta_{j-1} = \prod_{i=j}^{t}\alpha\frac{\eta_{i-1}^2}{\eta_i^2}\beta_{j-1} = \alpha^{t-(j-1)}\frac{\eta_{j-1}^2}{\eta_t^2}\beta_{j-1},$$

we have

$$\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right] \leqslant \beta_t + \sum_{j=1}^{t}\alpha^{t-(j-1)}\frac{\eta_{j-1}^2}{\eta_t^2}\beta_{j-1} = \sum_{j=1}^{t+1}\alpha^{t-(j-1)}\frac{\eta_{j-1}^2}{\eta_t^2}\beta_{j-1}.$$

Therefore,

$$\begin{aligned}
&\mathbf{E}\left[\left\|\tilde{e}_{t+1}\right\|^2\right] \\
&\leqslant \sum_{j=1}^{t+1}\alpha^{t-(j-1)}\frac{\eta_{j-1}^2}{\eta_t^2}\frac{2(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{k=0}^{j-1}\frac{\eta_{j-1-k}^2}{\eta_{j-1}^2}\alpha^k \\
&= \frac{2(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{j=1}^{t+1}\alpha^{t-(j-1)}\sum_{k=0}^{j-1}\frac{\eta_{j-1-k}^2}{\eta_t^2}\alpha^k \\
&= \frac{2(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{j=0}^{t}\alpha^{t-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_t^2}\alpha^k.
\end{aligned} \tag{20}$$

Substituting (16) and (20) to (10), we obtain

$$\begin{aligned}
&\mathbf{E}\left[\left\|\tilde{e}_{t+1} + \frac{1}{M}\sum_{i=1}^{M}e_{t+1,i}\right\|^2\right] \\
&\leqslant \frac{2(2-\delta)(1-\delta)G^2}{\delta}\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k + \frac{4(1-\delta)(2-\delta)^3G^2}{\delta^2}\sum_{j=0}^{t}\alpha^{t-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_t^2}\alpha^k,
\end{aligned}$$

as claimed in Theorem 2. □

As a sanity check, Theorem 2 matches the results given in [1] when the learning rate is nondecreasing. As a result, Theorems 1 and A agree when the learning rate is nondecreasing.

Corollary 1. *(Sanity check of Theorem 2, cf. Lemma 6 of [1] with $\mu = 0$) In Theorem 2, if $\{\eta_t\}$ is a nondecreasing sequence such that $\eta_t > 0, \forall t \geqslant 0$, then*

$$\mathbf{E}\left[\left\|\tilde{e}_{t+1} + \frac{1}{M}\sum_{i=1}^{M}e_{t+1,i}\right\|^2\right] \leqslant \frac{8(1-\delta)G^2}{\delta^2}\left(1+\frac{16}{\delta^2}\right).$$

Proof. Since $\{\eta_t\}$ is nondecreasing, we have $\frac{\eta_{t-k}^2}{\eta_t^2} \leqslant 1$. Moreover, since $\alpha = 1-\frac{\delta}{2} \in (0,1)$, we obtain

$$\sum_{k=0}^{t}\frac{\eta_{t-k}^2}{\eta_t^2}\alpha^k \leqslant \sum_{k=0}^{t}\alpha^k \leqslant \frac{2}{\delta} \tag{21}$$

and

$$\sum_{j=0}^{t}\alpha^{t-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_t^2}\alpha^k \leqslant \sum_{j=0}^{t}\alpha^{t-j}\sum_{k=0}^{j}\alpha^k \leqslant \sum_{j=0}^{t}\alpha^{t-j}\left(\frac{2}{\delta}\right) \leqslant \frac{4}{\delta^2}.$$

Replacing (21) and (22) to Theorem 2, we have the result stated in Corollary 1. □

5. CORRECTING THE CONVERGENCE THEOREM OF ZHENG *ET AL*. [1]

Because the error bound plays a crucial role in the proof of the convergence theorem of dis-EF-SGD, fixing [1, Lemma 2] as in Theorem 2 leads to the consequence that the convergence theorem need to be fixed as well.

