Analysis of Error Feedback in Federated Non-Convex Optimization with Biased Compression: Fast Convergence and Partial Participation

Xiaoyun Li, Ping Li
LinkedIn Ads
700 Bellevue Way NE, Bellevue, WA 98004, USA
{xiaoyli, pinli}@linkedin.com

Abstract

In practical federated learning (FL) systems, the communication cost between the clients and the central server can often be a bottleneck. In this paper, we focus on biased gradient compression in non-convex FL problems. In the classical distributed learning, the method of error feedback (EF) is a common technique to remedy the downsides of biased gradient compression, but the performance of EF in FL still lacks systematic investigation. In this work, we study a compressed FL scheme with error feedback, named Fed-EF, with two variants depending on the global model optimizer. While directly applying biased compression in FL leads to poor convergence, we show that Fed-EF is able to match the convergence rate of the full-precision FL counterpart with a linear speedup w.r.t. the number of clients. Experiments verify that Fed-EF achieves the same performance as the full-precision FL approach, at the substantially reduced communication cost.

Moreover, we develop a new analysis of the EF under partial participation (PP), an important scenario in FL. Under PP, the convergence rate of Fed-EF exhibits an extra slow-down factor due to a so-called "stale error compensation" effect, which is also justified in our experiments. Our results provide insights on a theoretical limitation of EF, and possible directions for improvements.

1. Introduction

Federated learning (FL) has seen numerous applications in, e.g., wireless communications, Internet of Things (IoT), sensor networks, financial fraud detection, input method editor (IME), advertising, online visual object detection, public health records (Hard et al., 2018; Yang et al., 2019a;b; Liu et al., 2020b; Rieke et al., 2020; Kairouz et al., 2021; Khan et al., 2021), etc. A centralized FL system includes multiple clients each with local data, and one central server that coordinates the training process. The goal of FL is for n clients to collaboratively learn a global model, parameterized by θ, such that

$$\theta^* = \arg\min_{\theta\in\mathbb{R}^d} f(\theta) := \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\theta), \qquad (1)$$

where $f_i(\theta) := \mathbb{E}_{D\sim\mathcal{D}_i}\big[F_i(\theta; D)\big]$ is an (in general) non-convex loss function for the i-th client w.r.t. the local data distribution $\mathcal{D}_i$. In a typical FL system design, in each training round, the server first broadcasts the model to the clients. Then, each client trains the model based on the local data, after which the updated local models are transmitted back to the server and aggregated (McMahan et al., 2017; Stich, 2019; Chen et al., 2020). The number of clients n can be tens/hundreds in some applications, e.g., *cross-silo* FL (Marfoq et al., 2020; Huang et al., 2021) where clients are companies/organizations. In other scenarios, n can be as large as millions or even billions, e.g., *cross-device* FL (Karimireddy et al., 2021; Khan et al., 2021), where clients can be personal/IoT devices.

There are three key features/challenges of FL:

- **Communication cost:** The model transmission cost can be a major challenge in FL systems with limited bandwidth (e.g., portable wireless devices), especially when training modern ML models with millions or billions of model parameters. Developing efficient methods for transmitting the model/gradient information in FL has gained growing research interest in wireless/satellite communication (Amiri and Gündüz, 2020; Niknam et al., 2020; Yang et al., 2020; Khan et al., 2021; Yang et al., 2021b; Chen et al., 2022a).
- **Data heterogeneity:** Unlike in the classic distributed training, the local data distribution in FL ($\mathcal{D}_i$ in (1)) can be different (non-iid), reflecting the practical scenarios where local data held by clients (e.g., app/website users) are highly personalized (Zhao et al., 2018; Kairouz et al.,

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2021; Li et al., 2022a). When multiple local training steps are taken, the local models becomes "biased" towards minimizing the local (instead of global) losses, which may hinder the global model to quickly converge to a good solution (Mohri et al., 2019; Li et al., 2020c; Hard et al., 2018; Li et al., 2020a;b; Yuan and Li, 2022).

- **Partial participation (PP):** Another practical issue of FL systems is the partial participation (PP) where the clients do not join training consistently, e.g., due to unstable connection or active selection (Li et al., 2020a).That is, only a fraction of clients are involved in each FL training round to update their local models and send the local update information. This may also slow down the convergence of the global model, intuitively because less data/information is used per round (Charles et al., 2021; Cho et al., 2022).

FL with compression. To reduce the communication cost, FL algorithms with compressed message passing have been proposed. Examples include FedPaQ (Reisizadeh et al., 2020), FedCOM (Haddadpour et al., 2021), FedZip (Malekijoo et al., 2021), etc. These algorithms are built upon directly compressing model updates communicated from clients to server. In particular, Reisizadeh et al. (2020); Haddadpour et al. (2021) applied unbiased stochastic compressors such as stochastic quantization (Alistarh et al., 2017) and stochastic sparsification (Wangni et al., 2018), and showed that with significantly communication saving (e.g., $>$ 30 - 100x compression rate), applying unbiased compression in FL could approach the performance of uncompressed FL algorithms in many learning tasks.

Error feedback (EF). Biased compression is also commonly used in communication-efficient distributed training (Lin et al., 2018; Beznosikov et al., 2020). Common biased compressors include fixed quantization (Dettmers, 2016), TopK sparsification (Aji and Heafield, 2017; Alistarh et al., 2018; Stich et al., 2018), SignSGD (Bernstein et al., 2018; 2019), etc. For these compressors, the output is a biased estimator of the true gradient. In distributed optimization literature, it has been observed that directly updating with the biased gradients may slow down the convergence, lead to worse generalisation performance, or even diverge (Seide et al., 2014; Karimireddy et al., 2019; Beznosikov et al., 2020; Malekijoo et al., 2021). A popular remedy is the *error feedback (EF)* strategy (Seide et al., 2014; Stich et al., 2018): in each iteration, the local worker sends a compressed gradient to the server and records the local compression error, which is used to adjust the gradient computed in next iteration, conceptually "correcting the bias" due to compression. With EF, using biased compression in distributed training can achieve the same theoretical convergence rate and empirical performance as the full-precision counterpart (Karimireddy et al., 2019; Li et al., 2022b).

Our contributions. Despite the rich literature on EF in classic distributed optimization, EF has not been fully explored in the context of federated learning. In this work, we provide a thorough analysis of EF in FL. The three key features of FL—local steps, data heterogeneity and partial participation, pose challenging questions regarding the performance of EF in federated learning: *(i) Can EF achieve the same linear speedup rate as full-precision FL algorithms, under highly non-iid local data distribution? (ii) How does partial participation affect the convergence of EF?*

- We study Fed-EF, a communication-efficient FL framework with biased compression and error feedback (EF), with two variants (Fed-EF-SGD and Fed-EF-AMS) depending on the global optimizer (SGD and adaptive AMSGrad (Reddi et al., 2018), respectively). Our investigation starts by providing a convergence analysis showing the non-convergence issue of directly applying biased compression in FL. Then, we develop sharp convergence analysis and prove that Fed-EF has asymptotic convergence rate $\mathcal{O}(\frac{1}{\sqrt{TKn}})$ when $T = \mathcal{O}(Kn)$, where T is the number of communication rounds, K is the number of local steps and n is the number of clients. Our result improves the communication complexity of the prior result on error compensated FL and matches the best-known rates of many full-communication and compressed FL methods (see Table 1 for a summary). Moreover, Fed-EF-AMS is the first compressed adaptive FL method in the literature.

- Error feedback has not been studied in the literature of distributed optimization under partial client participation. We initiate a new analysis of Fed-EF in this setting, considering both local steps and non-iid data situations. We show that under PP, Fed-EF exhibits a slow-down factor of $\sqrt{n/m}$ compared with the best full-precision rate, where m is the number of active clients per round. We argue that this potential theoretical limitation is caused by an intrinsic *"delayed error compensation"* effect of EF.

- Experiments are conducted to illustrate the effectiveness of Fed-EF. We show that Fed-EF matches the performance of full-precision FL with a significant reduction in communication cost, and the proposed method compares favorably with FL algorithms using unbiased compression without EF. Numerical examples are also provided to justify our theory.

2. Background and Related Work

Gradient compression methods. In distributed SGD training systems, extensive works have applied various compression techniques to the communicated gradients. For example, the so-called "unbiased compressors" are commonly

Table 1. Summary of theoretical convergence results from some existing works on distributed and federated learning with communication compression for non-convex optimization. "PP" stands for "partial participation", and "# of Rounds" is the number of communication rounds required to achieve linear speedup, which is a common measure of the communication complexity of FL algorithms. T is the number of communication rounds, K is the number of local steps, and n is the total number of clients.

