Detached Error Feedback for Distributed SGD with Random Sparsification

An Xu [1] **Heng Huang** [1]

Abstract

The communication bottleneck has been a critical problem in large-scale distributed deep learning. In this work, we study distributed SGD with random block-wise sparsification as the gradient compressor, which is ring-allreduce compatible and highly computation-efficient but leads to inferior performance. To tackle this important issue, we improve the communication-efficient distributed SGD from a novel aspect, that is, the trade-off between the variance and second moment of the gradient. With this motivation, we propose a new detached error feedback (DEF) algorithm, which shows better convergence bound than error feedback for non-convex problems. We also propose DEF-A to accelerate the generalization of DEF at the early stages of the training, which shows better generalization bounds than DEF. Furthermore, we establish the connection between communication-efficient distributed SGD and SGD with iterate averaging (SGD-IA) for the first time. Extensive deep learning experiments show significant empirical improvement of the proposed methods under various settings.

1. Introduction

Deep learning models are hard to train due to the heavy computation complexity and long training iterations. Therefore, distributed deep learning with multiple workers (GPUs) has become a prevalent practice to parallelize and accelerate the training for large-scale tasks, where the model and dataset sizes continue to grow nowadays (Simonyan & Zisserman, 2014; He et al., 2016; Deng et al., 2009).

Nevertheless, synchronous distributed training have difficulty in scaling up the number of workers for large deep learning models, as the gradient in each worker to be communicated per iteration is of the same dimension as the model size. It is also known as the communication bottleneck. Besides, it incurs imbalanced communication traffic in the parameter-server (Li et al., 2014a;b; 2013) architecture, where the server suffers from much larger communication burden than workers. To address the communication bottleneck issue, there have been numerous lines of works including asynchronous execution (Dean et al., 2012), gradient compression (Bernstein et al., 2018a;b; Wen et al., 2017; Alistarh et al., 2017; Aji & Heafield, 2017; Alistarh et al., 2018; Strom, 2015; Lin et al., 2018; Gao et al., 2021), communication scheduling (George & Gurram, 2020), infrequent communication (Stich, 2018), delayed gradient (Li et al., 2018; Zhu et al., 2021), decentralized training (Lian et al., 2017; Koloskova et al., 2019; Tang et al., 2018; Assran et al., 2019; Koloskova et al., 2019), model parallelism (Huang et al., 2019; Xu et al., 2020), etc.

In this work, we focus on synchronous distributed SGD with gradient compression, or more specifically, random block-wise gradient sparsification (RBGS) (Vogels et al., 2019; Xie et al., 2020). The most popular gradient sparsifier is probably the Top-$\mathcal{K}$ gradient sparsification (Alistarh et al., 2018; Lin et al., 2018), where each worker selects the largest $\mathcal{K}$ gradient components according to the absolute value as the sparsified gradient. However, Top-$\mathcal{K}$ has several drawbacks: 1) it requires extra communication overheads to communicate the gradient indices, 2) it is applied in parameter-server architecture but not ring-allreduce compatible, and most of all, 3) its computation overheads $\mathcal{O}(\mathcal{K}\log_2 d)$ for model $\theta \in \mathbb{R}^d$ may even outweigh its communication benefits (Song et al., 2021; Xie et al., 2020; Sahu et al., 2021) as it is efficient only for a small $\mathcal{K}$ for optimized implementations on GPU (Shanbhag et al., 2018). While in RBGS, we randomly sample a block of gradient as the sparsified gradient for communication among workers. To ensure the consistency of the sampling process, each worker will be pre-assigned the same random seed. In comparison to Top-$\mathcal{K}$, RBGS is highly computation-efficient ($\mathcal{O}(1)$) as we only need to uniformly and randomly sample one starting index of the gradient block. RBGS is also ring-allreduce compatible. However, RBGS results in inferior model performance in that its sparsified gradient usually does not include as

[1]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15213, USA. This work was partially supported by NSF IIS 1845666, 1852606, 1838627, 1837956, 1956002, IIA 2040588. Correspondence to: Heng Huang <heng.huang@pitt.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

many significant gradient components as Top-$\mathcal{K}$, leading to large compression error.

To address this important problem, we propose a novel detached error feedback method (DEF), while the vanilla error feedback (EF) method (Karimireddy et al., 2019; Zheng et al., 2019) fails to address it. We summarize our major contributions as follows.

- Our proposed DEF method is motivated by a novel insight that a trade-off between the gradient variance and second moment can improve the convergence bound related to compression error.
- We propose DEF-A to accelerate the generalization during the training with support from corresponding generalization analysis. It potentially demystifies why compression helps to improve the performance in some prior works (Avdiukhin & Yaroslavtsev, 2021; Zhao et al., 2020; Bernstein et al., 2018a;b).
- We find that SGD with iterate averaging (SGD-IA) (Polyak, 1990; Ruppert, 1988; Neu & Rosasco, 2018; Wu et al., 2020) can be viewed as a special case of communication-efficient distributed SGD for the first time. Consequently, our generalization analysis of DEF-A extends to SGD-IA, providing potential theoretical explanations for some other applications incorporating SGD-IA (He et al., 2020; Izmailov et al., 2018; Huang et al., 2017).
- Extensive deep image classification experiments on CIFAR-10/100 and ImageNet show significant improvements of DEF(-A) over existing works with RBGS.

2. Related Works

To begin with, suppose the training dataset $\mathcal{S} = \{\xi_n\}_{n=1}^N$ and we have the training objective function

$$F_{\mathcal{S}}(\theta) = \frac{1}{N}\sum_{n=1}^{N} f(\theta;\xi_n) = \mathbb{E}_{\xi\in\mathcal{S}} f(\theta;\xi) \tag{1}$$

to minimize, where $\theta \in \mathbb{R}^d$ denotes the model and f is the loss function. From now on, we will omit the subscript in $\mathbb{E}$ if the context is clear. For distributed SGD at iteration t, each worker k randomly selects one data sample $\xi_{k,t} \in \mathcal{S}$ and computes the stochastic gradient $g_{k,t} = \nabla f(\theta_t;\xi_{k,t})$. Then all the workers communication to get the average gradient $g_t = \frac{1}{K}\sum_{k=1}^K g_{k,t}$, where K is the total number of workers, and update the model via

$$\theta_{t+1} = \theta_t - \eta g_t\,, \tag{2}$$

where η is the learning rate.

Compression. Gradient compression includes quantization (Bernstein et al., 2018a;b; Wen et al., 2017; Alistarh et al., 2017), which reduces the 32-bit gradient component to as low as 1 bit (compression ratio ≤ 32), and sparsification (Aji & Heafield, 2017; Alistarh et al., 2018; Strom, 2015), which reduces the number of gradient components for communication. Let the compression function be $\mathcal{C}$, then the workers will communicate $\mathcal{C}(g_{k,t})$ instead of $g_{k,t}$. In general, sparsification achieves flexible and higher compression ratio than quantization. Besides Top-$\mathcal{K}$, random-$\mathcal{K}$ (Elibol et al., 2019; Stich et al., 2018) randomly selects $\mathcal{K}$ gradient components as the sparsified gradient. Dutta et al. (2020) selects gradient components larger than a threshold and is a variable-dimension compressor. Wangni et al. (2018); Song et al. (2021) propose to select each gradient component with a probability to keep the sparsified gradient unbiased. In this work, we consider RBGS (Vogels et al., 2019; Xie et al., 2020), which is most easy to implement, highly computation-efficient, but challenging to retain the model performance. Moreover, it is ring-allreduce compatible for SOTA GPU communication backend library (e.g., NCCL), i.e.,

$$\mathcal{C}(\Delta_1) + \mathcal{C}(\Delta_2) = \mathcal{C}(\Delta_1 + \Delta_2)\,. \tag{3}$$

Error Feedback. Error feedback (EF) (Karimireddy et al., 2019; Tang et al., 2019) method maintains local compression error $e_{k,t}$ at worker k, adds it to the current gradient before compression, and communicates to average $\mathcal{C}(\eta g_{k,t} + e_{k,t})$. The error is updated via

$$e_{k,t+1} = \eta g_{k,t} + e_{k,t} - \mathcal{C}(g_{k,t} + e_{k,t})\,. \tag{4}$$

Zheng et al. (2019) extends EF to momentum SGD (Polyak, 1964). EF works well for Top-$\mathcal{K}$ sparsifier but poorly for RBGS. Xie et al. (2020) proposes PSync to immediately apply local error to each worker's model for RBGS. However, we will show that PSync works better for Wide ResNet (Zagoruyko & Komodakis, 2016) but has scalability issue for other common model architectures. SAEF (Xu et al., 2021) proposes to apply the local error before computing gradient in the next iteration to accelerate the generalization during training. Other EF variants includes EF21 (Richtárik et al., 2021; Fatkhullin et al., 2021) which compresses the gradient difference (Mishchenko et al., 2019) but is evaluated only on logistic regression problems, acceleration for EF (Qian et al., 2021; Li et al., 2020), EF for variance reduction (Tang et al., 2021), etc.

Generalization Analysis. The generalization analysis of this work incorporates the uniform stability (Bousquet & Elisseeff, 2002; Hardt et al., 2016) approach, focusing on the inherent stability property of the learning algorithm. Bousquet & Elisseeff (2002) analyzes bagging methods. It is later used to analyze the generalization property of SGD (Hardt et al., 2016) and its momentum variants (Yan et al., 2018). Kuzborskij & Lampert (2018) establishes a data-dependent

Algorithm 1 Detached Error Feedback (DEF(-A)).

1: **Input:** training dataset $\mathcal{S}$, number of iterations T, number of workers K, learning rate η, ring-allreduce compressor $\mathcal{C}$, coefficient $\lambda \in [0, 1]$.
2: **Initialize:** model $x_0 = y_0$, local compression error $e_{k,0} = 0$, worker $k \in [K]$.
3: **for** $t = 0, 1, \cdots, T-1$ **do**
4: **for** worker $k \in [k]$ in parallel **do**
5: Randomly sample data $\xi_{k,t}$ from $\mathcal{S}$.
6: Compute $g_{k,t} = \nabla f(x_t - \lambda e_{k,t}; \xi_{k,t})$. *// detach*
7: $p_{k,t} = \eta g_{k,t} + e_{k,t}$. *// error feedback*
8: $e_{k,t+1} = p_{k,t} - \mathcal{C}(p_{k,t})$.
9: Ring-allreduce: $\mathcal{C}(p_t) = \mathcal{C}(\frac{1}{K}\sum_{k=1}^K p_{k,t}) = \frac{1}{K}\sum_{k=1}^K \mathcal{C}(p_{k,t})$.
10: Update $x_{t+1} = x_t - \mathcal{C}(p_t)$.
11: **end for**
12: **end for**
13: **Output:** $y_T = x_T - e_T = x_T - \frac{1}{K}\sum_{k=1}^K e_{k,T}$ for DEF and x_T for DEF-A.

notion of the stability to stress the distribution-dependent risk of the initialization point and make the generalization bounds more optimistic. Zhou et al. (2021) analyzes the generalization of the Lookahead optimizer (Zhang et al., 2019) with uniform stability.

As there are numerous works combining various techniques (Basu et al., 2019), in this work, we focus on random blockwise gradient sparsification (RBGS).

3. Detached Error Feedback

In this section, we described our proposed DEF method (Algorithm 1) in detail. As RBGS is a very aggressive compressor, the algorithm is crucial for better performance.

3.1. Motivation

In EF variants (Karimireddy et al., 2019; Zheng et al., 2019; Xie et al., 2020) for practical large-scale distributed training of deep learning models, Assumptions 3.1 and 3.2 are needed to bound the norm of the stochastic gradient

$$\|\nabla f(\theta; \xi)\|_2 \le G = \sqrt{\sigma^2 + M^2}\,. \tag{5}$$

Then G bounds the compression error

$$\frac{1}{K}\sum_{k=1}^K \|e_{k,t}\|_2^2 = \mathcal{O}(\sigma^2 + M^2) \tag{6}$$

at iteration t. Though Assumption 3.2 often appears in related literature, it is usually regarded as a strong assumption (Richtárik et al., 2021) because M^2 could be much larger than σ^2. Hereby, we propose a novel insight that if some trade-off coefficient α can be introduced to transform the compression error bound to a similar interpolation form as

$$(\alpha\sigma)^2 + ((1-\alpha)M)^2 \overset{\alpha=\frac{M^2}{\sigma^2+M^2}}{\ge} \frac{\sigma^2 M^2}{\sigma^2 + M^2} \overset{M\gg\sigma}{=} \sigma^2\,, \tag{7}$$

then the bound $\mathcal{O}(\sigma^2 + M^2)$ can be reduced to $\mathcal{O}(\sigma^2)$ when $M \to \infty$, i.e., Assumption 3.2 does not hold.

Assumption 3.1. (Bounded Variance) $\forall \theta \in \mathbb{R}^d$, the variance of the stochastic gradient satisfies $\mathbb{E}_{\xi\in\mathcal{S}}\|\nabla f(\theta;\xi) - \nabla F_{\mathcal{S}}(\theta)\|_2^2 \le \sigma^2$.

Assumption 3.2. (Bounded Second Moment) $\forall \theta \in \mathbb{R}^d$, the second moment of the full gradient satisfies $\|\nabla F_{\mathcal{S}}(\theta)\|_2^2 \le M^2$.

3.2. Algorithm

Assumption 3.3. (Ring-allreduce Compressor) $\forall \Delta_1, \Delta_2 \in \mathbb{R}^d$, the compressor $\mathcal{C}$ satisfies $\mathcal{C}(\Delta_1) + \mathcal{C}(\Delta_2) = \mathcal{C}(\Delta_1 + \Delta_2)$.

Firstly, the ring-allreduce communication architecture requires that the compressor should satisfy Assumption 3.3 such that Algorithm 1 line 9 holds. RBGS satisfies this assumption.

Secondly, DEF returns $y_T = x_T - e_T$ by default because we have

$$y_{t+1} = y_t - \eta g_t = y_t - \frac{1}{K}\sum_{k=1}^K g_{k,t}\,. \tag{8}$$

In particular, when $K = 1$ (single worker) and $\lambda = 1$, $\{y_t\}$ is identical to the SGD solution path. We note that averaging $e_T = \frac{1}{K}\sum_{k=1}^K e_{k,T}$ only incurs a one-time communication cost after the training concludes.