Proof. **(Proof of Theorem 1)** Following [1], we consider the iteration

$$\tilde{x}_t = x_t - \eta_{t-1}\left(\tilde{e}_t + \frac{1}{M}\sum_{i=1}^{M}e_{t,i}\right),$$

where x_t, $\tilde{e}_t$ and $e_{t,i}$ are generated from Algorithm 1. Then, by [1, Lemma 1],

$$\tilde{x}_{t+1} = \tilde{x}_t - \eta_t\frac{1}{M}\sum_{i=1}^{M}g_{t,i}. \tag{22}$$

Since f is L-smooth, by (3), we have Moreover, we have

$$\begin{aligned}\mathbf{E}_t\left[f(\tilde{x}_{t+1})\right] &\leqslant f(\tilde{x}_t) + \left\langle\nabla f(\tilde{x}_t), \mathbf{E}_t\left[\tilde{x}_{t+1}-\tilde{x}_t\right]\right\rangle + \frac{L}{2}\mathbf{E}_t\left[\left\|\tilde{x}_{t+1}-\tilde{x}_t\right\|^2\right]\\ &= f(\tilde{x}_t) - \eta_t\left\langle\nabla f(\tilde{x}_t), \mathbf{E}_t\left[\frac{1}{M}\sum_{i=1}^{M}g_{t,i}\right]\right\rangle + \frac{L\eta_t^2}{2}\mathbf{E}_f\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{t,i}\right\|^2\right],\end{aligned} \tag{23}$$

Moreover, we have

$$\mathbf{E}_t\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{t,i}\right\|^2\right] = \left\|\nabla f(x_t)\right\|^2 + \mathbf{E}_t\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{t,i}-\nabla f(x_t)\right\|^2\right], \tag{24}$$

which follows from the fact that $\mathbf{E}\left[\|X-\mathbf{E}[X]\|^2\right] = \mathbf{E}\left[\|X\|^2\right] - \|\mathbf{E}[X]\|^2$. Substituting (24) to (23), we obtain

$$\begin{aligned}\mathbf{E}_t\left[f(\tilde{x}_{t+1})\right] &\leqslant f(\tilde{x}_t) - \eta_t\left\langle\nabla f(\tilde{x}_t), \nabla f(x_t)\right\rangle + \frac{L\eta_t^2}{2}\left\|\nabla f(x_t)\right\|^2 + \frac{L\eta_t^2}{2}\mathbf{E}_t\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{t,i}-\nabla f(x_t)\right\|^2\right].\end{aligned} \tag{25}$$

Following [1], we assume that $\left\{g_{t,i}-\nabla f(x_t)\right\}_{1\leqslant i\leqslant M}$ are independent random vectors. Then the assumption $\mathbf{E}_t\left[g_{t,i}\right] = \nabla f(x_t)$ of Assumption 2 implies that $g_{t,i} - \nabla f(x_t)$ are random vectors with 0 means. Therefore

$$\mathbf{E}_t\left[\left\|\sum_{i=1}^{M}\left(g_{t,i}-\nabla f(x_t)\right)\right\|^2\right] = \sum_{i=1}^{M}\mathbf{E}_t\left[\left\|g_{t,i}-\nabla f(x_t)\right\|^2\right]$$

and hence we have

$$\begin{aligned}\mathbf{E}_t\left[\left\|\frac{1}{M}\sum_{i=1}^{M}g_{t,i}-\nabla f(x_t)\right\|^2\right] &= \mathbf{E}_t\left[\left\|\frac{1}{M}\sum_{i=1}^{M}\left(g_{t,i}-\nabla f(x_t)\right)\right\|^2\right]\\ &= \frac{1}{M^2}\mathbf{E}_t\left[\left\|\sum_{i=1}^{M}\left(g_{t,i}-\nabla f(x_t)\right)\right\|^2\right]\\ &= \frac{1}{M^2}\sum_{i=1}^{M}\mathbf{E}_t\left[\left\|g_{t,i}-\nabla f(x_t)\right\|^2\right]\\ &\leqslant \frac{\sigma^2 M}{M^2} = \frac{\sigma^2}{M}.\end{aligned}$$