Reference	Local Step	Non-iid Data	PP	Adaptive Opt.	Compression	# of Rounds
Jiang and Agrawal (2018) [a]		✓			Unbiased	$T = \mathcal{O}(n)$
Li et al. (2022b)		✓		✓	Biased + EF	$T = \mathcal{O}(n^3)$
Reisizadeh et al. (2020)	✓				Unbiased	– [b]
Haddadpour et al. (2021)	✓	✓			Unbiased	$T = \mathcal{O}(Kn)$
Basu et al. (2019)	✓				Biased + EF	$T = \mathcal{O}(K^3n^3)$
Gao et al. (2021)	✓				Biased + EF	$T = \mathcal{O}(Kn^3)$
Fed-EF (our paper) [c]	✓	✓	✓	✓	Biased + EF	$\boldsymbol{T = \mathcal{O}(Kn)}$

[a] The authors also provided analysis with local steps, but the result of $T = \mathcal{O}(K^2n^5)$ is worse than Haddadpour et al. (2021).
[b] This cell is invalid because the convergence rate does not have a linear speedup in n.
[c] The result in the table is under full participation (Corollary 4.9). The analysis for partial participation is provided in Section 4.3.

used, which include the stochastic rounding and QSGD (Alistarh et al., 2017; Zhang et al., 2017; Wu et al., 2018; Liu et al., 2020a; Xu et al., 2021), count-sketch (Charikar et al., 2004; Ivkin et al., 2019; Haddadpour et al., 2020; Li and Li, 2023), and the magnitude based random sparsification (Wangni et al., 2018). In an extreme case with highest compression rate, Seide et al. (2014); Bernstein et al. (2018; 2019); Jin et al. (2020) proposed to use only the stochastic sign (1-bit) information of the gradients. On the other hand, the "biased compressors" are also popular. Representative examples include the TopK compressor (Alistarh et al., 2018; Stich et al., 2018; Shi et al., 2019) (which only transmits gradient coordinates with largest magnitudes), fixed (or learned) quantization (Dettmers, 2016; Zhang et al., 2017; Yu et al., 2018; Malekijoo et al., 2021), and low-rank approximation (Vogels et al., 2019). See Beznosikov et al. (2020) for more discussion on biased compressors. Our work assumes a fairly general condition on the compressor which applies to a wide range of compression schemes.

Error feedback (EF). It has been shown that directly adopting biased compression in distributed SGD may lead to divergence, through empirical observations or counter examples (Seide et al., 2014; Karimireddy et al., 2019; Beznosikov et al., 2020). The error feedback (EF) method is proposed to fix this issue (Seide et al., 2014; Stich et al., 2018; Karimireddy et al., 2019). In particular, with EF, distributed SGD under biased compression can match the convergence rate of the full-precision distributed SGD (Alistarh et al., 2018; Jiang and Agrawal, 2018; Shen et al., 2018; Stich and Karimireddy, 2019; Zheng et al., 2019). Li et al. (2022b) showed that EF is also effective for distributed training with adaptive gradient methods. Among the limited related literature on applying EF to FL, the most relevant method is the QSparse-local-SGD (Basu et al., 2019), which is a special instance of the proposed Fed-EF-SGD method. The analysis of Basu et al. (2019) did not consider data heterogeneity and partial client participation, and their convergence rate is slower than our analysis. Gao et al. (2021) analyzed momentum SGD under the same setting and improved the communication complexity of Basu et al. (2019). Table 1 provides more details. Recently, Richtárik et al. (2021); Fatkhullin et al. (2021) proposed "EF21" as an alternative to the standard EF. Our work is different in that we study the standard EF (which is a different algorithm from EF21) and our analytical setup is more practical with local steps, non-iid data and partial participation.

Distributed adaptive gradient methods. Our proposed Fed-EF algorithm, in addition to SGD, also includes a variant with the base optimizer being AMSGrad (Reddi et al., 2018), an adaptive gradient method that is widely used in optimization and distributed learning (Chen et al., 2020; Karimi et al., 2021; Reddi et al., 2021; Li et al., 2022b; Chen et al., 2022b; Zhao et al., 2022). Adaptive gradient algorithms assign different implicit learning rates to different coordinates adaptively guided by the training trajectory, usually leading to faster convergence, less effort needed for parameter tuning, and better performance than SGD on some tasks (Zhang et al., 2020; Liu et al., 2021). Readers are referred to the extensive literature on adaptive gradient methods for more details, e.g., Duchi et al. (2011); Zeiler (2012); Kingma and Ba (2015); Chen et al. (2019); Zhou et al. (2020); Reddi et al. (2021); Wang et al. (2021).

3. Compressed FL with Error Feedback

3.1. Biased Compressors in Federated Learning

We introduce some existing and new compressors used in our paper. Let $[n]$ be the integer set $\{1, ..., n\}$. Let $\|\cdot\|$ denote the l_2 norm and $\|\cdot\|_1$ denote the l_1 norm.

Definition 3.1 ($q_{\mathcal{C}}$-deviate compressor)**.** *The biased $q_{\mathcal{C}}$-deviate compressor $\mathcal{C} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is defined such that for $\forall x \in \mathbb{R}^d$, $\exists\ 0 \leq q_{\mathcal{C}} < 1$ s.t. $\|\mathcal{C}(x) - x\|^2 \leq q_{\mathcal{C}}^2 \|x\|^2$. Two examples are Stich et al. (2018); Zheng et al. (2019):*

- *Let $\mathcal{S} = \{i \in [d] : |x_i| \geq t\}$ where t is the $(1-k)$-quantile of $|x_i|$, $i \in [d]$. The **TopK** compressor with compression rate k is defined as $\mathcal{C}(x)_i = x_i$, if $i \in \mathcal{S}$; $\mathcal{C}(x)_i = 0$ otherwise.*
- *Divide $[d]$ into M groups (e.g., neural network layers) with index sets $\mathcal{I}_i$, $i = 1, ..., M$, and $d_i := |\mathcal{I}_i|$. The **(Group) Sign** compressor is defined as $\mathcal{C}(x) = \left[\frac{\|x_{\mathcal{I}_1}\|_1}{d_1} sign(x_{\mathcal{I}_1}), ..., \frac{\|x_{\mathcal{I}_M}\|_1}{d_M} sign(x_{\mathcal{I}_M})\right]$, with $x_{\mathcal{I}_i}$ the sub-vector of x at indices $\mathcal{I}_i$.*

Larger $q_{\mathcal{C}}$ indicates heavier compression, and $q_{\mathcal{C}} = 0$ implies no compression, i.e. $\mathcal{C}(x) = x$. Additionally, these two compression operators can be combined to derive the so-called "**heavy-Sign**" compressor, where we first apply **TopK** and then **Sign**, for even higher compression rate. This is similar in spirit to the Sparse Binary Compression (Sattler et al., 2019) but with group-wise magnitude adjustment.

Definition 3.2 (Heavy-Sign)**.** *Let $\mathcal{C}_k(\cdot)$ and $\mathcal{C}_s(\cdot)$ be the **TopK** and **Sign** in Definition 3.1. Then the **Heavy-Sign** operator is defined as $\mathcal{C}_{hv}(x) := \mathcal{C}_s\big(\mathcal{C}_k(x)\big)$ for $x \in \mathbb{R}^d$.*

We can show that this compressor is also q-deviate.

Proposition 3.3. *The **heavy-Sign** compressor satisfies Definition 3.1 with $q_{\mathcal{C}}^2 = 1 - \min_{i\in[M]} \frac{k}{d_i}$.*

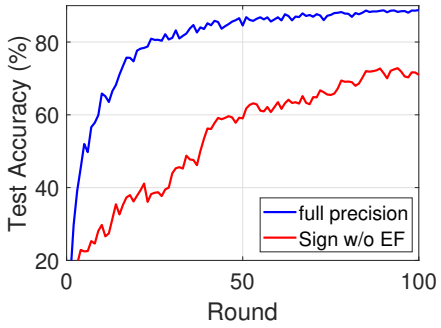

Figure 1. Test accuracy of MLP trained by Fed-SGD on MNIST dataset (LeCun et al., 1998): full-precision vs. **Sign** compressed communication, $\eta = 1$, $\eta_l = 0.01$, $n = 200$ non-iid clients.

Can we directly use biased compressors in FL? As an example, in Figure 1, we report the test accuracy of a multi-layer perceptron (MLP) trained on MNIST dataset in non-iid FL environment (see Section 5 for more descriptions), of Fed-SGD (Stich, 2019) with full communication (blue) versus compression using **Sign** (red), i.e., clients directly send compressed local model update to the server for aggregation (see Appendix B for details). We see a catastrophic performance downgrade of using biased compression directly. In Section 3, we will show that naively adopting biased compression leads to an undesirable asymptotically non-vanishing term in the convergence rate, theoretically justifying this empirical performance degradation.