Then, a major difference of DEF and EF is that we evaluate gradient at $x_t - \lambda e_{k,t}$, a point **detached** from the point x_t to evaluate gradient as in EF. This step does not incur any communication cost. From Eq. (8), our goal is to make sure that the point to evaluate gradient $g_{k,t}$ is as close to $y_t = x_t - e_t$ as possible. For EF, the distance is $\|x_t - y_t\|_2^2 = \|e_t\|_2^2 \le \frac{1}{K}\sum_{k=1}^K \|e_{k,t}\|_2^2$, while for DEF, the average distance to minimize regarding λ becomes

$$\frac{1}{K}\sum_{k=1}^K \|x_t - \lambda e_{k,t} - y_t\|_2^2 = \frac{1}{K}\sum_{k=1}^K \|e_t - \lambda e_{k,t}\|_2^2\,. \tag{9}$$

(1) When $\lambda = \lambda(k,t)$, it is obvious that $\lambda^*(k,t) = \frac{\langle e_t, e_{k,t}\rangle}{\|e_{k,t}\|_2^2}$, which is determined by the *projection* of e_t onto $e_{k,t}$. However, it is impractical to decide $\lambda^*(k,t)$ for worker k at iteration t as e_t is unknown ($e_t = \frac{1}{K}\sum_{k=1}^K e_{k,t}$ needs extra communication cost).

(2) When $\lambda = \lambda(t)$, we can derive $\lambda^*(t) = \frac{\|e_t\|_2^2}{\frac{1}{K}\sum_{k=1}^K \|e_{k,t}\|_2^2}$, which is still impractical due to unkown e_t.

(3) Therefore, we will regard λ as a tuned hyper-parameter, invariant regarding k and t. Then it becomes minimizing the sum of the errors $\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^K \|e_t - \lambda e_{k,t}\|_2^2$ which will appear in the convergence bound of DEF, similar to the suggestion in Sahu et al. (2021). Previously when λ is a function of t, it reduces to minimizing Eq. (9). In our CIFAR-10 VGG-16 experiments with $\lambda = 0.3$, we find that the new distance is $\times 1.7$ smaller than the distance in EF.

Relation to Motivation. Minimizing Eq. (9) is closely related to the motivation since

$$\|e_t - \lambda e_{k,t}\|_2^2 = \|(\underbrace{\eta g_{t-1} - \lambda\eta g_{k,t-1}}_{} + e_{t-1} - \lambda e_{k,t-1}) - \mathcal{C}(\underbrace{\eta g_{t-1} - \lambda\eta g_{k,t-1}}_{} + e_{t-1} - \lambda e_{k,t-1})\|_2^2 , \tag{10}$$

where $g_{t-1} - \lambda g_{k,t-1}$ is affected by the gradient variance and second moment trade-off via the choice of λ. For example, in extreme circumstances where $\sigma = 0$, in expectation, local errors on different workers are the same and $g_{t-1} - \lambda g_{k,t-1}$ is zero with $\lambda = 1$.

Momentum Variant. It is easy to extend DEF to momentum SGD variant. Let the momentum buffer on worker k be $m_{k,0} = 0$ and the momentum constant be μ. We only need to substitute Algorithm 1 line 7 with

$$m_{k,t+1} = \mu m_{k,t} + g_{k,t},\ p_{k,t} = \eta m_{k,t+1} + e_{k,t} . \tag{11}$$

DEF-A. Simply returning x_T can accelerate the generalization performance of DEF during training in that when $K = 1$, $\lambda = 1$ and $\mathcal{C}(\Delta) = \delta\Delta$ $(0 < \delta < 1)$, $\{y_t\}$ reduces to SGD and $\{x_t\}$ reduces to a special case of SGD-IA (Iterate Averaging, a combination of models in each iteration) (Wu et al., 2020):

$$x_t = \underbrace{(1-\delta)^t}_{P_0} y_0 + \sum_{t'=1}^{t} \underbrace{\delta(1-\delta)^{t-t'}}_{P_{t'}} y_{t'} , \tag{12}$$

where $P_0 + P_1 + \cdots + P_t = 1$. Note that for Polyak-Ruppert IA (Polyak, 1990), $P_0 = P_1 = \cdots = P_t = \frac{1}{t+1}$. While for geomeric Polyak-Ruppert IA (Neu & Rosasco, 2018), $P_{t'} = \frac{\beta^{t'}}{1+\beta+\cdots+\beta^t}$ where $0 < \beta < 1$ is some constant and $0 \le t' \le t$. However, this part is based on generalization analysis instead of convergence analysis as for DEF. Hence we leave the details of the general case in the next section.

4. Theoretical Analysis

In this section, we consider non-convex objective functions as our target is the deep learning model. All detailed proof can be found in the Appendix. Suppose that each ξ_n in the training dataset $\mathcal{S}$ is i.i.d drawn from an unknown data distribution $\mathcal{D}$ and $F_{\mathcal{D}}(\theta) = \mathbb{E}_{\xi\in\mathcal{D}} f(\theta;\xi)$. For generalization, we are interested in how the model $\theta_{\mathcal{A},\mathcal{S}}$, which is trained on $\mathcal{S}$ with a randomized algorithm $\mathcal{A}$, generalizes on $\mathcal{D}$ by measuring the well-known excess risk error ϵ.

$$\begin{aligned}\epsilon &= \mathbb{E}_{\mathcal{A},\mathcal{S}}[F_{\mathcal{D}}(\theta_{\mathcal{A},\mathcal{S}})] - \mathbb{E}_{\mathcal{A},\mathcal{S}}[F_{\mathcal{S}}(\theta^*_{\mathcal{S}})] \\ &= \underbrace{\mathbb{E}_{\mathcal{A},\mathcal{S}}[F_{\mathcal{S}}(\theta_{\mathcal{A},\mathcal{S}}) - F_{\mathcal{S}}(\theta^*_{\mathcal{S}})]}_{\text{optimization error } \epsilon_{\text{opt}}} \\ &\quad + \underbrace{\mathbb{E}_{\mathcal{A},\mathcal{S}}[F_{\mathcal{D}}(\theta_{\mathcal{A},\mathcal{S}}) - F_{\mathcal{S}}(\theta_{\mathcal{A},\mathcal{S}})]}_{\text{generalization error } \epsilon_{\text{gen}}}\end{aligned} \tag{13}$$

Assumption 4.1. (L-Lipschitz Smooth) $\forall\theta_1, \theta_2 \in \mathbb{R}^d$, the loss function satisfies

$$\|\nabla f(\theta_1;\xi) - \nabla f(\theta_2;\xi)\|_2 \le L\|\theta_1 - \theta_2\|_2 . \tag{14}$$

It also implies that

$$\|\nabla F(\theta_1) - \nabla F(\theta_2)\|_2 \le L\|\theta_1 - \theta_2\|_2 . \tag{15}$$

Assumption 4.2. (δ-approximate Compressor) $\forall\Delta \in \mathbb{R}^d$, the compressor $\mathcal{C}$ satisfies

$$\|\mathcal{C}(\Delta) - \Delta\|_2^2 \le (1-\delta)\|\Delta\|_2^2 , \tag{16}$$

where $0 < \delta < 1$ is related to the compression ratio.

This assumption is widely used in communication-efficient distributed SGD (Karimireddy et al., 2019; Zheng et al., 2019; Xie et al., 2020). For RBGS, we can take an expectation over the random compression and δ will be identical to the compression ratio.

4.1. Convergence Rate

In this section, we bound the gradient norm $\|\nabla F(\theta_{\mathcal{A},\mathcal{S}})\|_2^2$ for convergence rate analysis of the proposed DEF method.

Theorem 4.3. *(Convergence Rate of DEF, Appendix A) Let Assumptions 3.1, 3.2, 3.3, 4.1 and 4.2 hold. If $\eta \le \frac{1}{4L}$, we have*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2 \le \frac{4\mathbb{E}[F_{\mathcal{S}}(y_0) - F_{\mathcal{S}}(y^*)]}{\eta T} + \frac{2\eta L\sigma^2}{K} + \frac{4\eta^2L^2[\frac{K-1}{K^2}\sigma^2 + (\frac{1}{K}-\lambda)^2\sigma^2 + 2(1-\lambda)^2M^2]}{(\sqrt{(1-\delta/2)/(1-\delta)} - 1)^2} . \tag{17}$$

Remark 4.4. Suppose $\theta_{\mathcal{A},\mathcal{S}}$ is randomly chosen from the sequence $\{y_t\}_{t=0}^{T-1}$, $\eta = \mathcal{O}(\sqrt{\frac{K}{T}}) \le \frac{1}{4L}$, and $K = \mathcal{O}(T^{1/3})$ (i.e., T is large enough), we have $\mathbb{E}\|\nabla F_{\mathcal{S}}(\theta_{\mathcal{A},\mathcal{S}})]\|_2^2 = \mathcal{O}(\frac{1}{\sqrt{KT}} + \frac{K}{T}) = \mathcal{O}(\frac{1}{\sqrt{KT}})$. It matches the rate of SGD with linear speedup regarding the number of workers K.

Remark 4.5. The last term in Eq. (17) is determined by the compression error. When δ, σ, M are of interest, we have $\mathbb{E}\|\nabla F_{\mathcal{S}}(\theta_{\mathcal{A},\mathcal{S}})]\|_2^2 =$

$$\mathcal{O}(\frac{\frac{K-1}{K^2}\sigma^2 + (\frac{1}{K}-\lambda)^2\sigma^2 + 2(1-\lambda)^2M^2}{(\sqrt{(1-\delta/2)/(1-\delta)}-1)^2}). \quad (18)$$

(1) When $K = 1$ (single worker) and $\lambda = 1$, it *vanishes*, which is better than EF (Karimireddy et al., 2019; Zheng et al., 2019).

(2) When σ and M are of interest, following the motivation in the previous section and ignoring other constant factors, Eq. (18) becomes

$$\mathcal{O}(\frac{K-1}{K^2}\sigma^2 + \frac{2(1-\frac{1}{K})^2\sigma^2M^2}{\sigma^2+2M^2}). \quad (19)$$

when $\lambda = \frac{\frac{1}{K}\sigma^2+2M^2}{\sigma^2+2M^2}$. It further reduces to $\mathcal{O}(\frac{K-1}{K}\sigma^2)$ when $M \to \infty$ (i.e. Assumption 3.2 does not hold). Therefore, DEF is the first EF variant compressing gradient *without relying on the bound of the gradient second moment*.

(3) When K is large and σ and M are of interest, our bound improves $\mathcal{O}(\sigma^2 + M^2)$ (Karimireddy et al., 2019; Zheng et al., 2019; Xie et al., 2020) to

$$\mathcal{O}(\frac{2\sigma^2M^2}{\sigma^2+2M^2}). \quad (20)$$

Our empirical deep learning experiments suggest that $\sigma^2 \approx 0.3M^2$, which means that our bound is about $\times 5$ *smaller* ignoring other constant factors.

4.2. Generalization Rate

In this section, we consider non-convex objective functions under PL condition, which establishes the relation between the gradient norm and the optimization error ϵ_{opt} (Zhou et al., 2021). We bound the excess risk error $\epsilon = \epsilon_{\text{opt}} + \epsilon_{gen}$ for the generalization analysis of the proposed DEF(-A) method.

Polyak-Łojasiewicz (PL) Condition (Karimi et al., 2016). Let $\theta^* \in \min_{\theta\in\mathbb{R}^d} F_{\mathcal{S}}(\theta)$. The objective function $F_{\mathcal{S}}(\theta)$ satisfies μ-PL condition if $\forall\theta \in \mathbb{R}^d$, we have

$$2\mu[F_{\mathcal{S}}(\theta) - F_{\mathcal{S}}(\theta^*)] \le \|\nabla F_{\mathcal{S}}(\theta)\|_2^2. \quad (21)$$

Theorem 4.6. *(Excess Risk Error of DEF(-A), Appendix B) Let Assumptions 3.1, 3.2, 3.3, 4.1 and 4.2 hold. Suppose $\eta = \frac{c}{t+1}$, where $c > 0$ is some constant.*

(1) The generalization error of DEF

$$\epsilon_{gen} = \mathcal{O}(T^{(1-\frac{K}{N})Lc/((1-\frac{K}{N})Lc+1)}). \quad (22)$$

(2) Suppose $\eta \le \frac{1}{4L}$. The optimization error of DEF

$$\epsilon_{opt} = \widetilde{\mathcal{O}}(T^{-\frac{\mu c}{2}} + T^{-1}). \quad (23)$$

(3) For RBGS, the generalization error of DEF-A

$$\epsilon_{gen} = \mathcal{O}(T^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc/((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc+1)}). \quad (24)$$

(4) Suppose $\eta \le \frac{1}{8L}$. The optimization error of DEF-A

$$\epsilon_{opt} = \widetilde{\mathcal{O}}(T^{-\frac{\mu\delta c}{2}} + T^{-1} + (1/\sqrt{1-\delta}-1)^{-2}). \quad (25)$$

Remark 4.7. When $K = 1$, Eq. (22) matches the result of SGD in Hardt et al. (2016).

Remark 4.8. DEF-A has a better ϵ_{gen} but a worse ϵ_{opt} than DEF. Since $\epsilon = \epsilon_{\text{gen}} + \epsilon_{\text{opt}}$, DEF-A can achieve *better* generalization rate than DEF via a trade-off between ϵ_{gen} and ϵ_{opt} with a proper δ.

Remark 4.9. Theorem 4.6 provides a potential new theoretical insight for applications incorporating compression, though some of them were not related to communication-efficient distributed training. E.g., escaping saddle point with compressed gradient (Avdiukhin & Yaroslavtsev, 2021), feature quantization to improve GAN training (Zhao et al., 2020), SignSGD that empirically accelerates training (Bernstein et al., 2018a;b), etc.

4.3. Extension to Iterate Averaging (IA)

As SGD and SGD-IA is a special case of DEF and DEF-A respectively when $K = 1$, $\lambda = 1$, and $\mathcal{C}(\Delta) = \delta\Delta$, we immediately have the following Theorem 4.10.

Theorem 4.10. *(Excess Risk Error of SGD(-IA), Appendix C) Let Assumptions 3.1, 3.2, 3.3, 4.1 and 4.2 hold. Suppose $\eta = \frac{c}{t+1}$, where $c > 0$ is some constant.*

(1) The generalization error of SGD

$$\epsilon_{gen} = \mathcal{O}(T^{(1-\frac{1}{N})Lc/((1-\frac{1}{N})Lc+1)}). \quad (26)$$

(2) Suppose $\eta \le \frac{1}{4L}$. The optimization error of SGD

$$\epsilon_{opt} = \widetilde{\mathcal{O}}(T^{-\frac{\mu c}{2}} + T^{-1}). \quad (27)$$

(3) The generalization error of SGD-IA

$$\epsilon_{gen} = \mathcal{O}(T^{(1-\frac{1}{N})\delta Lc/((1-\frac{1}{N})\delta Lc+1)}). \quad (28)$$

(4) Suppose $\eta \le \frac{1}{8\delta L}$. The optimization error of SGD-IA

$$\epsilon_{opt} = \widetilde{\mathcal{O}}(T^{-\frac{\mu\delta c}{2}} + T^{-1} + (1/\sqrt{1-\delta}-1)^{-2}). \quad (29)$$

Remark 4.11. We have $\delta^{\frac{1}{2}}$ in Eq. (24) but δ in Eq. (28) because $\mathbb{E}_{\mathcal{C}}[\mathcal{C}(\Delta)] = \delta\Delta$ for RBGS but $\mathcal{C}(\Delta) = \delta\Delta$ for SGD-IA.