Substituting the above bound to (25) gives us

$$\mathbf{E}_t\left[f(\tilde{x}_{t+1})\right] \leqslant f(\tilde{x}_t) - \eta_t\left\langle\nabla f(\tilde{x}_t), \nabla f(x_t)\right\rangle + \frac{L\eta_t^2}{2}\left\|\nabla f(x_t)\right\|^2 + \frac{L\eta_t^2\sigma^2}{2M}. \tag{26}$$

Moreover, we have

$$\begin{aligned}&-\eta_t\left\langle\nabla f(\tilde{x}_t), \nabla f(x_t)\right\rangle\\ &= \eta_t\left\langle\nabla f(x_t)-\nabla f(\tilde{x}_t), \nabla f(x_t)\right\rangle - \eta_t\left\langle\nabla f(x_t), \nabla f(x_t)\right\rangle\\ &= \eta_t\left\langle\nabla f(x_t)-\nabla f(\tilde{x}_t), \nabla f(x_t)\right\rangle - \eta_t\left\|\nabla f(x_t)\right\|\\ &\leqslant \frac{\eta_t\rho}{2}\left\|\nabla f(x_t)\right\|^2 + \frac{\eta_t}{2\rho}\left\|\nabla f(x_t)-\nabla f(\tilde{x}_t)\right\|^2\\ &\quad -\eta_t\left\|\nabla f(x_t)\right\|^2\end{aligned} \tag{27}$$

$$\leqslant -\eta_t\left(1-\frac{\rho}{2}\right)\left\|\nabla f(x_t)\right\|^2+\frac{\eta_t L^2}{2\rho}\left\|x_t-\tilde{x}_t\right\|^2, \tag{28}$$

where (27) is by the fact that $\langle a,b\rangle \leq (\rho/2)\,||a||\,2+(\rho^{-1}/2)||b||^2$ for all a,b and real number $\rho>0$, and (28) is by Assumption 1. Replacing (28) to (26) gives us

$$\mathbf{E}_t\left[f(\tilde{x}_{t+1})\right] \leqslant f(\tilde{x}_l)-\eta_t\left(1-\tfrac{L\eta_t+\rho}{2}\right)\left\|\nabla f(x_l)\right\|^2 +\tfrac{\eta_t L^2}{2\rho}\left\|x_t-\tilde{x}_t\right\|^2+\tfrac{L\eta_t^2\sigma^2}{2M}.$$

Taking $\rho=\frac{1}{2}$, we have

$$\mathbf{E}_t\left[f(\tilde{x}_{t+1})\right] \leqslant f(\tilde{x}_t)-\eta_t\left(\tfrac{3}{4}-\tfrac{L\eta_t}{2}\right)\left\|\nabla f(x_t)\right\|^2 +\eta_t L^2\left\|x_t-\tilde{x}_t\right\|^2+\tfrac{L\eta_t^2\sigma^2}{2M}. \tag{29}$$

Since $x_t-\tilde{x}_t=\eta_{t-1}\left(\tilde{e}_t+\frac{1}{M}\sum_{i=1}^{M}e_{t,i}\right)$ by (22), after rearranging the terms and taking total expectation, we obtain

$$\begin{aligned}\eta_t\left(\tfrac{3}{4}-\tfrac{L\eta_t}{2}\right)\mathbf{E}\left[\left\|\nabla f(x_t)\right\|^2\right] &\leqslant \mathbf{E}\left[f(\tilde{x}_t)-f(\tilde{x}_{t+1})\right]+\tfrac{L\eta_t^2\sigma^2}{2M}\\ &\quad+\eta_t\eta_{t-1}^2L^2\mathbf{E}\left[\left\|\tilde{e}_t+\tfrac{1}{M}\sum_{i=1}^{M}e_{t,i}\right\|^2\right].\end{aligned}$$