Algorithm 1 Fed-EF: Compressed FL with Error Feedback

1: **Input**: learning rates η, η_l; parameters β_1, β_2, ϵ
2: **Initialize**: global model $\theta_1 \in \mathbb{R}^d \subseteq \mathbb{R}^d$; local error accumulator $e_{1,i} = \mathbf{0}$; $m_0 = \mathbf{0}$, $v_0 = \mathbf{0}$, $\hat{v}_0 = \mathbf{0}$
3: **for** $t = 1, \ldots, T$ **do**
4: **parallel for worker** $i \in [n]$ **do**:
5: Receive global model θ_t from server, set $\theta_{t,i}^{(1)} = \theta_t$
6: **for** $k = 1, \ldots, K$ **do**
7: Compute stochastic gradient $g_{t,i}^{(k)}$ at $\theta_{t,i}^{(k)}$
8: Local update $\theta_{t,i}^{(k+1)} = \theta_{t,i}^{(k)} - \eta_l g_{t,i}^{(k)}$
9: **end for**
10: Compute local update $\Delta_{t,i} = \theta_{t,i}^{(K+1)} - \theta_t$
11: Send $\widetilde{\Delta}_{t,i} = \mathcal{C}(\Delta_{t,i} + e_{t,i})$ to server
12: Update the error $e_{t+1,i} = e_{t,i} + \Delta_{t,i} - \widetilde{\Delta}_{t,i}$
13: **end parallel**
14: **Central server do:**
15: Global aggregation $\overline{\widetilde{\Delta}}_t = \frac{1}{n}\sum_{i=1}^n \widetilde{\Delta}_{t,i}$
16: Global update $\theta_{t+1} = \theta_t - \eta\overline{\widetilde{\Delta}}_t$ { Fed-EF-SGD }
17: $m_t = \beta_1 m_{t-1} + (1-\beta_1)\overline{\widetilde{\Delta}}_t$ { Fed-EF-AMS }
18: $v_t = \beta_2 v_{t-1} + (1-\beta_2)\overline{\widetilde{\Delta}}_t^2$, $\hat{v}_t = \max(v_t, \hat{v}_{t-1})$
19: Global update $\theta_{t+1} = \theta_t - \eta\frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$
20: **end for**

3.2. Fed-EF Algorithm

In Algorithm 1, we present the general compressed FL framework named Fed-EF. Error feedback (EF) is a popular tool in distributed training and can be adapted to federated learning. In round t: 1) The server broadcast the global model θ_t to all clients (line 5); 2) The i-th client performs K steps of local SGD updates to get local model $\theta_{t,i}^{(K)}$, compute the compressed local model update $\tilde{\Delta}_{t,i}$, updates the local error accumulator $e_{t,i}$, and sends the compressed $\tilde{\Delta}_{t,i}$ back to the server (line 6-12); 3) The server receives $\tilde{\Delta}_{t,i}$, $i \in [n]$ from all clients, takes the average, and perform a global model update using the averaged compressed local model updates (line 15-19).

Depending on the global optimizer, we propose two variants: Fed-EF-SGD (green) which applies SGD global updates, and Fed-EF-AMS (blue), whose global optimizer is AMSGrad (Reddi et al., 2018). In Fed-EF-AMS, we incorporate momentum (m_t) with different implicit dimension-wise

learning rates $\eta/\hat{v}_t$. For conciseness, the presented algorithm employs one-way compression (clients-to-server). In Appendix G, we provide Fed-EF with two-way compression and show that adding the server-to-clients compression does not affect the asymptotic convergence rates.

Algorithmic comparison with related work. Compared with EF in the classical distributed training, e.g., Stich et al. (2018); Karimireddy et al. (2019); Zheng et al. (2019); Liu et al. (2020a); Ghosh et al. (2021); Li et al. (2022b), our algorithm allows local steps and uses two-side learning rates. Note that when $\eta \equiv 1$, the Fed-EF-SGD method reduces to QSparse-local-SGD (Basu et al., 2019). In Section 4, we will demonstrate how the two-side learning rate schedule improves the convergence analysis of Basu et al. (2019). In addition, several recent works considered compressed FL using unbiased stochastic compressors. FedPaQ (Reisizadeh et al., 2020) applied stochastic quantization without error feedback to local SGD, which is improved by Haddadpour et al. (2021) using a gradient tracking trick that, however, requires communicating an extra vector from server to clients. Malekijoo et al. (2021) provided an empirical study on directly compressing the local updates without using EF. Mitra et al. (2021) proposed FedLin, which only uses compression for synchronizing a local memory term but still requires transmitting full-precision updates. It is also worth mentioning that, all the aforementioned works only studied the SGD prototype. To our knowledge, Fed-EF-AMS is the first compressed adaptive FL method in the literature[1].

4. Theoretical Results

Assumption 4.1 (Smoothness)**.** *For $\forall i \in [n]$, f_i is L-smooth: $\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|$, $\forall x, y \in \mathbb{R}^d$.*

Assumption 4.2 (Bounded variance)**.** *For $\forall t \in [T]$, $\forall i \in [n]$, $\forall k \in [K]$: (i) the stochastic gradient is unbiased: $\mathbb{E}\big[g_{t,i}^{(k)}\big] = \nabla f_i(\theta_{t,i}^{(k)})$; (ii) the **local variance** is bounded: $\mathbb{E}\big[\|g_{t,i}^{(k)} - \nabla f_i(\theta_{t,i}^{(k)})\|^2\big] < \sigma^2$; (iii) the **global variance** is bounded: $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\theta_t) - \nabla f(\theta_t)\|^2 \le \sigma_g^2$.*

Both assumptions are standard in the convergence analysis of stochastic gradient methods. The global variance bound σ_g^2 in Assumption 4.2 characterizes the difference among local loss functions, i.e., data heterogeneity.

Assumption 4.3 (Compression discrepancy)**.** *There exists some constant $q_{\mathcal{A}} < 1$, such that $\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n \mathcal{C}\big(\Delta_{t,i} + e_{t,i}\big) - \frac{1}{n}\sum_{i=1}^n(\Delta_{t,i} + e_{t,i})\|^2\big] \le q_{\mathcal{A}}^2 \mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n(\Delta_{t,i} + e_{t,i})\|^2\big]$ in every round $t \in [T]$.*

Assumption 4.3 is a common assumption in related work on compressed distributed learning. If we replace "the average of compression", $\frac{1}{n}\sum_{i=1}^n \mathcal{C}\big(\Delta_{t,i} + e_{t,i}\big)$, by "the compression of average", $\mathcal{C}\big(\frac{1}{n}\sum_{i=1}^n(\Delta_{t,i} + e_{t,i})\big)$, the statement immediately holds by Definition 3.1 with $q_{\mathcal{A}} = q_{\mathcal{C}}$. Thus, Assumption 4.3 basically says that the above two terms stay close during training. A similar assumption is used in Alistarh et al. (2018) for analyzing sparsified SGD. In Haddadpour et al. (2021), to achieve sharp convergence rate for FL with unbiased compression, a similar condition is also assumed with an absolute bound. See Appendix E for more discussion and justification.

Norm convergence. We bound the squared gradient norm of a uniformly sampled global model from $\theta_1, ..., \theta_T$: $\Psi = \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]$ which is a standard convergence measure in non-convex optimization.

4.1. Warm-up: Biased Compression Without EF

In Figure 1, we see that naively transmitting the condensed message by biased compressors performs poorly. The precise steps of this approach are summarized in Algorithm 2 in Appendix B. In each round t, after conducting local training to get the model update $\Delta_{t,i}$, the i-th client computes $\widetilde{\Delta}_{t,i} = \mathcal{C}(\Delta_{t,i})$ and sends it to the server. The server makes a global SGD update by the averaging all the compressed gradients. In Beznosikov et al. (2020), the authors provided toy counter-examples showing that solely using biased compressors may lead to divergence under a simple distributed learning setting with constant learning rate. In the following, we develop a general convergence analysis under our more complex FL setup, and present the result with specific learning rates. The proof can be found in Appendix F.5.

Theorem 4.4 (Fed-SGD with biased compression)**.** *Let $\theta^* = \arg\min f(\theta)$, and denote $q = \max\{q_{\mathcal{C}}, q_{\mathcal{A}}\}$. Consider Fed-SGD with biased compression (Algorithm 2). Under Assumptions 4.1 to 4.3, when $\eta_l \le \frac{1}{8KL\max\{1, 8(1+q^2)\eta\}}$, choosing $\eta_l = \Theta(\frac{1}{K\sqrt{T}})$ and $\eta = \Theta(\sqrt{Kn})$, we have*

$$\Psi = \mathcal{O}\Big(\frac{1+q^2}{\sqrt{TKn}} + \frac{1+q^2}{TK}(\sigma^2 + K\sigma_g^2) + \frac{q^2\sigma^2}{Kn}\Big).$$

Remark 4.5. *If unbiased compressors are used instead, we can remove the bias term in our analysis and recover the $\mathcal{O}(\frac{1}{\sqrt{TKn}})$ convergence rate in Haddadpour et al. (2021) for FL with unbiased compression.*

In Theorem 4.4, we see that larger q (i.e., higher compression rate) would slow down the convergence, as expected. The first term depends on the model initialization and the local variance. The second term containing σ_g^2 represents the influence of data heterogeneity. The last non-vanishing term (as $T \to \infty$) is a consequence of the gradient estimation bias, which shows that Fed-SGD does not converge to a stationary point when biased compression is directly applied without EF. Also, we note that this constant is independent

[1] We sincerely thank an anonymous reviewer and the Chair for suggesting us to discuss the results in Wang et al. (2022). As requested, we provide our comments in Appendix.

of the choice of the learning rates. This bias term decreases with smaller q (i.e., less compression); when $q = 0$, we recover the sharp rate of full-precision Fed-SGD.