Remark 4.12. SGD-IA can achieve better generalization rate than SGD with a proper δ. Neu & Rosasco (2018); Wu et al. (2020) theoretically only show that SGD-IA achieves adjustable regularization for strongly-convex objective functions, while SGD-IA applications such as averaging weights (Izmailov et al., 2018) and ensemble of models during training with cyclic learning rate (Huang et al., 2017) only empirically show better generalization than SGD.

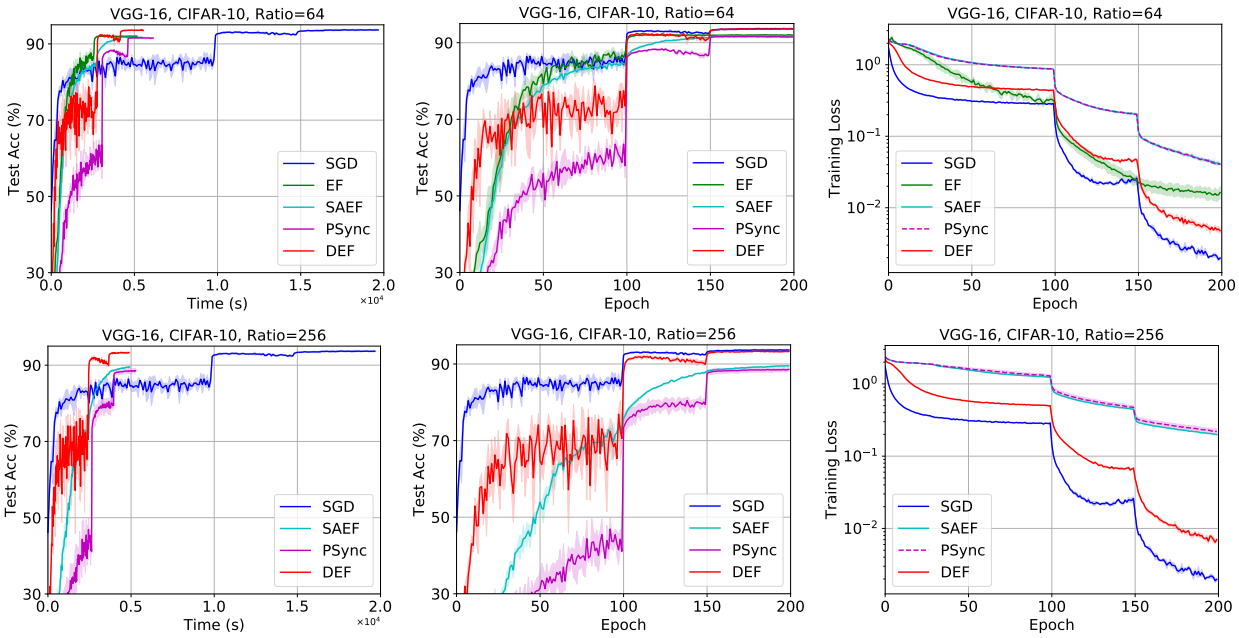

Figure 1. CIFAR-10 training curves of VGG-16. The compression ratio is 64 for the top row and 256 for the bottom row. EF is not plotted when the compression ratio is 256 due to divergence. From the left to right column, we plot the test accuracy (%) v.s. the wall-clock time, the test accuracy (%) v.s. training epochs, and the training loss v.s. training epochs respectively.

Table 1. The CIFAR-10 test accuracy (%) comparison of under various compression ratio settings with VGG-16.

Ratio	SGD	EF	SAEF	PSync	DEF	DEF-A
1	93.76 ± 0.14	—	—	—	—	—
16	—	93.04 ± 0.13	93.15 ± 0.04	93.31 ± 0.21	<u>93.61</u> ± 0.04	**93.66** ± 0.10
64	—	92.16 ± 0.06	91.88 ± 0.14	91.79 ± 0.17	**93.75** ± 0.12	<u>93.61</u> ± 0.07
256	—	diverge	89.59 ± 0.04	88.70 ± 0.61	**93.45** ± 0.11	<u>93.33</u> ± 0.26
512	—	diverge	87.83 ± 0.36	86.47 ± 0.14	<u>93.24</u> ± 0.08	**93.25** ± 0.18
1024	—	diverge	85.46 ± 0.80	84.27 ± 0.33	<u>93.03</u> ± 0.15	**93.06** ± 0.09

Remark 4.13. Compare with Theorem 4.6, we can see that DEF-A generalizes better than SGD with a proper δ.

Remark 4.14. Theorem 4.10 provides a new theoretical explanation for an important line of works in unsupervised learning - *momentum contrast* (He et al., 2020). In He et al. (2020), two sets of weights are maintained with a contrastive loss. One is the "query" y_t which is updated via SGD, and the other is the "key" x_t ($x_0 = y_0$) which is updated via

$$x_{t+1} = (1-\delta)x_t + \delta y_t\,. \tag{30}$$

The success of momentum contrast is explained as a "slowly progressing" key x_t (He et al., 2020) without theoretical guarantee. Interestingly, the above equation is identical to Eq. (12), i.e. SGD-IA. Therefore, our results suggests that the slowly progressing key x_t may actually have stabler and better generalization than the query y_t depending on δ.

5. Experiments

In this section, we conduct empirical experiments on benchmark deep learning tasks following settings in (Karimireddy et al., 2019; Zheng et al., 2019; Xie et al., 2020) to validate the performance of the proposed detached error feedback (DEF) method. We compare the following methods with RBGS as the gradient compressor: (1) SGD, which is the upper bound without gradient compression, (2) EF (Karimireddy et al., 2019; Zheng et al., 2019), (3) SAEF (Xu et al., 2021), (4) PSync (Xie et al., 2020), and (5) the proposed DEF(-A), where $\lambda = 0.3$ by default. We have also tested EF21 (Richtárik et al., 2021; Fatkhullin et al., 2021) on our deep learning tasks with RBGS, but it does not converge.

Settings. All experiments are implemented using PyTorch and conducted on a cluster of machines connected by. Each machine is equipped with 4 NVIDIA P40 GPUs and there

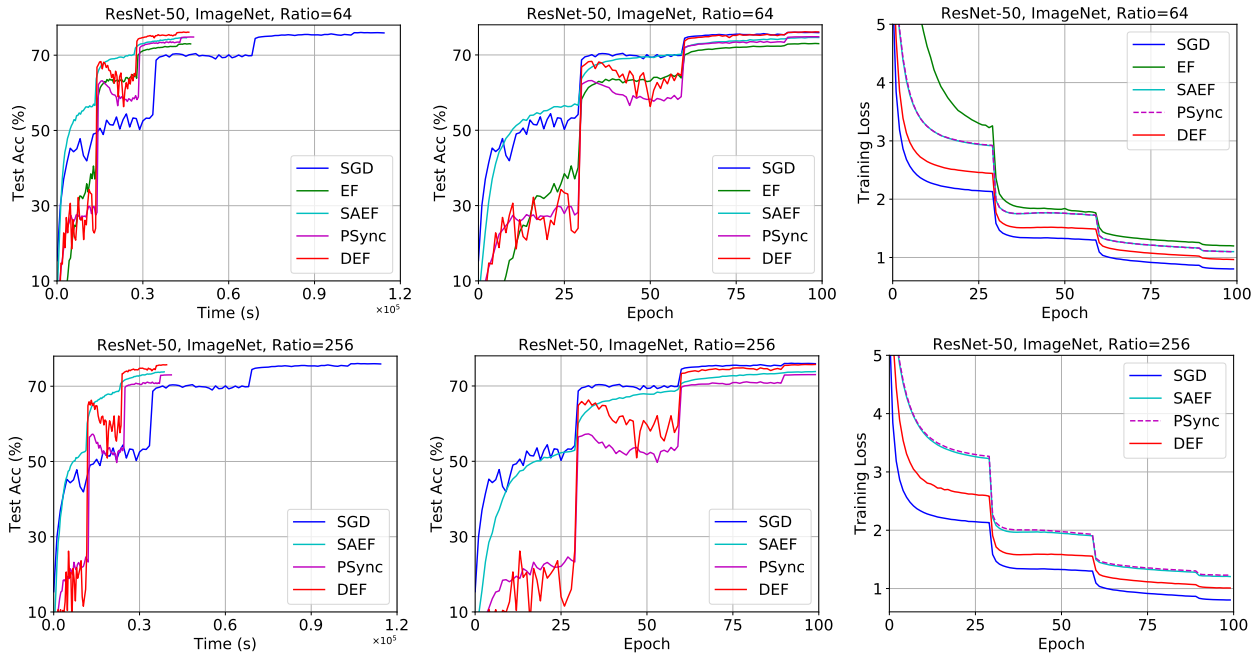

Figure 2. ImageNet training curves of ResNet-50. The compression ratio is 64 for the top row and 256 for the bottom row. From the left to right column, we plot the test accuracy (%) v.s. the wall-clock time, the test accuracy (%) v.s. training epochs, and the training loss v.s. training epochs respectively.

Table 2. The ImageNet test accuracy (%) comparison under various compression ratio settings with ResNet-50.

Ratio	SGD	EF	SAEF	PSync	DEF	DEF-A
1	76.04	—	—	—	—	—
16	—	75.29 (↓ 0.75)	75.83 (↓ 0.21)	75.63 (↓ 0.41)	75.98 (↓ 0.06)	**76.10** (↑ 0.06)
64	—	73.05 (↓ 2.99)	74.65 (↓ 1.39)	74.84 (↓ 1.20)	76.16 (↑ 0.12)	**76.37** (↑ 0.33)
128	—	63.80 (↓ 12.2)	74.26 (↓ 1.78)	74.12 (↓ 1.92)	**76.17** (↑ 0.13)	76.14 (↑ 0.10)
256	—	diverge	73.83 (↓ 2.21)	73.02 (↓ 3.02)	75.71 (↓ 0.33)	**76.00** (↓ 0.04)
512	—	diverge	73.00 (↓ 3.04)	72.60 (↓ 3.44)	75.52 (↓ 0.52)	**75.77** (↓ 0.27)
1024	—	diverge	71.89 (↓ 4.15)	71.82 (↓ 4.22)	**75.64** (↓ 0.40)	75.57 (↓ 0.47)

are 16 workers (GPUs) in total. We use NCCL as the backend of the PyTorch distributed package. The task-specific settings are as follows.

CIFAR. We train VGG-16 (Simonyan & Zisserman, 2014), ResNet-110 (He et al., 2016) and Wide ResNet (WRN-28-10) (Zagoruyko & Komodakis, 2016) models CIFAR-10/100 (Krizhevsky et al., 2009) image classification task. We report the mean and standard deviation metrics over 3 runs. The base learning rate is tuned from $\{\cdots, 0.1, 0.05, 0.01, \cdots\}$ and the batch size is 128. The momentum constant is 0.9 and the weight decay is 5×10^{-4}. The model is trained for 200 epochs with a learning rate decay of 0.1 at epoch 100 and 150. Random cropping, random flipping, and standardization are applied as data augmentation techniques.

ImageNet. We train the ResNet-50 model on ImageNet (Deng et al., 2009) image classification tasks. The model is trained for 100 epochs with a learning rate decay of 0.1 at epoch 30, 60, and 90. The base learning rate is tuned from $\{\cdots, 0.1, 0.05, 0.01, \cdots\}$ and the batch size is 256. The momentum constant is 0.9 and the weight decay is 1×10^{-4}. Similar data augmentation techniques as in CIFAR experiments are applied.

5.1. General Results

We plot the CIFAR-10 training curves of VGG-16 in Figure 1 and summarize the test numbers under various compression ratio settings in Table 1. From the curves, DEF achieves the best test acc and training loss among all the

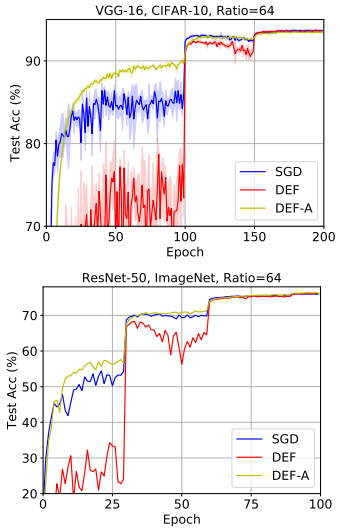

Figure 3. Accelerate the generalization with DEF-A. DEF-A significantly improves the test accuracy before the second learning rate decay compared with DEF.

communication-efficient methods. Compared with SGD, DEF achieves $\times 3.6$ and $\times 4.0$ speedup when the compression ratio is 64 and 256 respectively. For the test numbers in the table, DEF and DEF-A achieve the best results, which can be comparable to SGD for compression up to 256. When the compression ratio is high, DEF(-A) can significantly improve over the best counterpart by **8%**. Overall, DEF and DEF-A have similar final test performances. A significant improvement of the training loss over the existing EF variants can be observed, validating our lower bound in the convergence analysis of DEF.

The ImageNet training curves of ResNet-50 is shown in Figure 2 and the test numbers under various compression ratio settings are summarized in Table 2. We can reach similar conclusions as in CIFAR-10 experiments. Specifically, DEF achieves $\times 2.5$ and $\times 2.9$ speedup compared with SGD when the compression ratio is 64 and 256 respectively. For the test numbers in the table, DEF and DEF-A can be comparable to SGD for the compression ratio of 1024. For some smaller compression ratios, we may even see a slight improvement over SGD. When the compression ratio is high, DEF(-A) can significantly improve the best counterpart by **4%**.

For the concern of scalability, we also summarize the test numbers of VGG-16, ResNet-110, and WRN-28-10 on CIFAR-10/100 in Table 3 with 64 as the compression ratio. DEF-A achieves lossless performance compared with SGD and largely improves all the counterparts. We find that for VGG-16 on CIFAR-100 and ResNet-110 on CIFAR-10/100, DEF-A has a noticeable improvement over DEF. In particular, we find that PSync achieves closer performance to SGD on WRN as reported in Xie et al. (2020), but is much worse

Table 3. The CIFAR-10/100 test accuracy (%) comparison for various model architectures. The compression ratio is 1 for SGD and 64 for the other methods.