Applying Theorem 2 gives us

$$\eta_t\left(\tfrac{3}{4}-\tfrac{L\eta_t}{2}\right)\mathbf{E}\left[\left\|\nabla f(x_t)\right\|^2\right] \tag{30}$$

$$\begin{aligned}&\leqslant \mathbf{E}\left[f(\tilde{x}_t)-f(\tilde{x}_{t+1})\right]+\tfrac{L\eta_t^2\sigma^2}{2M}\\ &+\eta_t\eta_{t-1}^2\tfrac{2(2-\delta)(1-\delta)G^2L^2}{\delta}\sum_{k=0}^{t-1}\frac{\eta_{t-1-k}^2}{\eta_{t-1}^2}\alpha^k\\ &+\eta_t\eta_{t-1}^2\tfrac{4(1-\delta)(2-\delta)^3G^2L^2}{\delta^2}\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_{t-1}^2}\alpha^k,\end{aligned}$$

where $\alpha=1-\frac{\delta}{2}$. Since $\eta_t<\frac{3}{2L},\forall t$, we have

$$\sum_{k=0}^{T-1}\frac{\eta_k}{4}\left(3-2L\eta_k\right)>0.$$

Taking summation and dividing by $\sum_{k=0}^{T-1}\frac{\eta_k}{4}(3-2L\eta_k)$, (30) yields

$$\begin{aligned}&\sum_{t=0}^{T-1}\frac{\eta_t\left(3-2L\eta_t\right)\mathbf{E}\left[\left\|\nabla f(x_t)\right\|^2\right]}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\\ &\leqslant\frac{4(f(x_0)-f^*)}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\\ &+\tfrac{2L\sigma^2}{M}\sum_{t=0}^{T-1}\frac{\eta_t^2}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\\ &+\frac{8(2-\delta)(1-\delta)G^2L^2}{\delta\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\sum_{t=0}^{T-1}\eta_t\eta_{t-1}^2\sum_{k=0}^{t-1}\frac{\eta_{t-1-k}^2}{\eta_{t-1}^2}\alpha^k\\ &+\frac{16(1-\delta)(2-\delta)^3G^2L^2}{\delta^2\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\times\\ &\sum_{t=0}^{T-1}\eta_t\eta_{t-1}^2\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_{t-1}^2}\alpha^k.\end{aligned}$$

Following Zheng *et al.* [1], let $o\in\{0,\ldots,T-1\}$ be an index such that

$$\Pr(o=k)=\frac{\eta_k\left(3-2L\eta_k\right)}{\sum_{t-0}^{T-1}\eta_t\left(3-2L\eta_t\right)}$$

Then

$$\begin{aligned}&\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right]\\ &=\frac{1}{\sum_{k=0}^{T-1}\eta_k\left(3-2L\eta_k\right)}\sum_{t=0}^{T-1}\eta_t\left(3-2L\eta_t\right)\mathbf{E}\left[\left\|\nabla f(x_t)\right\|^2\right]\end{aligned}$$

and we obtain the result stated in Theorem 1 □

The following corollary establishes the convergence rate $O\left(\frac{1}{\sqrt{MT}}\right)$ of Algorithm 1 when the learning rate is decreasing.

Corollary 2. *(Convergence rate with decreasing learning rate) Under the assumptions of Theorem 1, if $\{\eta_t\}$ is a decreasing sequence such that*

$$\eta_t=\frac{1}{\frac{((t+1)T)^{1/4}}{\sqrt{M}}+T^{1/3}}$$

with sufficiently large T. Then

$$\begin{aligned}&\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right]\\ &\leqslant 2\left(\tfrac{1}{\sqrt{MT}}+\tfrac{1}{T^{2/3}}\right)\left[f(x_0)-f^*+L\sigma^2\right.\\ &\left.+\tfrac{4(1-\delta)(2-\delta)G^2L^2}{\delta^2}\left(1+\tfrac{4}{\delta^2}\right)\right],\end{aligned}$$

which yields $\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right]\leqslant O\left(\frac{1}{\sqrt{MT}}\right)$.