4.2. Fed-EF: Linear Speedup under Data Heterogeneity

We now analyze our Fed-EF algorithm which compensates the compression bias with error feedback (EF).

Theorem 4.6 (Fed-EF-SGD). *Let $\theta^* = \arg\min f(\theta)$, and denote $\triangle = f(\theta_1) - f(\theta^*)$, $q = \max\{q_{\mathcal{C}}, q_{\mathcal{A}}\}$, $C_1 := 2 + \frac{4q^2}{(1-q^2)^2}$. Under Assumptions 4.1 to 4.3, when $\eta_l \le \frac{1}{2KL\cdot\max\{4,\eta(C_1+1)\}}$, the squared gradient norm of Fed-EF-SGD iterates in Algorithm 1 satisfies*

$$\Psi \lesssim \frac{\triangle}{\eta\eta_l TK} + \frac{2\eta\eta_l C_1 L}{n}\sigma^2 + 10\eta\eta_l^3 C_1 K^2 L^3(\sigma^2 + 6K\sigma_g^2).$$

In our analysis of Fed-EF-AMS, we make the following additional assumption of bounded stochastic gradients, which is common in the convergence analysis of adaptive gradient methods, e.g., Reddi et al. (2018); Zhou et al. (2018); Chen et al. (2019); Reddi et al. (2021); Li et al. (2022b), among others. Note that this assumption is only used for Fed-EF-AMS, but not for Fed-EF-SGD.

Assumption 4.7 (Bounded gradients). *It holds that $\|g_{t,i}^{(k)}\| \le G$, $\forall t > 0$, $\forall i \in [n]$, $\forall k \in [K]$.*

We provide the convergence analysis of adaptive FL with error feedback (Fed-EF-AMS) as below.

Theorem 4.8 (Fed-EF-AMS). *With same notations as in Theorem 4.6, let $C_1 := \frac{\beta_1}{1-\beta_1} + \frac{2q}{1-q^2}$. Under Assumptions 4.1 to 4.7, if the learning rates satisfy $\eta_l \le \frac{\sqrt{\epsilon}}{8KL}\min\{\frac{1}{\sqrt{\epsilon}}, \frac{2(1-q^2)L}{(1+q^2)^{1.5}G}, \frac{1}{\max\{16,32C_1^2\}\eta}, \frac{1}{3\eta^{1/3}}\}$, the Fed-EF-AMS iterates in Algorithm 1 satisfy*

$$\Psi \lesssim \frac{\triangle}{\eta\eta_l TK} + \frac{\eta\eta_l L(6+4C_1^2)}{n\epsilon}\sigma^2 + \frac{2C_1G^2d}{T\sqrt{\epsilon}} + \frac{3\eta\eta_l C_1^2 LKG^2 d}{T\epsilon} + \Big[\frac{5\eta_l^2 KL^2}{2\sqrt{\epsilon}} + \frac{\eta\eta_l^3(30+20C_1^2)K^2L^3}{\epsilon}\Big](\sigma^2 + 6K\sigma_g^2).$$

With some properly chosen learning rates, we have the following simplified results.

Corollary 4.9 (Fed-EF, specific learning rates). *Suppose the conditions in Theorem 4.6 and Theorem 4.8 are satisfied, respectively. Choosing $\eta_l = \Theta(\frac{1}{K\sqrt{T}})$ and $\eta = \Theta(\sqrt{Kn})$, Fed-EF-SGD satisfies*

$$\Psi = \mathcal{O}\Big(\frac{\triangle}{\sqrt{TKn}} + \frac{1}{\sqrt{TKn}}\sigma^2 + \frac{\sqrt{n}}{T^{3/2}\sqrt{K}}(\sigma^2 + K\sigma_g^2)\Big),$$

and for Fed-EF-AMS, it holds that

$$\Psi = \mathcal{O}\Big(\frac{\triangle}{\sqrt{TKn}} + \frac{1}{\sqrt{TKn}}\sigma^2 + (\frac{1}{TK} + \frac{\sqrt{n}}{T^{\frac{3}{2}}\sqrt{K}})(\sigma^2 + K\sigma_g^2)\Big).$$

Discussion. From Corollary 4.9, when $T \ge \mathcal{O}(K)$, Fed-EF-AMS and Fed-EF-SGD have the same rate of convergence asymptotically. Therefore, our following discussion applies to the general Fed-EF scheme with both variants. In Corollary 4.9, when $T \ge Kn\frac{\sigma_g^2}{\sigma^2}$, the global variance term σ_g^2 vanishes and the convergence rate becomes $\mathcal{O}(1/\sqrt{TKn})$ [2]. Thus, the proposed Fed-EF enjoys a linear speedup w.r.t. the number of clients n, i.e., it reaches a δ-stationary point (i.e., $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \le \delta$) as long as $TK = \Theta(1/n\delta^2)$, which matches the recent results of the full-precision counterparts (Yang et al., 2021a; Reddi et al., 2021) (Note that Reddi et al. (2021) only analyzed the special case $\beta_1 = 0$, while our analysis is more general). Moreover, our result is also better than $\mathcal{O}(K^3n^3)$ of the federated momentum SGD analysis in Yu et al. (2019). By setting $K = \Theta(1/n\delta)$, Fed-EF only requires $T = \Theta(1/\delta)$ rounds of communication to converge. This matches one of the state-of-the-art FL communication complexity results of SCAFFOLD (Karimireddy et al., 2020).

Comparison with prior compressed FL results. Table 1 provides a summary of the related results of compressed FL in non-convex optimization. As a special case of Fed-EF-SGD ($\eta \equiv 1$), the analysis of QSparse-local-SGD (Basu et al., 2019) did not consider data heterogeneity, and their communication complexity to achieve linear speedup is $T = \mathcal{O}(K^3n^3)$, which is much worse than our $\mathcal{O}(Kn)$ result. They also imposed a bounded gradient assumption (for the SGD analysis) while our Theorem 4.6 does not. Gao et al. (2021) studied compressed local momentum SGD and proved the communication complexity $T = \mathcal{O}(Kn^3)$, but also limited to the homogeneous data setting and is worse than our rates. For FL with direct unbiased compression (without EF), the convergence rate of FedPaQ (Reisizadeh et al., 2020) is $\mathcal{O}(1/\sqrt{TK})$ which did not achieve linear speedup. Recently, Haddadpour et al. (2021) refined the analysis and algorithm of FedPaQ, and their result for unbiased compression matches our $\mathcal{O}(1/\delta)$ communication complexity when taking $K = \Theta(1/n\delta)$.

To conclude, our results provide the sharp linear speedup analysis of EF in FL (with local steps and data heterogeneity), and show that EF + FL can match the convergence rate of many state-of-the-art rates of uncompressed FL methods and FL methods with unbiased compression. Moreover, to our knowledge, Theorem 4.8 is the first result on compressed adaptive federated learning in the literature.

[2] This also implies that $K = \mathcal{O}(\frac{T\sigma^2}{n\sigma_g^2})$ is needed to achieve the $\mathcal{O}(\frac{1}{\sqrt{TKn}})$ asymptotic convergence rate which decreases in K (in other words, when more local steps help). The inverse relationship between K and σ_g^2 (the global variance) is because when σ_g^2 is large, the local losses might be very different from the global loss. Therefore, applying more local steps may not help with the global convergence in this case.

4.3. Analysis of Fed-EF under Partial Participation

Whilst being a popular strategy in classical distributed training, error feedback has not been analyzed under partial participation (PP), which is an important feature of FL. Next, we provide new analysis and results of EF under this setting, also considering both local steps and data heterogeneity. In each round t, assume only m randomly chosen clients (without replacement) indexed by $\mathcal{M}_t \subseteq [n]$ are active and participate in training (i.e., changing $i \in [n]$ to $i \in \mathcal{M}_t$ at line 4 of Algorithm 1). For the remaining $(n-m)$ inactive clients, nothing is changed and we simply set $e_{t,i} = e_{t-1,i}$, $\forall i \in [n] \setminus \mathcal{M}_t$. The convergence rate of Fed-EF under PP is given as below.