Method	VGG-16	ResNet-110	WRN-28-10
SGD	93.76 ± 0.14 / 72.50 ± 0.33	94.73 ± 0.06 / 76.78 ± 0.29	96.21 ± 0.07 / 80.81 ± 0.12
EF	92.16 ± 0.06 / 68.87 ± 0.21	diverge	diverge
SAEF	91.88 ± 0.14 / 67.00 ± 0.07	92.95 ± 0.17 / 71.19 ± 0.31	95.33 ± 0.01 / 79.04 ± 0.01
PSync	91.79 ± 0.17 / 65.68 ± 0.16	92.26 ± 0.04 / 69.21 ± 0.04	95.44 ± 0.13 / 79.60 ± 0.12
DEF	**93.75** ± 0.12 / $\underline{72.02} \pm 0.10$	$\underline{94.34} \pm 0.06$ / $\underline{76.43} \pm 0.12$	**96.26** ± 0.05 / $\underline{80.88} \pm 0.16$
DEF-A	$\underline{93.61} \pm 0.07$ / **72.38** ± 0.07	**94.66** ± 0.07 / **76.98** ± 0.21	$\underline{96.24} \pm 0.12$ / **80.95** ± 0.16

Table 4. The CIFAR-10 test accuracy (%) of DEF/DEF-A for various λ with VGG-16. The compression ratio is 64.

λ	0.1	0.2	**0.3**
DEF	92.83 ± 0.19	93.45 ± 0.06	**93.75** ± 0.12
DEF-A	92.78 ± 0.13	93.20 ± 0.14	$\underline{93.61} \pm 0.07$
λ	0.4	0.6	0.8
DEF	93.41 ± 0.12	93.26 ± 0.10	92.60 ± 0.19
DEF-A	93.41 ± 0.20	93.11 ± 0.20	92.59 ± 0.15

on VGG-16 and ResNet-110. Therefore, both the superior performance and scalability of DEF(-A) are validated.

5.2. Accelerate Generalization

Here we empirically validate the theoretical generalization analysis that DEF-A has a better generalization rate than DEF. We plot the training curves for VGG-16 on CIFAR-10 and ResNet-50 on ImageNet with compression ratio as 64 in Figure 3. We can see that DEF-A does have a much faster generalization rate than DEF. Specifically, the test accuracy improvement is about **15%** on CIFAR-10 and **25%** on ImageNet before the first learning rate decay, which validates the theoretical benefits in our generalization analysis. DEF-A can be even faster than full-precision SGD.

A significant improvement can still be observed before the second learning rate decay, but it becomes smaller when the learning rate is smaller. This matches our generalization analysis well. Let c be smaller such that the learning rate $\eta = \frac{c}{t+1}$ is smaller, then Eq. (22) is closer to Eq. (24), that is, the DEF-A's generalization error improvement over DEF becomes smaller. Then it is obvious that the excess risk error improvement will also become smaller.

5.3. Hyper-parameter λ

Here we explore DEF(-A) with various choices of the hyper-parameter λ with results summarized in Table 4. We can just set $\lambda = 0.3$ by default for the best performance. In comparison, an inappropriate choice of λ (e.g., 0.1 and 0.8) can lead to the performance degradation of about 1%. We also observe that a wide range of λ such as $0.2 \sim 0.6$ can result in fairly good performance compared with $\lambda = 0.3$, which means that the proposed DEF(-A) is not too sensitive to the hyper-parameter λ.

6. Conclusion

In this work, to address the performance loss issue for communication-efficient distributed SGD with the gradient sparsifier RBGS, we proposed a new DEF(-A) algorithm motivated by the trade-off between gradient variance and second moment. Our convergence analysis shows better bounds without relying on the bound of gradient second moment. We conduct the first generalization analysis for communication-efficient distributed training to show that DEF-A can generalize faster than DEF and SGD, which sheds light on other applications incorporating compression such as escaping saddle point, GAN training, and SignSGD training. We establish the connection to SGD-IA for the first time, thus our analysis provides potential theoretical explanations for SGD-IA applications such as averaging weights, ensemble, and momentum contrast in unsupervised learning. Last but not least, deep learning experiments validate the significant improvement of DEF(-A) over existing EF variants.

References

Aji, A. F. and Heafield, K. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 440–445, 2017.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pp. 5973–5983, 2018.

Assran, M., Loizou, N., Ballas, N., and Rabbat, M. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pp. 344–353. PMLR, 2019.

Avdiukhin, D. and Yaroslavtsev, G. Escaping saddle points with compressed sgd. *arXiv preprint arXiv:2105.10090*, 2021.

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

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018a.

Bernstein, J., Zhao, J., Azizzadenesheli, K., and Anandkumar, A. signsgd with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations*, 2018b.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Dean, J., Corrado, G. S., Monga, R., Chen, K., Devin, M., Le, Q. V., Mao, M. Z., Ranzato, M., Senior, A., Tucker, P., et al. Large scale distributed deep networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1223–1231, 2012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dutta, A., Bergou, E. H., Abdelmoniem, A. M., Ho, C.-Y., Sahu, A. N., Canini, M., and Kalnis, P. On the discrepancy between the theoretical analysis and practical implementations of compressed communication for distributed deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3817–3824, 2020.

Elibol, M., Lei, L., and Jordan, M. I. Variance reduction with sparse gradients. In *International Conference on Learning Representations*, 2019.

Fatkhullin, I., Sokolov, I., Gorbunov, E., Li, Z., and Richtárik, P. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Gao, H., Xu, A., and Huang, H. On the convergence of communication-efficient local sgd for federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence, Virtual*, pp. 18–19, 2021.

George, J. and Gurram, P. Distributed stochastic gradient descent with event-triggered communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7169–7178, 2020.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 1225–1234. PMLR, 2016.

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

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32:103–112, 2019.

Izmailov, P., Wilson, A., Podoprikhin, D., Vetrov, D., and Garipov, T. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pp. 876–885, 2018.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pp. 3252–3261. PMLR, 2019.

Koloskova, A., Lin, T., Stich, S. U., and Jaggi, M. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kuzborskij, I. and Lampert, C. Data-dependent stability of stochastic gradient descent. In *International Conference on Machine Learning*, pp. 2815–2824. PMLR, 2018.

Li, M., Zhou, L., Yang, Z., Li, A., Xia, F., Andersen, D. G., and Smola, A. Parameter server for distributed machine learning. In *Big Learning NIPS Workshop*, volume 6, pp. 2, 2013.

Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pp. 583–598, 2014a.

Li, M., Andersen, D. G., Smola, A. J., and Yu, K. Communication efficient distributed machine learning with the parameter server. *Advances in Neural Information Processing Systems*, 27:19–27, 2014b.

Li, Y., Yu, M., Li, S., Avestimehr, S., Kim, N. S., and Schwing, A. Pipe-sgd: A decentralized pipelined sgd framework for distributed deep net training. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Li, Z., Kovalev, D., Qian, X., and Richtarik, P. Acceleration for compressed gradient descent in distributed and federated optimization. In *International Conference on Machine Learning*, pp. 5895–5904. PMLR, 2020.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5336–5346, 2017.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, B. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.

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

Neu, G. and Rosasco, L. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pp. 3222–3242. PMLR, 2018.

Polyak, B. T. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Polyak, B. T. New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2, 1990.

Qian, X., Richtárik, P., and Zhang, T. Error compensated distributed SGD can be accelerated. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=dSqtddFibt2.

Richtárik, P., Sokolov, I., and Fatkhullin, I. Ef21: A new, simpler, theoretically better, and practically faster error feedback. *arXiv preprint arXiv:2106.05203*, 2021.

Ruppert, D. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

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

Shanbhag, A., Pirk, H., and Madden, S. Efficient top-k query processing on massively parallel hardware. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 1557–1570, 2018.

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

Song, L., Zhao, K., Pan, P., Liu, Y., Zhang, Y., Xu, Y., and Jin, R. Communication efficient sgd via gradient sampling with bayes prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12065–12074, 2021.

Stich, S. U. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified sgd with memory. In *Advances in Neural Information Processing Systems*, pp. 4447–4458, 2018.

Strom, N. Scalable distributed dnn training using commodity gpu cloud computing. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018.

Tang, H., Yu, C., Lian, X., Zhang, T., and Liu, J. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pp. 6155–6165. PMLR, 2019.

Tang, H., Li, Y., Liu, J., and Yan, M. Errorcompensatedx: error compensation for variance reduced algorithms. *Advances in Neural Information Processing Systems*, 34, 2021.

Vogels, T., Karinireddy, S. P., and Jaggi, M. Powersgd: Practical low-rank gradient compression for distributed optimization. *Advances In Neural Information Processing Systems 32 (Nips 2019)*, 32(CONF), 2019.

Wangni, J., Wang, J., Liu, J., and Zhang, T. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 1299–1309, 2018.

Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Wu, J., Braverman, V., and Yang, L. Obtaining adjustable regularization for free via iterate averaging. In *International Conference on Machine Learning*, pp. 10344–10354. PMLR, 2020.

Xie, C., Zheng, S., Koyejo, O. O., Gupta, I., Li, M., and Lin, H. Cser: Communication-efficient sgd with error reset. *Advances in Neural Information Processing Systems*, 33, 2020.

Xu, A., Huo, Z., and Huang, H. On the acceleration of deep learning model parallelism with staleness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2088–2097, 2020.

Xu, A., Huo, Z., and Huang, H. Step-ahead error feedback for distributed training with compressed gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10478–10486, 2021.

Yan, Y., Yang, T., Li, Z., Lin, Q., and Yang, Y. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 2955–2961. AAAI Press, 2018. ISBN 9780999241127.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, M., Lucas, J., Ba, J., and Hinton, G. E. Lookahead optimizer: k steps forward, 1 step back. *Advances in Neural Information Processing Systems*, 32:9597–9608, 2019.

Zhao, Y., Li, C., Yu, P., Gao, J., and Chen, C. Feature quantization improves gan training. In *International Conference on Machine Learning*, pp. 11376–11386. PMLR, 2020.

Zheng, S., Huang, Z., and Kwok, J. Communication-efficient distributed blockwise momentum sgd with error-feedback. *Advances in Neural Information Processing Systems*, 32:11450–11460, 2019.

Zhou, P., Yan, H., Yuan, X., Feng, J., and Yan, S. Towards understanding why lookahead generalizes better than sgd and beyond. *Advances in Neural Information Processing Systems*, 34, 2021.

Zhu, L., Lin, H., Lu, Y., Lin, Y., and Han, S. Delayed gradient averaging: Tolerate the communication latency for federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

A. Proof of Convergence of DEF (Theorem 4.3)

In this section, we consider general non-convex objective functions with δ-approximate and ring-allreduce compatible compressor. We want the following term to be as close to zero as possible.

$$\frac{1}{K}\sum_{k=1}^{K}\|\frac{1}{K}\sum_{k'=1}^{K}e_{k',t}-\lambda_t e_{k,t}\|_2^2\,. \tag{31}$$

For ease of notation, let $e_t \coloneqq \frac{1}{K}\sum_{k=1}^{K}e_{k,t}$, $\text{Var}(e_t) \coloneqq \frac{1}{K}\sum_{k=1}^{K}\|e_{k,t}-e_t\|_2^2$. We have $\frac{1}{K}\sum_{k=1}^{K}\|e_{k,t}\|_2^2 = \|e_t\|_2^2+\text{Var}(e_t)$. Then

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\|\frac{1}{K}\sum_{k'=1}^{K}e_{k,t}-\lambda_t e_{k,t}\|_2^2 = \frac{1}{K}\sum_{k=1}^{K}\|e_t-\lambda_t e_{k,t}\|^2 &= \frac{1}{K}\sum_{k=1}^{K}(\|e_t\|_2^2-2\lambda_t\langle e_t,e_{k,t}\rangle+\lambda_t^2\|e_{k,t}\|_2^2)\\ &= \frac{1}{K}\sum_{k=1}^{K}\|e_{k,t}\|_2^2\lambda_t^2-2\|e_t\|_2^2\lambda_t+\|e_t\|_2^2\\ &= (\|e_t\|_2^2+\text{Var}(e_t))\lambda_t^2-2\|e_t\|_2^2\lambda_t+\|e_t\|_2^2\,.\end{aligned} \tag{32}$$

When $\lambda_t^* = \frac{\|e_t\|_2^2}{\|e_t\|_2^2+\text{Var}(e_t)}$, it has the minimum $\frac{\|e_t\|_2^2\cdot\text{Var}(e_t)}{\|e_t\|_2^2+\text{Var}(e_t)}$.

If we allow individual coefficient $\lambda_{k,t}$ for each worker $k\in[K]$, then $\lambda_{k,t}^* = \frac{\langle e_t,e_{k,t}\rangle}{\|e_{k,t}\|_2^2}$ by minimizing $\|e_t-\lambda_{k,t}e_{k,t}\|_2^2$. However, it is impractical due to the additional K hyper-parameters to tune when K is large.

For simplicity, let $\lambda_t\to\lambda$ because it is hard to manually tune λ_t during the training.

A.1. Lemmas

For ease of notation, let $g_{k,t}$ denotes the stochastic gradient computed at iteration t for worker k and $g_t \coloneqq \frac{1}{K}\sum_{k=1}^{K}g_{k,t}$.