Proof. Following [1], assume that $T \geqslant 16L^4M^2$, we have

$$\eta_t = \frac{1}{\frac{((t+1)T)^{1/4}}{\sqrt{M}} + T^{1/3}} \leqslant \frac{\sqrt{M}}{((t+1)T)^{1/4}} \leqslant \frac{\sqrt{M}}{(t+1)^{1/4}(16L^4M^2)^{1/4}} = \frac{1}{(t+1)^{1/4}2L} \leqslant \frac{1}{2L}.$$

Therefore $\eta_t < \frac{3}{2L} \forall t \geqslant 0$, which satisfies the assumption of Theorem 1 on $\{\eta_t\}$. Recall that by Theorem 1, we have

$$\begin{aligned}
&\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right] \\
&\leqslant \frac{4(f(x_0)-f^*)}{\sum_{k=0}^{T-1}\eta_k(3-2L\eta_k)} + \frac{2L\sigma^2}{M\sum_{k=0}^{T-1}\eta_k(3-2L\eta_k)}\sum_{t=0}^{T-1}\eta_t^2 \\
&\quad + \frac{8(1-\delta)(2-\delta)G^2L^2}{\delta\sum_{k=0}^{T-1}\eta_k(3-2L\eta_k)}\sum_{t=0}^{T-1}\eta_t\eta_{t-1}^2\sum_{k=0}^{t-1}\frac{\eta_{t-1-k}^2}{\eta_{t-1}^2}\alpha^k \\
&\quad + \frac{16(1-\delta)(2-\delta)^3G^2L^2}{\delta^2\sum_{k=0}^{T-1}\eta_k(3-2L\eta_k)}\times \\
&\qquad \sum_{t=0}^{T-1}\eta_t\eta_{t-1}^2\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_{t-1}^2}\alpha^k,
\end{aligned} \tag{31}$$

where $\alpha = 1 - \frac{\delta}{2}$ and $o \in \{0, \dots, T-1\}$ is an index such that

$$\Pr(o = k) = \frac{\eta_k(3-2L\eta_k)}{\sum_{k=0}^{T-1}\eta_t(3-2L\eta_t)}, \forall k = 0, \dots, T-1.$$

Since

$$3 - 2L\eta_t \geqslant 3 - \frac{1}{(t+1)^{1/4}} \geqslant 2,$$

we have

$$\frac{1}{\sum_{k=0}^{T-1}\eta_t(32L\eta_t)} \leqslant \frac{1}{2\sum_{k=0}^{T-1}\eta_t}. \tag{32}$$

Moreover, we have

$$\eta_t\eta_{t-1}^2\sum_{k=0}^{t-1}\frac{\eta_{t-1-k}^2}{\eta_{t-1}^2}\alpha^k = \sum_{k=0}^{t-1}\eta_t\eta_{t-1-k}^2\alpha^k \tag{33}$$

and

$$\begin{aligned}
&\eta_t\eta_{t-1}^2\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\frac{\eta_{j-k}^2}{\eta_{t-1}^2}\alpha^k \\
&= \sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\eta_t\eta_{j-k}^2\alpha^k.
\end{aligned} \tag{34}$$

Substituting (32), (33), and (34) to (31) gives us

$$\begin{aligned}
&\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right] \\
&\leqslant \frac{2(f(x_0)-f^*)}{\sum_{t=0}^{T-1}\eta_t} + \frac{L\sigma^2}{M\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\eta_t^2 \\
&\quad + \frac{4(1-\delta)(2-\delta)G^2L^2}{\delta\sum_{t=0}^{T-1}\eta_t}\sum_{t=0}^{T-1}\sum_{k=0}^{t-1}\eta_t\eta_{t-1-k}^2\alpha^k \\
&\quad + \frac{8(1-\delta)(2-\delta)^3G^2L^2}{\delta^2\sum_{t=0}^{T-1}\eta_t}\times \\
&\qquad \sum_{t=0}^{T-1}\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\eta_t\eta_{j-k}^2\alpha^k,
\end{aligned} \tag{35}$$