Theorem 4.10 (Fed-EF, partial participation)**.** *In each round, suppose m randomly chosen clients in $\mathcal{M}_t$ participate in the training. Denote $\triangle = f(\theta_1) - f(\theta^*)$. Under Assumptions 4.1 to 4.3, suppose the learning rates satisfy $\eta_l \leq \min\left\{\frac{1}{6}, \frac{m}{96C'\eta}, \frac{m^2}{53760(n-m)C_1\eta}, \frac{1}{4\eta}, \frac{1}{32C_1\eta}\right\}\frac{1}{KL}$. Then, Fed-EF-SGD admits*

$$\Psi \lesssim \frac{\triangle}{\eta\eta_l TK} + \Big[\frac{\eta\eta_l L}{m} + \frac{8\eta\eta_l C_1 Ln}{m^2}\Big]\sigma^2 + \frac{3\eta\eta_l C'KL}{m}\sigma_g^2 + \Big[\frac{5\eta_l^2 KL^2}{2} + \frac{15\eta\eta_l^3 C'K^2L^3}{m} + \frac{560\eta\eta_l C_1(n-m)L}{m^2}\Big](\sigma^2 + 6K\sigma_g^2)$$

with constants $C_1 = \frac{q^2}{(1-q^2)^3}$ and $C' = \frac{n-m}{n-1}$. Choosing $\eta = \Theta(\sqrt{Km})$, $\eta_l = \Theta(\frac{\sqrt{m}}{K\sqrt{Tn}})$, we have

$$\Psi = \mathcal{O}\Big(\frac{\sqrt{n}}{\sqrt{m}}\Big(\frac{f(\theta_1) - f(\theta^*)}{\sqrt{TKm}} + \frac{1}{\sqrt{TKm}}\sigma^2 + \frac{\sqrt{K}}{\sqrt{Tm}}\sigma_g^2\Big)\Big).$$

Remark 4.11. *We present Fed-EF-SGD for simplicity. With more complicated analysis, the same asymptotic convergence rate also applies to Fed-EF-AMS.*

Remark 4.12. *When $m = n$, Theorem 4.10 reduces to the $\mathcal{O}(1/\sqrt{TKn})$ rate in Corollary 4.9. When $q = 0$ (no compression), we can recover the $\mathcal{O}(\sqrt{K/Tm})$ rate of full-precision Fed-SGD under PP (Yang et al., 2021a).*

The convergence rate in Theorem 4.10 contains m in the denominator, instead of n as in Corollary 4.9, which is a result of larger gradient estimation variance due to client sampling. Importantly, compared with the $\mathcal{O}(\sqrt{K/Tm})$ rate of the full-precision local SGD under partial participation, we extract an additional slow-down factor of $\sqrt{n/m}$ of Fed-EF under partial participation. Hence, Theorem 4.10 points out a potential theoretical limitation of EF that has not been demonstrated before.

Effect of delayed error compensation. We argue that this slow-down factor is a consequence of the mechanism of error feedback as a "stateful" compression scheme. Intuitively, EF itself can, to a large extent, be regarded as subtly "delaying" the "untransmitted" gradient information $(\mathcal{C}(\Delta_t) - \Delta_t)$ to the next iteration. However, under partial participation, in round t, the error accumulator of a chosen client actually contains the latest information from round $t - s$, where s can be viewed as the "lag" which follows a geometric distribution with $\mathbb{E}[s] = n/m$. In some sense, this "delayed error compensation" effect shares similar spirit to asynchronous distributed optimization with delayed gradients (e.g., Agarwal and Duchi (2011); Lian et al. (2015)). The delayed error information in Fed-EF under PP is likely to pull the model away from heading towards a stationary point (i.e., slower down the convergence), especially for highly non-convex loss functions. In Section 5, we propose a simple strategy called "error restarting" to empirically justify (and to an extent mitigate) the negative impact of the stale error compensation on the norm convergence.

5. Experiments

We provide empirical results to show the efficacy of Fed-EF and justify our theoretical convergence results. More implementation details and results are placed in Appendix D.

Datasets. We conduct experiments on three popular FL datasets. The MNIST dataset (LeCun et al., 1998) contains 60000 training examples and 10000 test samples of 28×28 gray-scale hand-written digits from 0 to 9. The FMNIST dataset (Xiao et al., 2017) has the same input size and train/test split as MNIST, but the samples are fashion products (e.g., clothes and bags). We describe the CIFAR dataset and its ResNet experiments in Appendix D.

Training setup. We test $n = 200$ clients. The clients' local data are set to be highly non-iid (heterogeneous). Precisely, we split the data into $2n = 400$ shards where each shard only contain samples from one class. Then, we randomly assign each client to shards of data, so that the local data samples of each client contains at most two classes. We run $T = 100$ rounds, where one FL training round is finished after all the clients perform one epoch of local training. The local mini-batch size is 32, implying 10 local iterations per round per client. We uniformly randomly sample $m = 20, 100$ clients in each round (i.e., participation rate $p = 0.1, 0.5$ respectively). For all the methods, we tune the global and local learning rates η, η_l over a fine grid. The reported results are averaged over 10 independent runs.

Methods and compressors. We compare the following algorithms in our experiments (see Algorithm 2 in Appendix B for details of the competing methods):

- **SGD-full:** Fed-SGD with full-precision communication and two-sided learning rates (Yang et al., 2021a).
- **SGD-Stoc:** Fed-SGD with unbiased stochastic quantization (QSGD) (Alistarh et al., 2017) **without error feedback**, see (2) for the details. This algorithm is

equivalent to FedPaQ/FedCOM (Reisizadeh et al., 2020; Haddadpour et al., 2021). For this compressor, we test parameter $b \in \{1, 2, 4\}$.

- **SGD-sign, SGD-topk, SGD-hv-sign:** our proposed Fed-EF-SGD method with three compressors, respectively. For **TopK**, we test compression level $k \in \{0.001, 0.01, 0.05\}$. For **heavy-Sign** compressor (Definition 3.2) where **Sign** is applied after **TopK** (i.e., a further 32x compression over **TopK** under same sparsity), we test $k \in \{0.01, 0.05, 0.1\}$. Note that the Qsparse-local-sgd (Basu et al., 2019) method is also included in our experiments as a special case of Fed-EF when the global learning rate $\eta = 1$.
- **AMS-full:** full-precision adaptive federated learning with two-sided learning rates (Reddi et al., 2021).
- **AMS-Stoc:** unbiased stochastic quantization applied to the client-to-server communication in AMS-full.
- **AMS-sign, AMS-optk, AMS-hv-sign:** the proposed Fed-EF-AMS with AMSGrad as the global optimizer and three compressors respectively.

5.1. Fed-EF Matches Full-Precision FL with Substantially Less Communication

Firstly, we provide a general evaluation of Fed-EF in practical FL tasks. We train a ReLU activated CNN with two convolutional layers followed by one max-pooling, one dropout and two fully-connected layers before the softmax output. We test each compression strategy with different compression ratios. To measure the communication cost, we report the accumulated number of bits transmitted from the client to server (averaged over all clients), assuming that full-precision gradients are 32-bit encoded. In Figure 2 (participation rate $p = 0.5$) and Figure 3 ($p = 0.1$), we present the test accuracy chosen as follows: for each method, we present the curve with highest compression rate that achieves the best full-precision test accuracy (within 0.1%); if the method does not match the full-precision performance, we present the curve with the highest test accuracy. The accuracy tables with standard deviations and full results with all compression rates can be found in Appendix D. From the experiments, we observe the following:

- Both variants of Fed-EF achieve the same performance as the full-precision methods with substantial communication reduction, e.g., **heavy-Sign** and **TopK** reduce the communication by more than 100x without losing accuracy; **Sign** also provides 30x compression with matching accuracy as full-precision training.
- In Figure 2, on MNIST, the test accuracy of **Stoc** (without EF) is slightly lower than Fed-EF-SGD with **heavy-Sign**, yet requiring more communication.

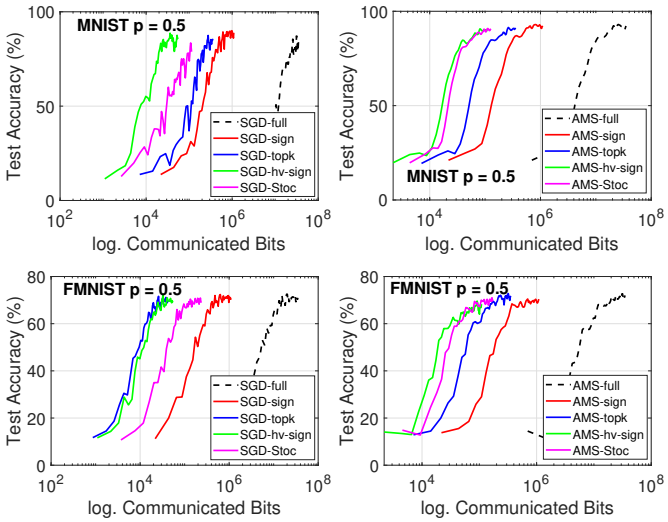

Figure 2. Test accuracy vs. communicated bits on MNIST and FMNIST, participation rate $p = 0.5$. "sign", "topk" and "hv-sign" are with Fed-EF; "Stoc" is stochastic quantization without EF.