Lemma A.1. *Let Assumptions 3.1, 3.2, and 4.1 hold. Let $B_1 = (\frac{1}{K}-\lambda)^2+\frac{K-1}{K^2}$ and $B_2 = |\frac{1}{K}-\lambda|+\frac{K-1}{K}$. We have*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|g_t-\lambda g_{k,t}\|_2^2 \le B_1\sigma^2+2(1-\lambda)^2M^2+2B_2^2L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2\,, \tag{33}$$

Proof. We know that $g_{k,t} = \nabla f(x_t-\lambda e_{k,t};\xi_{k,t})$. Then

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|g_t-\lambda g_{k,t}\|_2^2\\ &= \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}[\nabla f(x_t-\lambda e_{k',t};\xi_{k',t})-\nabla F_{\mathcal{S}}(x_t-\lambda e_{k',t})]-\lambda[\nabla f(x_t-\lambda e_{k,t};\xi_{k,t})-\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})]\\ &\qquad+\frac{1}{K}\sum_{k'=1}^{K}[\nabla F_{\mathcal{S}}(x_t-\lambda e_{k',t})-\nabla F_{\mathcal{S}}(y_t)]-\lambda[\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)]+(1-\lambda)\nabla F_{\mathcal{S}}(y_t)\|_2^2\\ &\overset{(a)}{=} \underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}[\nabla f(x_t-\lambda e_{k',t};\xi_{k',t})-\nabla F_{\mathcal{S}}(x_t-\lambda e_{k',t})]-\lambda[\nabla f(x_t-\lambda e_{k,t};\xi_{k,t})-\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})]\|_2^2}_{①}\\ &\qquad+\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}[\nabla F_{\mathcal{S}}(x_t-\lambda e_{k',t})-\nabla F_{\mathcal{S}}(y_t)]-\lambda[\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)]+(1-\lambda)\nabla F_{\mathcal{S}}(y_t)\|_2^2\\ &\overset{(b)}{\le} ①+\underbrace{\frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}[\nabla F_{\mathcal{S}}(x_t-\lambda e_{k',t})-\nabla F_{\mathcal{S}}(y_t)]-\lambda[\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)]\|_2^2}_{②}+2(1-\lambda)^2M^2\,,\end{aligned} \tag{34}$$

where (a) is due to $\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})=\mathbb{E}g_{k,t}=\mathbb{E}\nabla f(x_t-\lambda e_{k,t};\xi_{k,t})$, (b) follows Assumption 3.2. Now we consider term ①. For simplicity, let $a_k=\nabla f(x_t-\lambda e_{k,t};\xi_{k,t})-\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})$ and we will have $\mathbb{E}\langle a_k,a_{k'}\rangle=0$ when $k\neq k'$. Then

$$① = \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|(\frac{1}{K}-\lambda)a_k+\frac{1}{K}\sum_{k'=1,k'\neq k}^{K}a_{k'}\|_2^2=\frac{1}{K}\sum_{k=1}^{K}\underbrace{[(\frac{1}{K}-\lambda)^2+\frac{K-1}{K^2}]}_{B_1}\|a_k\|_2^2\overset{(a)}{\leq}B_1\sigma^2\,, \tag{35}$$

where (a) follows Assumption 3.1. $B_1=0$ if and only if $K=1$ and $\lambda=1$. Now we consider term ②. For simplicity, let $b_k=\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)$ and $B_2=|\frac{1}{K}-\lambda|+\frac{K-1}{K}$. $B_2=0$ if and only if $K=1$ and $\lambda=1$. Then

$$\begin{aligned}
② &= \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}b_{k'}-\lambda b_k\|_2^2=\frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\|(\frac{1}{K}-\lambda)b_k+\frac{1}{K}\sum_{k'=1,k'\neq k}^{K}b_{k'}\|_2^2\\
&=\frac{2}{K}\sum_{k=1}^{K}B_2^2\mathbb{E}\|\frac{\frac{1}{K}-\lambda}{B_2}b_k+\frac{1}{KB_2}\sum_{k'=1,k'\neq k}^{K}b_{k'}\|_2^2\\
&\leq\frac{2}{K}\sum_{k=1}^{K}B_2^2\mathbb{E}(\frac{|\frac{1}{K}-\lambda|}{B_2}\|b_k\|_2^2+\frac{1}{KB_2}\sum_{k'=1,k'\neq k}^{K}\|b_{k'}\|_2^2)\\
&=\frac{2}{K}\sum_{k=1}^{K}B_2^2(\frac{|\frac{1}{K}-\lambda|}{B_2}+\frac{K-1}{KB_2})\mathbb{E}\|b_k\|_2^2=\frac{2B_2^2}{K}\sum_{k=1}^{K}\mathbb{E}\|b_k\|_2^2\\
&\overset{(a)}{\leq}2B_2^2L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2\,,
\end{aligned} \tag{36}$$

where (a) follows Assumption 4.1. Substitute terms ① and ② with their bounds and we can complete the proof.

□

Lemma A.2. *Let Assumptions 3.1, 3.2, 4.1, 4.2 and 3.3 hold. Let $B_1=(\frac{1}{K}-\lambda)^2+\frac{K-1}{K^2}$ and $\eta<\frac{1}{2L}$, we have*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}e_{k',t}-\lambda e_{k,t}\|_2^2\leq\frac{1}{(\sqrt{\frac{1-\delta/2}{1-\delta}}-1)^2}\eta^2[B_1\sigma^2+2(1-\lambda)^2M^2]\,. \tag{37}$$

Proof. We have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\frac{1}{K}\sum_{k'=1}^{K}e_{k',t}-\lambda e_{k,t}\|_2^2=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2\\
&\overset{(a)}{=}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|(\eta g_{t-1}+e_{t-1})-\mathcal{C}(\eta g_{t-1}+e_{t-1})-\lambda(\eta g_{k,t-1}+e_{k,t-1})+\lambda\mathcal{C}(\eta g_{k,t-1}+e_{k,t-1})\|_2^2\\
&\overset{(a)}{=}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|(\eta g_{t-1}-\lambda\eta g_{k,t-1}+e_{t-1}-\lambda e_{k,t-1})-\mathcal{C}(\eta g_{t-1}-\lambda\eta g_{k,t-1}+e_{t-1}-\lambda e_{k,t-1})\|_2^2\\
&\overset{(b)}{=}(1-\delta)\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\eta g_{t-1}-\lambda\eta g_{k,t-1}+e_{t-1}-\lambda e_{k,t-1}\|_2^2\\
&\leq(1-\delta)(1+\beta)\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{t-1}-\lambda e_{k,t-1}\|_2^2+(1-\delta)(1+\frac{1}{\beta})\eta^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|g_{t-1}-\lambda g_{k,t-1}\|_2^2\\
&\overset{(c)}{=}(1-\delta)(1+\beta)\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{t-1}-\lambda e_{k,t-1}\|_2^2+(1-\delta)(1+\frac{1}{\beta})\eta^2[B_1\sigma^2+2(1-\lambda)^2M^2]\\
&\quad+(1-\delta)(1+\frac{1}{\beta})2B_2^2\eta^2L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{t-1}-\lambda e_{k,t-1}\|_2^2\,,
\end{aligned} \tag{38}$$

where (a) follows Assumption 3.3, (b) follows Assumption 4.2, and (c) follows Lemma A.1. β is a constant such that $0<\beta<\frac{\delta}{1-\delta}$, i.e., $(1-\delta)(1+\beta)<1$. Let $B_3=(1-\delta)(1+\frac{1}{\beta})2B_2^2\eta^2L^2<1-(1-\delta)(1+\beta)$, i.e., $B_3+(1-\delta)(1+\beta)<1$, then

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2\\
&\leq [B_3+(1-\delta)(1+\beta)]\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{t-1}-\lambda e_{k,t-1}\|_2^2+(1-\delta)(1+\frac{1}{\beta})\eta^2[B_1\sigma^2+2(1-\lambda)^2M^2]\\
&=(1-\delta)(1+\frac{1}{\beta})[B_1\sigma^2+2(1-\lambda)^2M^2]\sum_{t'=0}^{t-1}[B_3+(1-\delta)(1+\beta)]^{t-1-t'}\eta^2\\
&<\underbrace{\frac{(1-\delta)(1+\frac{1}{\beta})}{1-B_3-(1-\delta)(1+\beta)}}_{h(\beta)}\eta^2[B_1\sigma^2+2(1-\lambda)^2M^2]\,.
\end{aligned}\tag{39}$$

Now we consider the minimum value of $h(\beta)$. Its gradient regarding β is

$$\frac{\partial h(\beta)}{\partial\beta}=\frac{1-\delta}{\beta^2[1-B_3-(1-\delta)(1+\beta)]^2}[(1-\delta)\beta^2+2(1-\delta)\beta+B_3-\delta]\,.\tag{40}$$

Therefore,

$$\begin{aligned}
\beta^*&=-1+\sqrt{\frac{1-B_3}{1-\delta}}\to B_3=1-(1-\delta)(1+\beta^*)^2<1-(1-\delta)(1+\beta^*),\text{ valid,}\\
h(\beta^*)&=\frac{1-\delta}{(\sqrt{1-B_3}-\sqrt{1-\delta})^2}=\frac{1}{(\sqrt{\frac{1-B_3}{1-\delta}}-1)^2}\,,\\
\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2&\leq\frac{1}{(\sqrt{\frac{1-B_3}{1-\delta}}-1)^2}\eta^2[B_1\sigma^2+2(1-\lambda)^2M^2]\,,
\end{aligned}\tag{41}$$

which completes the proof. For simplicity we can just set $B_3\leq\delta/2$ (we can choose a constant >1 other than 2), which is valid as it leads to $\beta^*\leq-1+\sqrt{\frac{1-\delta/2}{1-\delta}}$ and $-1+\sqrt{\frac{1-\delta/2}{1-\delta}}<\frac{\delta}{1-\delta}$ holds. Based on the definition of B_3, it also requires that

$$2B_2^2\eta^2L^2<\frac{\delta/2}{(1-\delta)(-1+\sqrt{\frac{1-\delta/2}{1-\delta}})}=\frac{\delta/2}{-(1-\delta)+\sqrt{(1-\delta)(1-\delta/2)}}\,,\tag{42}$$

where the R.H.S. is monotonically increasing for $0<\delta<1$. Therefore, for all conditions above to hold, we only need to assume that

$$2B_2^2\eta^2L^2\leq 4(1+(1-\lambda)^2)\eta^2L^2<\lim_{\delta\to0}\frac{\delta/2}{-(1-\delta)+\sqrt{(1-\delta)(1-\delta/2)}}=2\,.\tag{43}$$

As $0<\lambda<1$, we can simply assume $\eta<\frac{1}{2L}$. □

Lemma A.3. *Let Assumptions 3.1, 3.2, 4.2, and 3.3 hold. We have* $\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}\|_2^2\leq\frac{\eta^2(\sigma^2+M^2)}{(\sqrt{1/(1-\delta)}-1)^2}$.

Proof. We have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}\|_2^2 \overset{(a)}{=} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\eta g_{k,t-1}+e_{k,t-1}-\mathcal{C}(\eta g_{k,t-1}+e_{k,t-1})\|_2^2\\
&\overset{(b)}{\leq} \frac{1-\delta}{K}\sum_{k=1}^{K}\mathbb{E}\|\eta g_{k,t-1}+e_{k,t-1}\|_2^2\\
&\leq (1-\delta)(1+\beta)\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t-1}\|_2^2+(1-\delta)(1+\frac{1}{\beta})\eta^2(\sigma^2+M^2)\\
&\leq (1-\delta)(1+\frac{1}{\beta})(\sigma^2+M^2)\sum_{t'=0}^{t-1}[(1-\delta)(1+\beta)]^{t-1-t'}\eta^2\\
&\leq \frac{(1-\delta)(1+\frac{1}{\beta})}{1-(1-\delta)(1+\beta)}\eta^2(\sigma^2+M^2)\\
&= \frac{1}{(\sqrt{1/(1-\delta)}-1)^2}\eta^2(\sigma^2+M^2)\,,
\end{aligned} \tag{44}$$

where $\beta=-1+\frac{1}{\sqrt{1-\delta}}$, (a) follows Assumption 3.3, and (b) follows Assumption 4.2. □

A.2. Main Proof

In this section, we need Assumptions 3.1, 3.2, 4.1, 4.2, 3.3, and $\eta\leq\frac{1}{4L}$.

Firstly, we have the update rule of y_t

$$\begin{aligned}
y_{t+1} &= x_{t+1}-e_{t+1} = x_{t+1}-\frac{1}{K}\sum_{k=1}^{K}e_{k,t+1}\\
&= x_t-\frac{1}{K}\sum_{k=1}^{K}\mathcal{C}(\eta g_{k,t}+e_{k,t})-\frac{1}{K}\sum_{k=1}^{K}(\eta g_{k,t}+e_{k,t}-\mathcal{C}(\eta g_{k,t}+e_{k,t}))\\
&= x_t-\frac{1}{K}\sum_{k=1}^{K}(\eta g_{k,t}+e_{k,t}) = y_t-\frac{1}{K}\sum_{k=1}^{K}\eta g_{k,t} = y_t-\eta g_t\,.
\end{aligned} \tag{45}$$

According to the Liptschitz gradient assumption,

$$\begin{aligned}
&\mathbb{E}[F_{\mathcal{S}}(y_{t+1})-F_{\mathcal{S}}(y_t)] \leq \mathbb{E}\langle\nabla F_{\mathcal{S}}(y_t), y_{t+1}-y_t\rangle+\frac{L}{2}\mathbb{E}\|y_{t+1}-y_t\|_2^2\\
&= \mathbb{E}\langle\nabla F_{\mathcal{S}}(y_t), -\frac{\eta}{K}\sum_{k=1}^{K}g_{k,t})\rangle+\frac{\eta^2L}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}g_{k,t}\|_2^2\\
&= \underbrace{-\frac{\eta}{K}\sum_{k=1}^{K}\mathbb{E}\langle\nabla F_{\mathcal{S}}(y_t), \nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})\rangle}_{①}+\frac{\eta^2L}{2}\underbrace{\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})\|_2^2}_{②}+\frac{\eta^2L\sigma^2}{2K}\,.
\end{aligned} \tag{46}$$

For term ①, we have

$$\begin{aligned}
① &= -\eta\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2-\frac{\eta}{K}\sum_{k=1}^{K}\mathbb{E}\langle\nabla F_{\mathcal{S}}(y_t), \nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)\rangle\\
&\leq -\frac{\eta}{2}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2+\frac{\eta}{2K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)\|_2^2\\
&\leq -\frac{\eta}{2}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2+\frac{\eta L^2}{2K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2\,.
\end{aligned} \tag{47}$$

For term ②, we have

$$\begin{aligned} ② &\leq \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[2\|\nabla F_{\mathcal{S}}(x_t-\lambda e_{k,t})-\nabla F_{\mathcal{S}}(y_t)\|_2^2+2\|\nabla F_{\mathcal{S}}(y_t)\|_2^2] \\ &\leq 2\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2+\frac{2L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2\,. \end{aligned} \tag{48}$$

Replace ① and ② with their bounds and we have

$$\begin{aligned} \mathbb{E}[f(y_{t+1})-f(y_t)] &\leq (-\frac{\eta}{2}+\eta^2 L)\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2+(\frac{\eta L^2}{2}+\eta^2L^3)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2+\frac{\eta^2L\sigma^2}{2K} \\ &\overset{(a)}{\leq} -\frac{\eta}{4}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2+\eta L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2+\frac{\eta^2L\sigma^2}{2K}\,, \end{aligned} \tag{49}$$

where (a) is due to the assumption $\eta \leq \frac{1}{4L}$ for simplicity. Rearrange and sum from $t=0$ to $T-1$, we will have

$$\begin{aligned} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2 &\leq \frac{4\mathbb{E}[F_{\mathcal{S}}(y_0)-F_{\mathcal{S}}(y_T)]}{\eta T}+\frac{2\eta L\sigma^2}{K}+4L^2\cdot\frac{1}{KT}\sum_{t=0}^{T-1}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2 \\ &\overset{(a)}{\leq} \frac{4\mathbb{E}[F_{\mathcal{S}}(y_0)-F_{\mathcal{S}}(y^*)]}{\eta T}+\frac{2\eta L\sigma^2}{K}+\frac{4}{(\sqrt{\frac{1-\delta/2}{1-\delta}}-1)^2}\eta^2L^2[B_1\sigma^2+2(1-\lambda)^2M^2] \\ &= \frac{4\mathbb{E}[F_{\mathcal{S}}(y_0)-F_{\mathcal{S}}(y^*)]}{\eta T}+\frac{2\eta L\sigma^2}{K}+\frac{4}{(\sqrt{\frac{1-\delta/2}{1-\delta}}-1)^2}\eta^2L^2[\frac{K-1}{K^2}\sigma^2+(\frac{1}{K}-\lambda)^2\sigma^2+2(1-\lambda)^2M^2]\,, \end{aligned} \tag{50}$$

where (a) follows Lemma A.2. Let $\eta=\mathcal{O}(\sqrt{\frac{K}{T}})$, we have the convergence rate