Because

$$\begin{aligned}
\eta_t\eta_{t-1-k}^2 &\leqslant \eta_{t-1-k}^3 \\
&= \frac{1}{\left(\frac{((t-k)T)^{1/4}}{\sqrt{M}} + T^{1/3}\right)^3} \\
&\leqslant \frac{1}{(T^{1/3})^3} = \frac{1}{T}
\end{aligned} \tag{36}$$

and

$$\sum_{k=0}^{t-1}\alpha^k \leqslant \sum_{k\geqslant 0}\alpha^k = \frac{1}{1-\alpha} = \frac{2}{\delta}, \tag{37}$$

we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1}\sum_{k=0}^{t-1}\eta_t\eta_{t-1-k}^2\alpha^k &\leqslant \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t-1}\alpha^k \\
&\leqslant \frac{1}{T}\sum_{t=0}^{T-1}\frac{2}{\delta} = \frac{2}{\delta}.
\end{aligned}$$

By the same reason as in (36) and (37), we have $\eta_t\eta_{j-k}^2 \leqslant \frac{1}{T}$, $\sum_{j=0}^{t-1}\alpha^{t-1-j} \leqslant \frac{2}{\delta}$, and $\sum_{k=0}^{j}\alpha^k \leqslant \frac{2}{\delta}$. Therefore

$$\begin{aligned}
&\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\eta_t\eta_{j-k}^2\alpha^k \\
&\leqslant \frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}\alpha^{t-1-j}\sum_{k=0}^{j}\alpha^k \\
&\leqslant \frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}\alpha^{t-1-j}\left(\frac{2}{\delta}\right) \\
&\leqslant \frac{1}{T}\sum_{t=0}^{T-1}\frac{4}{\delta^2} = \frac{4}{\delta^2}.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}\sum_{t=0}^{T-1}\eta_t^2 &= \sum_{t=0}^{T-1}\frac{1}{\left(\frac{((t+1)T)^{1/4}}{\sqrt{M}}+T^{1/3}\right)^2}\\ &\leqslant \sum_{t=0}^{T-1}\frac{1}{\left(\frac{((t+1)T)^{1/4}}{\sqrt{M}}\right)^2}\\ &= \sum_{t=0}^{T-1}\frac{M}{[(t+1)T]^{1/2}}\\ &= \frac{M}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\\ &\leqslant 2M,\end{aligned}$$

where the last inequality is by the fact that $\sum_{t=1}^{T}\frac{1}{\sqrt{t}} \leqslant 2\sqrt{T}$.

Therefore

$$\begin{aligned}\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right] \leqslant \frac{2\left(f(x_0)-f^\star\right)}{\sum_{t=0}^{T-1}\eta_t} + \frac{2L\sigma^2}{\sum_{t=0}^{T-1}\eta_t}\\ +\frac{8(1-\delta)(2-\delta)G^2L^2}{\delta^2\sum_{t=0}^{T-1}\eta_t}\\ +\frac{32(1-\delta)(2-\delta)^3G^2L^2}{\delta^4\sum_{t=0}^{T-1}\eta_t}.\end{aligned}$$

Furthermore, becauce

$$\begin{aligned}\sum_{t=0}^{T-1}\eta_t = \sum_{t=0}^{T-1}\frac{1}{\frac{((t+1)T)^{1/4}}{(\sqrt{M})}+T^{1/3}} \geqslant \sum_{t=0}^{T-1}\frac{1}{\frac{\sqrt{T}}{\sqrt{M}}+T^{1/3}}\\ = \frac{1}{\frac{1}{\sqrt{MT}}+T^{-2/3}},\end{aligned}$$

we obtain

$$\frac{1}{\sum_{t=0}^{T-1}\eta_t} \leqslant \frac{1}{\sqrt{MT}}+\frac{1}{T^{2/3}}.$$