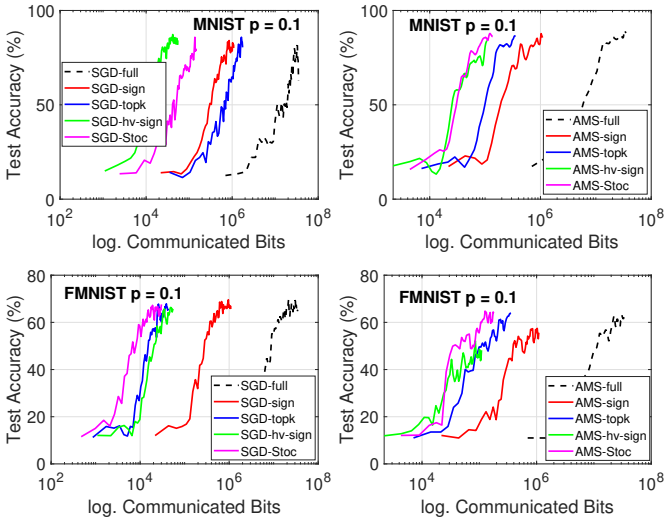

Figure 3. Training loss and test accuracy vs. communicated bits on MNIST and FMNIST datasets, participation rate $p = 0.1$.

- With more aggressive participation rate $p = 0.1$ (Figure 3), Fed-EF with a proper compressor still performs on a par with uncompressed methods. We notice that fixed sign-based compressors (**Sign** and **heavy-Sign**) are outperformed by **TopK** for Fed-EF-AMS on FMNIST. We conjecture that this is because when p is small, sign-based compressors tend to assign equal implicit learning rates across coordinates, making adaptive method less effective. In contrast, magnitude-preserving compressors (e.g., **TopK** and **Stoc**) may better exploit the adaptivity of AMSGrad.

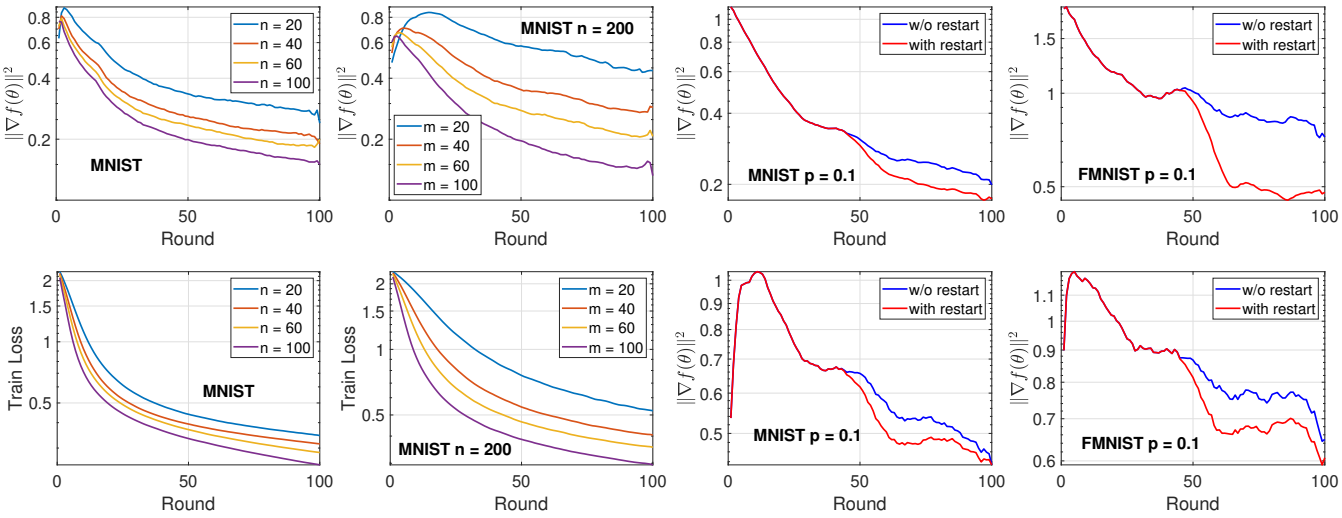

Figure 4. MLP on MNIST with **TopK-0.01** Fed-EF: Squared gradient norm and train loss. 1st column: full participation with increasing n. 2nd column: partial participation with increasing m.

Figure 5. Squared gradient norm of Fed-EF (**TopK-0.01**) under PP with error restarting, $S = 10$. 1st row: logistic regression. 2nd row: MLP. $n = 200$, $\eta = 1$, $\eta_l = 0.1$.

5.2. Linear Speedup and Delayed Error Compensation

We evaluate the norm convergence to verify the linear speedup of Fed-EF and the effect of delayed error compensation in partial participation (PP). From Theorem 4.9, under full participation, reaching a δ-stationary point requires running $\Theta(1/n\delta^2)$ rounds of Fed-EF. When n is fixed, Theorem 4.10 implies that under PP the speedup should be super-linear against m, the number of active clients. In other words, altering m under PP should have more impact on the convergence than altering n with full participation.

We train an MLP (which is also used for Figure 1) with one hidden layer with 200 neurons. In Figure 4, we report the squared gradient norm and the training loss on MNIST under the same non-iid setting above (the results on FMNIST are similar). In the full participation case, we implement Fed-EF-SGD with $n = 20, 40, 60, 100$; for PP, we fix $n = 200$ and alter $m = 20, 40, 60, 100$. According to our theory, we set $\eta = 0.1\sqrt{n}$ (or $0.1\sqrt{m}$) and $\eta_l = 0.1$. We see that: 1) the convergence of Fed-EF is faster with increasing n and m, which confirms the speedup property; 2) The gaps among curves in the PP setting are larger than those with the full-participation, suggesting that the acceleration brought by increasing m under PP is more significant than that of increasing n in full-participation, which is consistent with the implications of theoretical rates.

Error restarting. To further embody the intuitive impact of delayed error compensation under PP, we test a simple strategy called *"error restarting"* as follows: for each client i in round t, if the error accumulator was last updated more than S rounds ago, we simply restart its error accumulator by setting $e_{t,i} = 0$, which effectively eliminates the error information that is "too old". In Figure 5, we report the squared gradient norm for training logistic regression and MLP with non-iid data. We first run Fed-EF for 50 rounds, and then trigger error restarting with $S = 10$. As we see, after the 50-th round, the gradient norm with error restarting is smaller than that of the standard Fed-EF. These results illustrate the impact of the stale error compensation of Fed-EF, and suggests that properly handling this staleness might be a promising direction for improvement in the future.

6. Conclusion

We study Fed-EF, a federated learning (FL) framework with biased communication compression and error feedback (EF). We consider two variants, Fed-EF-SGD and Fed-EF-AMS, where the global optimizers are the SGD and the adaptive AMSGrad method, respectively. Theoretically, we demonstrate the non-convergence issue of directly using biased compression in FL, and provide new analysis showing that Fed-EF is able to achieve linear speedup in convergence rate as state-of-the-art full-precision FL algorithms. In the literature, the proposed Fed-EF-AMS variant is the first compressed adaptive federated learning method.

Moreover, we develop new analysis of EF in distributed optimization under partial participation (PP). Our result reveals an additional slow-down factor related to the participation rate due to the delayed error compensation of the EF mechanism. Experiments validate that Fed-EF achieves a significant communication reduction without performance drop. In summary, our work provides a thorough theoretical investigation of the error feedback technique in federated learning, and analyzes its convergence under practical FL settings. Our new analysis of partial participation also reveals a potential theoretical limitation of EF.

References

Alekh Agarwal and John C. Duchi. Distributed delayed stochastic optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 873–881, Granada, Spain, 2011.

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 440–445, Copenhagen, Denmark, 2017.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1709–1720, Long Beach, CA, 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5973–5983, Montréal, Canada, 2018.

Mohammad Mohammadi Amiri and Deniz Gündüz. Federated learning over wireless fading channels. *IEEE Trans. Wirel. Commun.*, 19(5):3546–3557, 2020.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas N. Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14668–14679, Vancouver, Canada, 2019.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SIGNSGD: compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 559–568, Stockholmsmässan, Stockholm, Sweden, 2018.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.

Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*, 2020.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyian, and Virginia Smith. On large-cohort training for federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 20461–20475, virtual, 2021.