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2=\mathcal{O}(\frac{1}{\sqrt{KT}}+\frac{K}{T})\overset{K=\mathcal{O}(T^{1/3})}{=}\mathcal{O}(\frac{1}{\sqrt{KT}})\,. \tag{51}$$

If we are only interested in δ, σ^2 and M^2, let $\lambda=\frac{\frac{1}{K}\sigma^2+2M^2}{\sigma^2+2M^2}$ and we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2=\mathcal{O}(\frac{1}{(\sqrt{(1-\delta/2)/(1-\delta)}-1)^2}\cdot(\frac{K-1}{K^2}\sigma^2+\frac{2(1-\frac{1}{K})^2\sigma^2M^2}{\sigma^2+2M^2}))\,. \tag{52}$$

B. Proof of Generalization of DEF(-A) (Theorem 4.6)

We use uniform stability to bound the generalization error. Let $\mathcal{S}=\{\xi_1,\xi_2,\cdots,\xi_N\}$ be the training dataset of size N, where each data ξ_n is sampled from distribution $\mathcal{D}$. Let $\mathcal{S}^{(n)}=\{\xi_1',\xi_2',\cdots,\xi_N'\}=\{\xi_1,\xi_2,\cdots,\xi_{n-1},\xi_n',\xi_{n+1},\cdots,\xi_N\}$ be another training datasets of size N. We can see that $\mathcal{S}$ and $\mathcal{S}^{(n)}$ only differs in the n^{th} data. Let the models trained on $\mathcal{S}$ and $\mathcal{S}^{(i)}$ be x_t and $\widetilde{x}_t$ respectively for DEF-A. For DEF, they will be y_t and $\widetilde{y}_t$ correspondingly.

In each iteration t and under the same random sampling procedure, all workers select the same data from $\mathcal{S}$ and $\mathcal{S}^{(n)}$ with probability $\binom{N-1}{K}/\binom{N}{K}=\frac{N-K}{N}$, while one of the workers selects different data from $\mathcal{S}$ and $\mathcal{S}^{(n)}$ with probability $1-\binom{N-1}{K}/\binom{N}{K}=\frac{K}{N}$. For simplicity, let $G=\sqrt{\sigma^2+M^2}$, $B_2=|\frac{1}{K}-\lambda|+\frac{K-1}{K}$. Following Hardt et al. (2016) (Theorem 3.8), we only need to bound (x_t for DEF-A and y_t for DEF)

$$\mathbb{E}[f(y_t;\xi)-f(\widetilde{y}_T;\xi)]\leq\frac{Kt_0}{N}+G\mathbb{E}[\|y_T-\widetilde{y}_T\|_2|y_{t_0}-\widetilde{y}_{t_0}=0]\,. \tag{53}$$

B.1. Generalization Error of DEF

In this section, we consider non-convex objective functions. We need Assumptions 3.1, 3.2, 4.1, 4.2, 3.3 and $\eta_t \leq \frac{c}{t+1}$. We first consider y_t selecting the same data at iteration t.

$$
\begin{aligned}
&\|y_{t+1} - \widetilde{y}_{t+1}\|_2 = \|y_t - \eta_t g_t - \widetilde{y}_t + \eta_t \widetilde{g}_t\|_2 \leq \|y_t - \widetilde{y}_t\|_2 + \eta_t \|g_t - \widetilde{g}_t\|_2 \\
&= \|y_t - \widetilde{y}_t\|_2 + \eta_t \|\frac{1}{K}\sum_{k=1}^{K}[\nabla f(x_t - \lambda e_{k,t}; \xi_{k,t}) - \nabla f(\widetilde{x}_t - \lambda \widetilde{e}_{k,t}; \xi_{k,t})]\|_2 \\
&\leq \|y_t - \widetilde{y}_t\|_2 + \frac{\eta_t}{K}\sum_{k=1}^{K}\|\nabla f(y_t + e_t - \lambda e_{k,t}; \xi_{k,t}) - \nabla f(\widetilde{y}_t + \widetilde{e}_t - \lambda \widetilde{e}_{k,t}; \xi_{k,t})\|_2 \\
&\leq (1 + \eta_t L)\|y_t - \widetilde{y}_t\|_2 + \frac{\eta_t L}{K}\sum_{k=1}^{K}\|(e_t - \lambda e_{k,t}) - (\widetilde{e}_t - \lambda \widetilde{e}_{k,t})\|_2 \\
&= (1 + \eta_t L)\|y_t - \widetilde{y}_t\|_2 + \frac{\eta_t L}{K}\sum_{k=1}^{K}\|(\frac{1}{K} - \lambda)(e_{k,t} - \widetilde{e}_{k,t}) + \frac{1}{K}\sum_{k'=1, k'\neq k}^{K}(e_{k',t} - \widetilde{e}_{k',t})\|_2 \\
&\leq (1 + \eta_t L)\|y_t - \widetilde{y}_t\|_2 + \frac{\eta_t L B_2}{K}\sum_{k=1}^{K}\underbrace{\|e_{k,t} - \widetilde{e}_{k,t}\|_2}_{①}\,.
\end{aligned}
\tag{54}
$$

Now we consider term ①. Following the same procedures in Lemma A.3, but let $\eta_t \leq \frac{c}{t+1}$ and $\beta = \frac{\delta}{2(1-\delta)}$, we have

$$
\begin{aligned}
&\mathbb{E}\|e_{k,t} - \widetilde{e}_{k,t}\|_2^2 \leq (1-\delta)(1 + \frac{1}{\beta}) \cdot 2(\sigma^2 + M^2)\sum_{t'=0}^{t-1}[(1-\delta)(1+\beta)]^{t-1-t'}\eta_{t'}^2 \\
&\leq 2(1-\delta)(1+\frac{1}{\beta})(\sigma^2 + M^2)c^2\sum_{t'=0}^{t-1}\frac{1}{(t'+1)^2} \\
&\leq 2(1-\delta)(1+\frac{1}{\beta})(\sigma^2 + M^2)c^2[1 + (-\frac{1}{t'+1})|_0^{t-1}] \\
&\leq 4(1-\delta)(1+\frac{1}{\beta})(\sigma^2 + M^2)c^2 \\
&= \frac{4(1-\delta)(2-\delta)}{\delta}G^2c^2\,.
\end{aligned}
\tag{55}
$$

Because $(\mathbb{E}\|e_{k,t} - \widetilde{e}_{k,t}\|_2)^2 \leq \mathbb{E}\|e_{k,t} - \widetilde{e}_{k,t}\|_2^2$, we have

$$
\mathbb{E}① \leq \sqrt{\mathbb{E}\|e_{k,t} - \widetilde{e}_{k,t}\|_2^2} \leq 2Gc\underbrace{\sqrt{\frac{(1-\delta)(2-\delta)}{\delta}}}_{B_4} = 2B_4Gc\,.
\tag{56}
$$

At iteration t, if a worker $k' \in [K]$ selects different data from $\mathcal{S}$ and $\mathcal{S}^{(n)}$, we have

$$
\|y_{t+1} - \widetilde{y}_{t+1}\|_2 \leq \|y_t - \widetilde{y}_t\|_2 + \eta_t\|g_t - \widetilde{g}_t\|_2 \leq \|y_t - \widetilde{y}_t\|_2 + 2G\eta_t\,.
\tag{57}
$$

When we consider both circumstances, we have

$$
\begin{aligned}
&\mathbb{E}\|y_{t+1} - \widetilde{y}_{t+1}\|_2 \leq (1 - \frac{K}{N})[(1+\eta_t L)\mathbb{E}\|y_t - \widetilde{y}_t\|_2 + \eta_t L B_2 \cdot 2B_4Gc] + \frac{K}{N}[\mathbb{E}\|y_t - \widetilde{y}_t\|_2 + 2G\eta_t] \\
&= [1 + (1 - \frac{K}{N})\eta_t L]\mathbb{E}\|y_t - \widetilde{y}_t\|_2 + \underbrace{[\frac{2K}{N}G + 2(1 - \frac{K}{N})LB_2B_4Gc]}_{B_5}\eta_t \\
&\overset{(a)}{\leq} \exp((1 - \frac{K}{N})\eta_t L)\mathbb{E}\|y_t - \widetilde{y}_t\|_2 + B_5\eta_t\,.
\end{aligned}
\tag{58}
$$

Unwind the recurrence with $t = 0, 1, \cdots, T-1$, we have

$$\begin{aligned}
&\mathbb{E}\|y_T - \widetilde{y}_T\|_2 \leq \sum_{t=t_0}^{T-1} B_5 \frac{c}{t+1} \prod_{t'=t+1}^{T-1} \exp((1-\frac{K}{N})\frac{c}{t+1}L) \\
&= B_5 c \sum_{t=t_0}^{T-1} \frac{1}{t+1} \exp((1-\frac{K}{N})Lc \sum_{t'=t+1}^{T-1} \frac{1}{t+1}) \\
&\overset{(a)}{\leq} B_5 c \sum_{t=t_0}^{T-1} \frac{1}{t+1} \exp((1-\frac{K}{N})Lc \log \frac{T}{t+1}) \\
&= B_5 c T^{(1-\frac{K}{N})Lc} \sum_{t=t_0}^{T-1} (t+1)^{-(1-\frac{K}{N})Lc-1} \\
&= B_5 c T^{(1-\frac{K}{N})Lc} \frac{t_0^{-(1-\frac{K}{N})Lc} - T^{-(1-\frac{K}{N})Lc}}{(1-\frac{K}{N})Lc} \\
&\leq \frac{B_5}{(1-\frac{K}{N})L} (\frac{T}{t_0})^{(1-\frac{K}{N})Lc}
\end{aligned} \tag{59}$$

where (a) is due to $\sum_{t'=t+1}^{T-1} \frac{1}{t'+1} \leq \int_t^{T-1} \log(t'+1)dt'$. Following Eq. (53),

$$\mathbb{E}[f(y_t;\xi) - f(\widetilde{y}_T;\xi)] \leq \frac{Kt_0}{N} + \underbrace{\frac{B_5 G}{(1-\frac{K}{N})L}}_{B_6} (\frac{T}{t_0})^{(1-\frac{K}{N})Lc}. \tag{60}$$

The R.H.S is minimized when

$$t_0 = [(\frac{N}{K}-1)LcB_6]^{1/((1-\frac{K}{N})Lc+1)} T^{(1-\frac{K}{N})Lc/((1-\frac{K}{N})Lc+1)}, \tag{61}$$

which gives us

$$\mathbb{E}[f(y_T;\xi) - f(\widetilde{y}_T;\xi)] \leq [\frac{K}{N} + \frac{1}{(\frac{N}{K}-1)Lc}][(\frac{N}{K}-1)LcB_6]^{1/((1-\frac{K}{N})Lc+1)} T^{(1-\frac{K}{N})Lc/((1-\frac{K}{N})Lc+1)}. \tag{62}$$

Note that when $K = 1$, $\lambda = 1$, we will have $B_2 = 0$, $B_5 = \frac{2K}{N}G$, and $B_6 = \frac{2G^2}{(N-1)L}$. Then the R.H.S. equals

$$\begin{aligned}
&[\frac{1}{N} + \frac{1}{(N-1)Lc}](2cG^2)^{1/((1-\frac{1}{N})Lc+1)} T^{(1-\frac{1}{N})Lc/((1-\frac{1}{N})Lc+1)} \\
&= \frac{1+1/(Lc)}{N-1} T (\frac{2cG^2}{T})^{1/((1-\frac{1}{N})Lc+1)} \\
&\overset{(a)}{\leq} \frac{1+1/(Lc)}{N-1} T (\frac{2cG^2}{T})^{1/(Lc+1)} \\
&= \frac{1+1/(Lc)}{N-1} (2cG^2)^{1/(Lc+1)} T^{Lc/(Lc+1)},
\end{aligned} \tag{63}$$

which matches the result in Hardt et al. (2016) for SGD. (a) is due to $\frac{2cG^2}{T} \leq 1$ when $t_0 \leq T$.

B.2. Generalization Error of DEF-A

In this section, we consider non-convex objective functions and random sparsification which satisfies Assumptions 4.2 and 3.3. We need Assumptions 3.1, 3.2, 4.1, and $\eta_t \leq \frac{c}{t+1}$.