Therefore

$$\begin{aligned}&\mathbf{E}\left[\left\|\nabla f(x_o)\right\|^2\right]\\ &\leqslant 2\left(\frac{1}{\sqrt{MT}}+\frac{1}{T^{2/3}}\right)\left[f(x_0)-f^\star+L\sigma^2\right.\\ &\quad\left.+\frac{4(1-\delta)(2-\delta)G^2L^2}{\delta^2}\left(1+\frac{4}{\delta^2}\right)\right]\end{aligned}$$

and hence Corollary 2 follows. □

6. CONCLUSION

We show that the convergence proof of dist-EF-SGD of Zheng *et al.* [1] is problematic when the sequence of learning rate is decreasing. We explicitly provide counter-examples with certain decreasing sequences of learning rate to show the issue in the proof of Zheng *et al.* [1]. We fix the issue by providing a new error bound and a new convergence theorem for the dist-EF-SGD algorithm, which helps recover its mathematical foundation.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Tran Thi Phuong established the research direction. Both authors contributed to the technical contents, and to the edition of the manuscript. Both authors read, revised, and approved the final manuscript.

ACKNOWLEDGMENTS

We are grateful to Shuai Zheng for his communication and verification. We also thank the anonymous reviewers for their careful comments. The work of Le Trieu Phong was supported in part by JST CREST under Grant JPMJCR19F6.

REFERENCES

[1] S. Zheng, Z. Huang, J.T. Kwok, Communication-efficient distributed blockwise momentum SGD with error-feedback, in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. 2019, pp. 11446–11456. https://arxiv.org/abs/1905.10936

[2] J. Bernstein, J. Zhao, K. Azizzade-nesheli, A. Anandkumar, signSGD with majority vote is communication efficient and fault tolerant, in 7th International Conference on Learning Representations (ICLR 2019), New Orleans, LA, USA. 2019.

[3] D. Basu, D. Data, C. Karakus, S.N. Diggavi, Qsparse-local-SGD: distributed SGD with quantization, sparsification and local computations, in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. 2019, pp. 14668–14679.

[4] T. Vogels, S.P. Karimireddy, M. Jaggi, PowerSGD: practical low-rank gradient compression for distributed optimization, in Advances in Neural Information Processing Systems, Curran Associates, Inc., Vancouver, Canada. 2019, pp. 14236–14245.

[5] S.U. Stich, J.-B. Cordonnier, M. Jaggi, Sparsified SGD with memory, in Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, Montréal, Canada. 2018, pp. 4452–4463.

[6] H. Tang, C. Yu, X. Lian, T. Zhang, J. Liu, DoubleSqueeze: parallel stochastic gradient descent with double-pass error-compensated compression, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, vol. 97 of Proceedings of Machine Learning Research (PMLR), PMLR, Long Beach, California, USA. 2019, pp. 6155–6165.

[7] X. Liu, Y. Li, J. Tang, M. Yan, A double residual compression algorithm for efficient distributed learning, in Proceedings of Machine Learning Research (PMLR), Palermo, Italy. 2020, pp. 133–143.

[8] T.T. Phuong, L.T. Phong, Distributed SGD with flexible gradient compression, IEEE Access. 8 (2020), 64707–64717.

[9] S.P. Karimireddy, Q. Rebjock, S.U. Stich, M. Jaggi, Error feedback fixes signSGD and other gradient compression schemes, in Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, California, USA. 2019, pp. 3252–3261. https://arxiv.org/abs/1901.09847

[10] Y. Nesterov, Introductory Lectures on Convex Optimization, vol. 87, Springer Science and Business Media, Boston, MA, USA, 2004.