Hao Chen, Ming Xiao, and Zhibo Pang. Satellite-based computing networks with federated learning. *IEEE Wirel. Commun.*, 29(1):78–84, 2022a.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of A class of adam-type algorithms for non-convex optimization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.

Xiangyi Chen, Xiaoyun Li, and Ping Li. Toward communication efficient adaptive gradient method. In *Proceedings of the ACM-IMS Foundations of Data Science Conference (FODS)*, pages 119–128, Virtual Event, USA, 2020.

Xiangyi Chen, Belhal Karimi, Weijie Zhao, and Ping Li. On the convergence of decentralized adaptive gradient methods. In *Proceedings of The 14th Asian Conference on Machine Learning (ACML)*, Hyderabad, India, 2022b.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 10351–10375, Virtual Event, 2022.

Tim Dettmers. 8-bit approximations for parallelism in deep learning. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016.

John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, 2011.

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

Hongchang Gao, An Xu, and Heng Huang. On the convergence of communication-efficient local SGD for federated learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 7510–7518, Virtual Event, 2021.

Avishek Ghosh, Raj Kumar Maity, Swanand Kadhe, Arya Mazumdar, and Kannan Ramchandran. Communication-efficient and byzantine-robust distributed learning with error feedback. *IEEE J. Sel. Areas Inf. Theory*, 2(3): 942–953, 2021.

Farzin Haddadpour, Belhal Karimi, Ping Li, and Xiaoyun Li. Fedsketch: Communication-efficient and private federated learning via sketching. *arXiv preprint arXiv:2008.04975*, 2020.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2350–2358, Virtual Event, 2021.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, 2016.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-IID data. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 7865–7873, Virtual Event, 2021.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman Arora. Communication-efficient distributed SGD with sketching. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13144–13154, Vancouver, Canada, 2019.

Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2530–2541, Montréal, Canada, 2018.

Richeng Jin, Yufan Huang, Xiaofan He, Huaiyu Dai, and Tianfu Wu. Stochastic-sign SGD for federated learning with theoretical guarantees. *arXiv preprint arXiv:2002.10940*, 2020.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2): 1–210, 2021.

Belhal Karimi, Ping Li, and Xiaoyun Li. Fed-LAMB: Layerwise and dimension-wise locally adaptive federated learning. *arXiv preprint arXiv:2110.00532*, 2021.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 3252–3261, Long Beach, CA, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 5132–5143, Virtual Event, 2020.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 28663–28676, virtual, 2021.

Latif U. Khan, Walid Saad, Zhu Han, Ekram Hossain, and Choong Seon Hong. Federated learning for internet of things: Recent advances, taxonomy, and open challenges. *IEEE Commun. Surv. Tutorials*, 23(3):1759–1799, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Ping Li and Xiaoyun Li. OPORP: One permutation + one random projection. *arXiv preprint arXiv:2302.03505*, 2023.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *Proceedings of the 38th IEEE International Conference on Data Engineering (ICDE)*, pages 965–978, Kuala Lumpur, Malaysia, 2022a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, Austin, TX, 2020b.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 2020c.

Xiaoyun Li, Belhal Karimi, and Ping Li. On distributed adaptive optimization with gradient compression. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, Virtual Event, 2022b.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2737–2745, Montreal, Canada, 2015.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for efficient distributed learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 133–143, Online [Palermo, Sicily, Italy], 2020a.

Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. FedVision: An online visual object detection platform powered by federated learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 13172–13179, New York, NY, 2020b.

Zechun Liu, Zhiqiang Shen, Shichao Li, Koen Helwegen, Dong Huang, and Kwang-Ting Cheng. How do adam and training strategies help BNNs optimization. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 6936–6946, Virtual Event, 2021.

Amirhossein Malekijoo, Mohammad Javad Fadaeieslam, Hanieh Malekijou, Morteza Homayounfar, Farshid Alizadeh-Shabdiz, and Reza Rawassizadeh. FedZIP: A compression framework for communication-efficient federated learning. *arXiv preprint arXiv:2102.01593*, 2021.

Othmane Marfoq, Chuan Xu, Giovanni Neglia, and Richard Vidal. Throughput-optimal topology design for cross-silo federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282, Fort Lauderdale, FL, 2017.

Aritra Mitra, Rayana H. Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14606–14619, virtual, 2021.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 4615–4625, Long Beach, CA, 2019.

Solmaz Niknam, Harpreet S. Dhillon, and Jeffrey H. Reed. Federated learning for wireless communications: Motivation, opportunities, and challenges. *IEEE Commun. Mag.*, 58(6):46–51, 2020.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2021–2031, Online [Palermo, Sicily, Italy], 2020.

Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4384–4396, virtual, 2021.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Sparse binary compression: Towards distributed deep learning with minimal communication. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, Budapest, Hungary, 2019. IEEE.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1058–1062, Singapore, 2014.

Zebang Shen, Aryan Mokhtari, Tengfei Zhou, Peilin Zhao, and Hui Qian. Towards more efficient stochastic decentralized learning: Faster convergence and sparse communication. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 4631–4640, Stockholmsmässan, Stockholm, Sweden, 2018.

Shaohuai Shi, Kaiyong Zhao, Qiang Wang, Zhenheng Tang, and Xiaowen Chu. A convergence analysis of distributed SGD with communication-efficient gradient sparsification. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3411–3417, Macao, China, 2019.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, 2019.

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4452–4463, Montréal, Canada, 2018.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14236–14245, Vancouver, Canada, 2019.

Jun-Kun Wang, Xiaoyun Li, Belhal Karimi, and Ping Li. An optimistic acceleration of AMSGrad for nonconvex optimization. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, pages 422–437, Virtual Event, 2021.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. In *Proceeding of the International Conference on Machine Learning (ICML)*, pages 22802–22838, Baltimore, MD, 2022. URL https://proceedings.mlr.press/v162/wang22o/wang22o.pdf.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1306–1316, Montréal, Canada, 2018.

Jiaxiang Wu, Weidong Huang, Junzhou Huang, and Tong Zhang. Error compensated quantized SGD and its applications to large-scale distributed optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5321–5329, Stockholmsmässan, Stockholm, Sweden, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zhiqiang Xu, Dong Li, Weijie Zhao, Xing Shen, Tianbo Huang, Xiaoyun Li, and Ping Li. Agile and accurate CTR prediction model training for massive-scale online advertising systems. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 2404–2409, Virtual Event, China, 2021.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021a.

Kai Yang, Tao Jiang, Yuanming Shi, and Zhi Ding. Federated learning via over-the-air computation. *IEEE Trans. Wirel. Commun.*, 19(3):2022–2035, 2020.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):12:1–12:19, 2019a.

Wensi Yang, Yuhang Zhang, Kejiang Ye, Li Li, and Cheng-Zhong Xu. FFD: A federated learning based method for credit card fraud detection. In *Proceedings of the 8th International Congress on Big Data*, pages 18–32, San Diego, CA, 2019b.

Zhaohui Yang, Mingzhe Chen, Kai-Kit Wong, H. Vincent Poor, and Shuguang Cui. Federated learning for 6G: Applications, challenges, and opportunities. *Engineering*, 2021b.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7184–7193, Long Beach, CA, 2019.

Mingchao Yu, Zhifeng Lin, Krishna Narra, Songze Li, Youjie Li, Nam Sung Kim, Alexander G. Schwing, Murali Annavaram, and Salman Avestimehr. GradiVeQ: Vector quantization for bandwidth-efficient gradient aggregation in distributed CNN training. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5129–5139, Montréal, Canada, 2018.

Xiaotong Yuan and Ping Li. On convergence of FedProx: Local dissimilarity invariant bounds, non-smoothness and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, 2022.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. ZipML: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 4035–4043, Sydney, Australia, 2017.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J. Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Weijie Zhao, Xuewu Jiao, Mingqing Hu, Xiaoyun Li, Xiangyu Zhang, and Ping Li. Communication-efficient terabyte-scale model training framework for online advertising. In *Proceedings of the IEEE International Conference on Big Data (IEEE BigData)*, Osaka, Japan, 2022.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-IID data. *arXiv preprint arXiv:1806.00582*, 2018.

Shuai Zheng, Ziyue Huang, and James T. Kwok. Communication-efficient distributed blockwise momentum SGD with error-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11446–11456, Vancouver, Canada, 2019.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Yiqi Tang, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.

Yingxue Zhou, Belhal Karimi, Jinxing Yu, Zhiqiang Xu, and Ping Li. Towards better generalization of adaptive gradient methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, virtual, 2020.