Now we consider x_t.

$$\begin{aligned}
&\mathbb{E}\|x_{t+1}-\widetilde{x}_{t+1}\|_2 \overset{(a)}{=} \mathbb{E}\|x_t-\mathcal{C}(\eta_t g_t+e_t)-\widetilde{x}_t+\mathcal{C}(\eta_t\widetilde{g}_t+\widetilde{e}_t)\|_2\\
&\overset{(a)}{\leq} \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\mathbb{E}\|\mathcal{C}(\eta_t g_t-\eta_t\widetilde{g}_t)\|_2+\mathbb{E}\|\mathcal{C}(e_t-\widetilde{e}_t)\|_2\\
&\leq \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\mathbb{E}\sqrt{\mathbb{E}_{\mathcal{C}}\|\mathcal{C}(\eta_t g_t-\eta_t\widetilde{g}_t)\|_2^2}+\mathbb{E}\sqrt{\mathbb{E}_{\mathcal{C}}\|\mathcal{C}(e_t-\widetilde{e}_t)\|_2^2}\\
&\overset{(a)}{=} \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\mathbb{E}\sqrt{\delta\eta_t^2\|g_t-\widetilde{g}_t\|_2^2}+\mathbb{E}\sqrt{\delta\|e_t-\widetilde{e}_t\|_2^2}\\
&= \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\delta^{\frac{1}{2}}\eta_t\mathbb{E}\|g_t-\widetilde{g}_t\|_2+\delta^{\frac{1}{2}}\mathbb{E}\|e_t-\widetilde{e}_t\|_2\\
&\leq \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\delta^{\frac{1}{2}}\eta_t\mathbb{E}\|g_t-\widetilde{g}_t\|_2+\delta^{\frac{1}{2}}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2\,.
\end{aligned} \tag{64}$$

where (a) is due to the random sparsification. When selecting the same data at iteration t, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t+1}-\widetilde{x}_{t+1}\|_2\\
&\leq \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{\delta^{\frac{1}{2}}}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2+\delta^{\frac{1}{2}}\eta_t\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}[\nabla f(x_t-\lambda e_{k,t};\xi_{k,t})-\nabla f(\widetilde{x}_t-\lambda\widetilde{e}_{k,t};\xi_{k,t})]\|_2\\
&\leq \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{\delta^{\frac{1}{2}}}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2+\frac{\delta^{\frac{1}{2}}\eta_t L}{K}\sum_{k=1}^{K}\|x_t-\lambda e_{k,t}-\widetilde{x}_t+\lambda\widetilde{e}_{k,t}\|_2\\
&\leq (1+\delta^{\frac{1}{2}}\eta_t L)\mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{\delta^{\frac{1}{2}}(1+\eta_t L\lambda)}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2\,.
\end{aligned} \tag{65}$$

When selecting different data at iteration t, we have

$$\mathbb{E}\|x_{t+1}-\widetilde{x}_{t+1}\|_2\leq \mathbb{E}\|x_t-\widetilde{x}_t\|_2+\delta^{\frac{1}{2}}\eta_t\cdot 2G+\frac{\delta^{\frac{1}{2}}}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2\,. \tag{66}$$

When we consider both circumstances, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1}-\widetilde{x}_{t+1}\|_2 &\leq (1-\frac{K}{N})[(1+\delta^{\frac{1}{2}}\eta_t L)\mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{\delta^{\frac{1}{2}}(1+\eta_t L\lambda)}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2]\\
&\quad+\frac{K}{N}[\mathbb{E}\|x_t-\widetilde{x}_t\|_2+\delta^{\frac{1}{2}}\eta_t\cdot 2G+\frac{\delta^{\frac{1}{2}}}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2]\\
&= [1+(1-\frac{K}{N})\delta^{\frac{1}{2}}\eta_t L]\mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{K}{N}2\delta^{\frac{1}{2}}\eta_t G+\delta^{\frac{1}{2}}[1+(1-\frac{K}{N})\eta_t L\lambda]\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}-\widetilde{e}_{k,t}\|_2\\
&\overset{(a)}{\leq} [1+(1-\frac{K}{N})\delta^{\frac{1}{2}}\eta_t L]\mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{K}{N}2\delta^{\frac{1}{2}}\eta_t G+\delta^{\frac{1}{2}}[1+(1-\frac{K}{N})\eta_t L\lambda]\cdot 2B_4 Gc\\
&\overset{(b)}{\leq} \exp((1-\frac{K}{N})\delta^{\frac{1}{2}}\eta_t L)\mathbb{E}\|x_t-\widetilde{x}_t\|_2+\frac{K}{N}2\delta^{\frac{1}{2}}\eta_t G+\delta^{\frac{1}{2}}[1+(1-\frac{K}{N})\eta_t L\lambda]\cdot 2B_4 Gc\,,
\end{aligned} \tag{67}$$

where (a) follows the Eq. (56). Let $\eta_t \leq \frac{c}{t+1}$, we have

$$\begin{aligned}
\mathbb{E}\|x_T - \widetilde{x}_T\|_2 &\leq \sum_{t=t_0}^{T-1}[\frac{K}{N}2\delta^{\frac{1}{2}}\eta_t G + \delta^{\frac{1}{2}}(1+(1-\frac{K}{N})\eta_t L\lambda)\cdot 2B_4 Gc]\prod_{t'=t+1}^{T-1}\exp((1-\frac{K}{N})\delta^{\frac{1}{2}}\eta_{t'}L) \\
&\leq 2\delta^{\frac{1}{2}}G\sum_{t=t_0}^{T-1}[\frac{K}{N}\frac{c}{t+1}+(1+(1-\frac{K}{N})\frac{c}{t+1}L\lambda)B_4 c]\exp((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc\sum_{t''=t+1}^{T-1}\frac{1}{t''+1}) \\
&\leq 2\delta^{\frac{1}{2}}G\sum_{t=t_0}^{T-1}[\frac{K}{N}\frac{c}{t+1}+(1+(1-\frac{K}{N})\frac{c}{t+1}L\lambda)B_4 c]\exp((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc\log(\frac{T}{t+1})) \\
&= 2\delta^{\frac{1}{2}}G\sum_{t=t_0}^{T-1}[\frac{K}{N}\frac{c}{t+1}+(1+(1-\frac{K}{N})\frac{c}{t+1}L\lambda)B_4 c](\frac{T}{t+1})^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc} \\
&\leq 2\delta^{\frac{1}{2}}Gc[\frac{K}{N}+1+(1-\frac{K}{N})Lc\lambda B_4]T^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc}\sum_{t=t_0}^{T-1}(t+1)^{-(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc} \\
&\leq 2\delta^{\frac{1}{2}}Gc[\frac{K}{N}+1+(1-\frac{K}{N})Lc\lambda B_4]T^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc}\frac{t_0^{-1-(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc}-T^{-1-(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc}}{1+(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc} \\
&\leq \frac{2\delta^{\frac{1}{2}}G[\frac{K}{N}+1+(1-\frac{K}{N})Lc\lambda B_4]}{1+(1-\frac{K}{N})\delta^{\frac{1}{2}}L}(\frac{T}{t_0})^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc}.
\end{aligned} \tag{68}$$

Following Eq. (53),

$$\mathbb{E}[f(x_T;\xi)-f(\widetilde{x}_T;\xi)] \leq \frac{Kt_0}{N}+\underbrace{\frac{2\delta^{\frac{1}{2}}G[\frac{K}{N}+1+(1-\frac{K}{N})Lc\lambda B_4]}{1+(1-\frac{K}{N})\delta^{\frac{1}{2}}L}}_{B_7}(\frac{T}{t_0})^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc}. \tag{69}$$

The R.H.S is minimized when

$$t_0 = [(\frac{N}{K}-1)LcB_7]^{1/((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc+1)}T^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc/((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc+1)}, \tag{70}$$

which gives us

$$\mathbb{E}[f(x_T;\xi)-f(\widetilde{x}_T;\xi)] \leq [\frac{K}{N}+\frac{1}{(\frac{N}{K}-1)Lc}][(\frac{N}{K}-1)LcB_7]^{1/((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc+1)}T^{(1-\frac{K}{N})\delta^{\frac{1}{2}}Lc/((1-\frac{K}{N})\delta^{\frac{1}{2}}Lc+1)}. \tag{71}$$

B.3. Optimization Error of DEF

In this section, we consider non-convex objective functions satisfying the Polyak-Łojasiewicz (PL) condition, which establishes the relation between the objective function and the gradient norm. We consider δ-approximate and ring-allreduce compatible compressor. We need Assumptions 3.1, 3.2, 4.1, 4.2, 3.3, and $\eta_t = \frac{c}{t+1} \leq \frac{1}{4L}$. From Eq. (49) and the PL condition, we have

$$\begin{aligned}
\mathbb{E}[F_{\mathcal{S}}(y_{t+1})-F_{\mathcal{S}}(y_t)] &\leq -\frac{\eta_t}{4}\mathbb{E}\|\nabla F_{\mathcal{S}}(y_t)\|_2^2+\eta_t L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2+\frac{\eta_t^2 L\sigma^2}{2K} \\
&\leq -\frac{\mu\eta_t}{2}\mathbb{E}[F_{\mathcal{S}}(y_t)-F_{\mathcal{S}}(y^*)]+\eta_t L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2+\frac{\eta_t^2 L\sigma^2}{2K},
\end{aligned} \tag{72}$$

where we need $\eta_t \leq \frac{1}{4L}$ following Section A. Rearrange,

$$\begin{aligned}
\mathbb{E}[F_{\mathcal{S}}(y_{t+1})-F_{\mathcal{S}}(y^*)] &\leq (1-\frac{\mu\eta_t}{2})\mathbb{E}[F_{\mathcal{S}}(y_t)-F_{\mathcal{S}}(y^*)]+\eta_t L^2\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_t-\lambda e_{k,t}\|_2^2+\frac{\eta_t^2 L\sigma^2}{2K} \\
&\overset{(a)}{\leq} (1-\frac{\mu\eta_t}{2})\mathbb{E}[F_{\mathcal{S}}(y_t)-F_{\mathcal{S}}(y^*)]+\frac{B_1\sigma^2+2(1-\lambda)^2M^2}{(\sqrt{\frac{1-\delta/2}{1-\delta}}-1)^2}\eta_t^3L^2+\frac{\eta_t^2L\sigma^2}{2K},
\end{aligned} \tag{73}$$

where (a) follows Lemma A.2. Let $\eta_t = \frac{c}{t+1}$, we have

$$\begin{aligned}&\mathbb{E}[F_{\mathcal{S}}(y_T) - F_{\mathcal{S}}(y^*)]\\ &\leq \mathbb{E}[F_{\mathcal{S}}(y_t) - F_{\mathcal{S}}(y^*)] \underbrace{\prod_{t=0}^{T-1}(1-\frac{\mu c}{2(t+1)})}_{①} + [\frac{B_1\sigma^2 + 2(1-\lambda)^2M^2}{(\sqrt{\frac{1-\delta/2}{1-\delta}}-1)^2} + \frac{\sigma^2}{2K}]L \underbrace{\sum_{t=0}^{T-1}\frac{c^2}{(t+1)^2}\prod_{t'=t+1}^{T-1}(1-\frac{\mu c}{2(t+1)})}_{②},\end{aligned} \tag{74}$$

where

$$① \leq \prod_{t=0}^{T-1}\exp(-\frac{\mu c}{2(t+1)}) = \exp(-\frac{\mu c}{2})\exp(-\frac{\mu c}{2}\sum_{t=1}^{T-1}\frac{1}{t+1}) \leq \exp(-\frac{\mu c}{2})\exp(-\frac{\mu c}{2}\log T) = \exp(-\frac{\mu c}{2})T^{-\frac{\mu c}{2}}, \tag{75}$$

$$② \leq \sum_{t=0}^{T-1}\frac{c^2}{(t+1)^2}\exp(-\frac{\mu c}{2}\sum_{t'=t+1}^{T-1}\frac{1}{t'+1}) \leq \sum_{t=0}^{T-1}\frac{c^2}{(t+1)^2}\exp(-\frac{\mu c}{2}\log\frac{T}{t+1}) = c^2T^{-\frac{\mu c}{2}}\sum_{t=0}^{T-1}(t+1)^{\frac{\mu c}{2}-2}. \tag{76}$$

When $\frac{\mu c}{2} - 2 \geq 0$,

$$② \leq \frac{c^2T^{-\frac{\mu c}{2}}}{\frac{\mu c}{2}-1}((T+1)^{\frac{\mu c}{2}-1}-1) \leq \frac{c^2}{\frac{\mu c}{2}-1}(\frac{T+1}{T})^{\frac{\mu c}{2}-1}T^{-1}. \tag{77}$$

When $\frac{\mu c}{2} - 2 \leq 0$ and $\frac{\mu c}{2} - 2 \neq -1$,

$$② \leq \frac{c^2T^{-\frac{\mu c}{2}}}{\frac{\mu c}{2}-1}(1+T^{\frac{\mu c}{2}-1}-1) = \frac{c^2}{\frac{\mu c}{2}-1}T^{-1}. \tag{78}$$

When $\frac{\mu c}{2} - 2 = -1$,

$$② \leq c^2T^{-1}(1+\log T). \tag{79}$$

Hence,

$$\mathbb{E}[F_{\mathcal{S}}(y_T) - F_{\mathcal{S}}(y^*)] = \widetilde{\mathcal{O}}(T^{-\frac{\mu c}{2}} + T^{-1}). \tag{80}$$

B.4. Optimization Error of DEF-A

In this section, we consider non-convex objective functions satisfying the Polyak-Łojasiewicz (PL) condition, which establishes the relation between the objective function and the gradient norm. The compressor we consider here is random sparsification which satisfies Assumptions 4.2 and 3.3. We need Assumptions 3.1, 3.2, 4.1, and $\eta_t = \frac{c}{t+1} \leq \frac{1}{8L}$.

$$\begin{aligned}&\mathbb{E}[F_{\mathcal{S}}(x_{t+1}) - F_{\mathcal{S}}(x_t)] \leq \mathbb{E}\langle\nabla F_{\mathcal{S}}(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\mathbb{E}\|x_{t+1} - x_t\|_2^2\\ &\overset{(a)}{=} \underbrace{\mathbb{E}\langle\nabla F_{\mathcal{S}}(x_t), -\mathcal{C}(\eta_t g_t + e_t)\rangle}_{①} + \frac{L}{2}\underbrace{\mathbb{E}\|\mathcal{C}(\eta_t g_t + e_t)\|_2^2}_{②},\end{aligned} \tag{81}$$

where (a) follows Assumption 3.3. For term ①,

$$
\begin{aligned}
① &\overset{(a)}{=} -\mathbb{E}\langle \nabla F_{\mathcal{S}}(x_t), \delta(\eta_t g_t + e_t)\rangle = -\delta\eta_t \mathbb{E}\langle \nabla F_{\mathcal{S}}(x_t), \frac{1}{K}\sum_{k=1}^{K} \nabla F_{\mathcal{S}}(x_t - \lambda e_{k,t}) + \frac{e_t}{\eta_t}\rangle \\
&= -\delta\eta_t \mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 - \delta\eta_t \mathbb{E}\langle \nabla F_{\mathcal{S}}(x_t), \frac{1}{K}\sum_{k=1}^{K} \nabla F_{\mathcal{S}}(x_t - \lambda e_{k,t}) + \frac{e_t}{\eta_t} - \nabla F_{\mathcal{S}}(x_t)\rangle \\
&\le -\frac{\delta\eta_t}{2}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{\delta\eta_t}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K} \nabla F_{\mathcal{S}}(x_t - \lambda e_{k,t}) - \nabla F_{\mathcal{S}}(x_t) + \frac{e_t}{\eta_t}\|_2^2 \\
&\le -\frac{\delta\eta_t}{2}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{\delta\eta_t L^2}{2K}\sum_{k=1}^{K}\mathbb{E}\|\lambda e_{k,t}\|_2^2 + \frac{\delta}{2\eta_t}\mathbb{E}\|e_t\|_2^2 \\
&\le -\frac{\delta\eta_t}{2}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{\delta(\eta_t^2 L^2\lambda^2 + 1)}{2\eta_t K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}\|_2^2 ,
\end{aligned} \tag{82}
$$

where (a) is due to the random sparsification compressor. For term ②,

$$
\begin{aligned}
② &\overset{(a)}{=} \delta\mathbb{E}\|\eta_t g_t + e_t\|_2^2 \le 2\delta\eta_t^2\mathbb{E}\|g_t\|_2^2 + 2\delta\mathbb{E}\|e_t\|_2^2 \\
&= 2\delta\eta_t^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F_{\mathcal{S}}(x_t - \lambda e_{k,t})\|_2^2 + \frac{2\delta\eta_t^2\sigma^2}{K} + 2\delta\mathbb{E}\|e_t\|_2^2 \\
&\le \frac{2\delta\eta_t^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t - \lambda e_{k,t}) - \nabla F_{\mathcal{S}}(x_t) + \nabla F_{\mathcal{S}}(x_t)\|_2^2 + 2\delta\mathbb{E}\|e_t\|_2^2 + \frac{2\delta\eta_t^2\sigma^2}{K} \\
&\le 4\delta\eta_t^2\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{4\delta\eta_t^2 L^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\lambda e_{k,t}\|_2^2 + 2\delta\mathbb{E}\|e_t\|_2^2 + \frac{2\delta\eta_t^2\sigma^2}{K} \\
&= 4\delta\eta_t^2\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{2\delta(2\eta_t^2 L^2\lambda^2 + 1)}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}\|_2^2 + \frac{2\delta\eta_t^2\sigma^2}{K} ,
\end{aligned} \tag{83}
$$

where (a) is due to the random sparsification compressor. Put them together, let $\eta_t \le \frac{1}{8L}$, and we have

$$
\begin{aligned}
&\mathbb{E}[F_{\mathcal{S}}(x_{t+1}) - F_{\mathcal{S}}(x_t)] \\
&\le -\frac{\delta\eta_t}{2}(1 - 4\eta_t L)\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{\delta(1 + 2\eta_t L)(1 + 2\eta_t^2 L^2\lambda^2)}{2\eta_t K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}\|_2^2 + \frac{\delta\eta_t^2 L\sigma^2}{K} \\
&\le -\frac{\delta\eta_t}{4}\mathbb{E}\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 + \frac{\delta(1+\lambda^2)}{\eta_t}\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|e_{k,t}\|_2^2 + \frac{\delta\eta_t^2 L\sigma^2}{K} \\
&\overset{(a)}{\le} -\frac{\mu\delta\eta_t}{2}\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x^*)] + \frac{\delta(1+\lambda^2)\eta_t(\sigma^2 + M^2)}{(1/\sqrt{1-\delta} - 1)^2} + \frac{\delta\eta_t^2 L\sigma^2}{K} ,
\end{aligned} \tag{84}
$$

where (a) is due to $\|\nabla F_{\mathcal{S}}(x_t)\|_2^2 \ge 2\mu(F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x^*))$ according to the PL condition and Lemma A.3. Rearrange,

$$
\mathbb{E}[F_{\mathcal{S}}(x_{t+1}) - F_{\mathcal{S}}(x^*)] \le (1 - \frac{\mu\delta\eta_t}{2})\mathbb{E}[F_{\mathcal{S}}(x_t) - F_{\mathcal{S}}(x^*)] + \frac{\delta(1+\lambda^2)\eta_t(\sigma^2 + M^2)}{(1/\sqrt{1-\delta} - 1)^2} + \frac{\delta\eta_t^2 L\sigma^2}{K} . \tag{85}
$$

Let $\eta_t = \frac{c}{t+1}$ and $G = \sqrt{\sigma^2 + M^2}$. Take this recurrence from $t = 0$ to $T-1$ and we have

$$\mathbb{E}[F_{\mathcal{S}}(x_T) - F_{\mathcal{S}}(x^*)] \leq \mathbb{E}[F_{\mathcal{S}}(x_0) - F_{\mathcal{S}}(x^*)] \underbrace{\prod_{t=0}^{T-1}(1 - \frac{\mu\delta c}{2(t+1)})}_{③}$$
$$+ \frac{\delta(1+\lambda^2)G^2}{(1/\sqrt{1-\delta}-1)^2} \underbrace{\sum_{t'=0}^{T-1} \frac{c}{t'+1} \underbrace{\prod_{t=t'+1}^{T-1}(1 - \frac{\mu\delta c}{2(t+1)})}_{④}}_{⑤} + \frac{\delta L\sigma^2}{K} \underbrace{\sum_{t'=0}^{T-1} \frac{c^2}{(t'+1)^2} \underbrace{\prod_{t=t'+1}^{T-1}(1 - \frac{\mu\delta c}{2(t+1)})}_{④}}_{⑥}, \tag{86}$$

where

$$\begin{aligned} ③ &\leq \prod_{t=0}^{T-1} \exp(-\frac{\mu\delta c}{2(t+1)}) = \exp(-\frac{\mu\delta c}{2}\sum_{t=0}^{T-1}\frac{1}{t+1}) \leq \exp(-\frac{\mu\delta c}{2})\exp(-\frac{\mu\delta c}{2}\sum_{t=1}^{T-1}\frac{1}{t+1}) \\ &\leq \exp(-\frac{\mu\delta c}{2})\exp(-\frac{\mu\delta c}{2}\log(T)) = \exp(-\frac{\mu\delta c}{2})T^{-\frac{\mu\delta c}{2}}. \end{aligned} \tag{87}$$

$$④ \leq \prod_{t=t'+1}^{T-1} \exp(-\frac{\mu\delta c}{2(t+1)}) = \exp(-\frac{\mu\delta c}{2}\sum_{t=t'+1}^{T-1}\frac{1}{t+1}) \leq \exp(-\frac{\mu\delta c}{2}\log(\frac{T}{t'+1})) = (\frac{T}{t'+1})^{-\frac{\mu\delta c}{2}}, \tag{88}$$

When $\frac{\mu\delta c}{2} \geq 1$,

$$⑤ \leq cT^{-\frac{\mu\delta c}{2}}\sum_{t'=0}^{T-1}(t'+1)^{\frac{\mu\delta c}{2}-1} \leq cT^{-\frac{\mu\delta c}{2}} \cdot \frac{2}{\mu\delta c}[(T+1)^{\frac{\mu\delta c}{2}} - 1] \leq \frac{2}{\mu\delta}(\frac{T+1}{T})^{\frac{\mu\delta c}{2}}. \tag{89}$$

When $0 < \frac{\mu\delta c}{2} \leq 1$,

$$⑤ \leq cT^{-\frac{\mu\delta c}{2}}\sum_{t'=0}^{T-1}(t'+1)^{\frac{\mu\delta c}{2}-1} \leq cT^{-\frac{\mu\delta c}{2}} \cdot [1 + \frac{2}{\mu\delta c}(T^{\frac{\mu\delta c}{2}} - 1)] \leq \frac{2}{\mu\delta}. \tag{90}$$

When $\frac{\mu\delta c}{2} \geq 2$,

$$⑥ \leq c^2T^{-\frac{\mu\delta c}{2}}\sum_{t'=0}^{T-1}(t'+1)^{\frac{\mu\delta c}{2}-2} \leq c^2T^{-\frac{\mu\delta c}{2}} \cdot \frac{1}{\mu\delta c/2-1}[(T+1)^{\frac{\mu\delta c}{2}-1} - 1] \leq \frac{c^2}{\mu\delta c/2-1}(\frac{T+1}{T})^{\frac{\mu\delta c}{2}-1}T^{-1}. \tag{91}$$

When $\frac{\mu\delta c}{2} \leq 2$, $\frac{\mu\delta c}{2} \neq 1$,

$$⑥ \leq c^2T^{-\frac{\mu\delta c}{2}}\sum_{t'=0}^{T-1}(t'+1)^{\frac{\mu\delta c}{2}-2} \leq c^2T^{-\frac{\mu\delta c}{2}} \cdot [1 + \frac{1}{\mu\delta c/2-1}(T^{\frac{\mu\delta c}{2}-1} - 1)] \leq \frac{c^2}{\mu\delta c/2-1}T^{-1}. \tag{92}$$

When $\frac{\mu\delta c}{2} = 1$,

$$⑥ \leq c^2T^{-1}\sum_{t'=0}^{T-1}(t'+1)^{-1} \leq c^2T^{-1} \cdot (1 + \log T). \tag{93}$$

Hence,

$$\mathbb{E}[F_{\mathcal{S}}(x_T) - F_{\mathcal{S}}(x^*)] = \widetilde{\mathcal{O}}(T^{-\frac{\mu\delta c}{2}} + T^{-1} + (1/\sqrt{1-\delta}-1)^{-2}). \tag{94}$$

C. Proof of Generalization of SGD-(IA) (Theorem 4.10)

For consistency, let $\{y_t\}$ be the SGD solution path, i.e.,

$$y_{t+1} = y_t - \eta_t \nabla f(y_t; \xi_t) = y_t - \eta_t g_t \,. \tag{95}$$

Let $\{x_t\}$ be the SGD-IA solution path with $x_0 = y_0$, where IA denotes momentum iterative averaging, i.e.,

$$x_{t+1} = (1-\delta)x_t + \delta y_{t+1} = x_t + \delta(y_{t+1} - x_t) \text{ and } x_0 = y_0 \,, \tag{96}$$

where $0 < 1 - \delta < 1$ is the momentum constant. Then,

$$x_t = (1-\delta)^t x_0 + \sum_{t'=1}^{t} \delta(1-\delta)^{t-t'} y_{t'} = (1-\delta)^t y_0 + \sum_{t'=1}^{t} \delta(1-\delta)^{t-t'} y_{t'} \,. \tag{97}$$

Let $K = 1$ and $\lambda = 1$, then DEF is identical to SGD. Following Lemma C.1, SGD-IA is a special case of DEF-A with $\mathcal{C}(\Delta) = \delta\Delta$. Note that this compressor does not compress the message volume.

Lemma C.1. *Let $g_t = \nabla f(y_t; \xi_t)$, $x_0 = y_0$, and $\mathcal{C}(-\Delta) = \mathcal{C}(\Delta)$. If $y_{t+1} = y_t - \eta_t g_t$ and $x_{t+1} = x_t + \mathcal{C}(y_{t+1} - x_t)$, then the update rule of x_t is identical to DEF-A when $K = 1$, $\lambda = 1$ with compressor $\mathcal{C}$, i.e.,*

$$\begin{aligned} x_{t+1} &= x_t - \mathcal{C}(\eta_t g_t + e_t) \,, \\ e_{t+1} &= \eta_t g_t + e_t - \mathcal{C}(\eta_t g_t + e_t),\ e_0 = 0 \\ y_{t+1} &= y_t - \eta_t g_t \,. \end{aligned} \tag{98}$$

Proof. We just need to verify $x_{t+1} = x_t + \mathcal{C}(y_{t+1} - x_t)$ with the 3 equations above. We have

$$x_{t+1} = x_0 - \sum_{t'=0}^{t} \mathcal{C}(\eta_{t'} g_{t'} + e_{t'}),\ y_{t+1} = y_0 - \sum_{t'=0}^{t} \eta_t g_{t'} \,. \tag{99}$$

Then

$$\begin{aligned} x_{t+1} - e_{t+1} &= x_0 - e_{t+1} - \sum_{t'=0}^{t} \mathcal{C}(\eta_{t'} g_{t'} + e_{t'}) \\ &= x_0 - e_{t+1} - \mathcal{C}(\eta_t g_t + e_t) - \sum_{t'=0}^{t-1} \mathcal{C}(\eta_{t'} g_{t'} + e_{t'}) \\ &= x_0 - \eta_t g_t - e_t - \sum_{t'=0}^{t-1} \mathcal{C}(\eta_{t'} g_{t'} + e_{t'}) \\ &= \cdots = x_0 - \sum_{t'=0}^{t} \eta_t g_t = y_{t+1} \,, \end{aligned} \tag{100}$$

$$\begin{aligned} x_t + \mathcal{C}(y_{t+1} - x_t) &= x_t + \mathcal{C}(x_{t+1} - e_{t+1} - x_t) \\ &= x_t + \mathcal{C}(-\mathcal{C}(\eta_t g_t + e_t) - e_{t+1}) \\ &= x_t + \mathcal{C}(-\eta_t g_t - e_t) = x_{t+1} \,, \end{aligned} \tag{101}$$

which completes the proof. □

C.1. Generalization Error of SGD

In this section, we need Assumptions 3.1, 3.2, 4.1, and $\eta_t \le \frac{c}{t+1}$. Following Sec. B.1 with $K = 1$ and $\lambda = 1$, we have

$$\mathbb{E}[f(y_T; \xi) - f(\widetilde{y}_T; \xi)] = \mathcal{O}(T^{(1-\frac{1}{N})Lc/((1-\frac{1}{N})Lc+1)}) \,. \tag{102}$$

C.2. Generalization Error of SGD-IA

In this section, we need Assumptions 3.1, 3.2, 4.1, and $\eta_t \leq \frac{c}{t+1}$. Following Sec. B.2 with $K = 1$, $\lambda = 1$ and the compressor replaced with $\mathcal{C}(\Delta) = \delta\Delta$ which satisfies Assumptions 4.2 and 3.3, we have $\delta^{\frac{1}{2}} \to \delta$ in Eq. (64). All the other procedures are the same, thus

$$\mathbb{E}[f(x_T;\xi) - f(\widetilde{x}_t;\xi)] = \mathcal{O}(T^{(1-\frac{1}{N})\delta Lc/((1-\frac{1}{N})\delta Lc+1)}) \,. \tag{103}$$

C.3. Optimization Error of SGD

In this section, we need Assumptions 3.1, 3.2, 4.1, and $\eta_t = \frac{c}{t+1} \leq \frac{1}{4L}$. Following Sec. B.3 with $K = 1$ and $\lambda = 1$, we have

$$\mathbb{E}[F_{\mathcal{S}}(y_T) - F_{\mathcal{S}}(y^*)] = \widetilde{\mathcal{O}}(T^{-\frac{\mu c}{2}} + T^{-1}) \,. \tag{104}$$

C.4. Optimization Error of SGD-IA

In this section, we need Assumptions 3.1, 3.2, 4.1, and $\eta_t = \frac{c}{t+1} \leq \frac{1}{8\delta L}$. Following Sec. B.4 with $K = 1$, $\lambda = 1$ and the compressor replaced with $\mathcal{C}(\Delta) = \delta\Delta$ which satisfies Assumptions 4.2 and 3.3, we have the same bound as Eq. (82), but $② = \delta^2\mathbb{E}\|\eta_t g_t + e_t\|_2^2$ in Eq. (83), which leads to the need for $\eta_t \leq \frac{1}{8\delta L}$. All the other procedures are the same, therefore

$$\mathbb{E}[F_{\mathcal{S}}(x_T) - F_{\mathcal{S}}(x^*)] = \widetilde{\mathcal{O}}(T^{-\frac{\mu\delta c}{2}} + T^{-1} + (1/\sqrt{1-\delta} - 1)^{-2}) \,. \tag{105}$$