Appendix

Contents (Appendix)

A. Comments on Wang et al. (2022)

We sincerely thank an anonymous reviewer and the Chair for strongly suggesting us to discuss the results in Wang et al. (2022) https://proceedings.mlr.press/v162/wang22o/wang22o.pdf which also studied compressed FL with EF. Our paper and Wang et al. (2022) both applied the setup and techniques in Li et al. (2022b) on compressed distributed adaptive optimization. During the intensive discussion/rebuttal period, it became clear that Wang et al. (2022) had an error in the critical step of the proof of the main result and it is not obvious how to fix the error. Specifically, in their equation (C.15), the $\mathbb{E}[\|\triangle_t^i\|^2]$ term on the RHS of the first inequality cannot be bounded by their Lemma C.5, since Lemma C.5 is for the averaged term $\|\triangle_t\|^2$. Also, their equation (C.16) does not hold because the bound on $\mathbb{E}[\|e_t^i\|^2]$ is not the one in their Lemma C.3.

Wang et al. (2022) also included an analysis for EF in FL with partial participation in their appendix. Their rate, however, does not appear to be meaningful. Precisely, in their Theorem B.2, the convergence rate contains two dominant terms (of asymptotic order $\mathcal{O}(1/\sqrt{T})$) that sum up to $\mathcal{O}(\frac{1}{\eta\eta_l KT} + \eta\eta_l K)$. Hence, one term decreasing in m, the number of participating clients, would result in another term increasing in m. This means there is a region where their convergence rate decreases with larger m, which contradicts the common expectation that more active clients should improve the convergence. Also, their rate does not recover the full participation rate when $m = n$. In contrast, in our Theorem 4.10, the convergence rate of Fed-EF under PP decreases with larger m, and recovers the full participation result when $m = n$ and recovers the result for full-precision FL when $q = 0$.

We again highly appreciate the intensive and detailed technical discussion during the rebuttal phase. As suggested by the Reviewer, those discussions should be made available to the community.

B. Compressed FL Algorithms Without Error Feedback

Algorithm 2 A general framework for more algorithms considered in this paper

Input: learning rates η, η_l, hyper-parameters β_1, β_2, ϵ
Initialize: central server parameter $\theta_1 \in \mathbb{R}^d \subseteq \mathbb{R}^d$; $e_{1,i} = \mathbf{0}$ the accumulator for each worker; $m_0 = \mathbf{0}$, $v_0 = \mathbf{0}$, $\hat{v}_0 = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  parallel for worker $i \in [n]$ **do**:
    Receive model parameter θ_t from central server, set $\theta_{t,i}^{(1)} = \theta_t$
    for $k = 1, \ldots, K$ **do**
      Compute stochastic gradient $g_{t,i}^{(k)}$ at $\theta_{t,i}^{(k)}$
      Local update $\theta_{t,i}^{(k+1)} = \theta_{t,i}^{(k)} - \eta_l g_{t,i}^{(k)}$
    end for
    Compute the local model update $\Delta_{t,i} = \theta_{t,i}^{(K+1)} - \theta_t$
    Send quantized model update $\widetilde{\Delta}_{t,i} = \mathcal{Q}(\Delta_{t,i})$ to central server using (2)
  end parallel
  Central server do:
  Global aggregation $\overline{\widetilde{\Delta}}_t = \frac{1}{n}\sum_{i=1}^n \widetilde{\Delta}_{t,i}$
  Update the global model $\theta_{t+1} = \theta_t - \eta\overline{\widetilde{\Delta}}_t$ { **Stoc** or biased compression with SGD }
  $m_t = \beta_1 m_{t-1} + (1-\beta_1)\overline{\widetilde{\Delta}}_t$ { **Stoc** or biased compression with AMSGrad }
  $v_t = \beta_2 v_{t-1} + (1-\beta_2)\overline{\widetilde{\Delta}}_t^2, \quad \hat{v}_t = \max(v_t, \hat{v}_{t-1})$
  Update the global model $\theta_{t+1} = \theta_t - \eta\frac{m_t}{\sqrt{\hat{v}_t+\epsilon}}$
end for

For completeness, in Algorithm 2, we give the details of more algorithms, such as compressed FL without error feedback and FL with stochastic quantization. Similar to Fed-EF, we may also design two variants depending on the global optimizer.

When the compressor is biased. In Algorithm 2, if the compressor $\mathcal{Q}$ is biased (e.g., **Sign** and **TopK**), then the SGD variant of Algorithm 2 is essentially the algorithm considered in Theorem 4.4, the Fed-SGD method directly using biased compression without EF.

When the compressor is unbiased, Algorithm 2 becomes the **Stoc** baseline in our experiments, which directly compresses the transmitted vector from clients to server by unbiased stochastic quantization $\mathcal{Q}(\cdot)$ proposed by Alistarh et al. (2017). For a vector $x \in \mathbb{R}^d$, the operator $\mathcal{Q}(\cdot)$ is defined as

$$\textbf{QSGD (Stoc):} \qquad \mathcal{Q}_b(x) = \|x\| \cdot sign(x) \cdot \xi(x, b), \tag{2}$$

where $b \geq 1$ is number of bits per non-zero entry of the compressed vector $\mathcal{Q}(x)$. Suppose $0 \leq l < 2^{b-1}$ is the integer such that $|x_i|/\|x\|$ is contained in the interval $[l/2^{b-1}, (l+1)/2^{b-1}]$. The random variable $\xi(x, b)$ is defined by

$$\xi(x,b) = \begin{cases} l/s, & \text{with probability } 1 - g(\frac{|x_i|}{\|x\|}, b), \\ (l+1)/s, & \text{otherwise,} \end{cases}$$

with $g(a, b) = a \cdot 2^{b-1} - l$ for $a \in [0, 1]$. Simply, 0 is always quantized to 0. We can verify that the **Stoc** quantizer is unbiased, i.e., $\mathbb{E}[\mathcal{Q}(x)|x] = x$. In addition, it also introduces sparsity to the compressed vector in a probabilistic way, with $\mathbb{E}[\|\mathcal{Q}(x)\|_0] \leq 2^b + 2^{b-1}\sqrt{d}$.

Additionally, we mention that **Stoc** also has two corresponding variants, one using SGD and one using AMSGrad as the global optimizer. For the SGD variant, **Stoc** is equivalent to the FedCOM method in Haddadpour et al. (2021), which is also the FedPaQ algorithm (Reisizadeh et al., 2020) with tunable global learning rate.

When there is no compression, for the full-precision algorithms, we simply set $\widetilde{\Delta}_{t,i} = \Delta_{t,i}$ in line 11 of Algorithm 2. For SGD, it is the one studied in Yang et al. (2021a) which is the standard local SGD (McMahan et al., 2017) with global learning rate. For the AMSGrad variant, it becomes FedAdam (Reddi et al., 2021). Note that (Reddi et al., 2021) used Adam, while we use AMSGrad (with the max operation) for better stability. Empirically, the performance of these two options are very similar.

C. Analysis of Biased Compressors (Proof of Proposition 3.3)

The following Proposition C.1 is well-known, and we include the proof for clarity and completeness.

Proposition C.1. *For the **TopK** compressor which selects top k-percent of coordinates, we have $q_{\mathcal{C}}^2 = 1 - k$. For the **(Group) Sign** compressor, $q_{\mathcal{C}}^2 = 1 - \min_{i\in[M]} \frac{1}{d_i}$.*

Proof. For **TopK**, the proof is trivial: since $\mathcal{C}(x) - x$ only contain $(1-k)d$ coordinates with lowest magnitudes, we know $\|C(x) - x\|^2/\|x^2\| \leq 1 - k$.

For **Sign**, recall that $\mathcal{I}_i$ is the index set of block (group) i. By definition, for the i-th block (group) $x_{\mathcal{I}_i} \in \mathbb{R}^{d_i}$, we have

$$\begin{aligned}\|\mathcal{C}(x_{\mathcal{I}_i}) - x_{\mathcal{I}_i}\|^2 &= \|x_{\mathcal{I}_i} - \frac{\|x_{\mathcal{I}_i}\|_1}{d_i} sign(x_{\mathcal{I}_i})\|^2 \\ &= \|x_{\mathcal{I}_i}\|^2 + \frac{\|x_{\mathcal{I}_i}\|_1^2}{d_i^2} \cdot d_i - \frac{2\|x_{\mathcal{I}_i}\|_1^2}{d_i} \\ &= \|x_{\mathcal{I}_i}\|^2 - \|x_{\mathcal{I}_i}\|_1^2/d_i.\end{aligned}$$

Since we have M blocks, concatenating the blocks leads to

$$\begin{aligned}\|\mathcal{C}(x) - x\|^2 &= \sum_{i=1}^{M} \Big(\|x_{\mathcal{I}_i}\|^2 - \|x_{\mathcal{I}_i}\|_1^2/d_i\Big) \\ &= \|x\|^2 - \sum_{i=1}^{M} \|x_{\mathcal{I}_i}\|_1^2/d_i \\ &= \big(1 - \frac{\sum_{i=1}^{M} \|x_{\mathcal{I}_i}\|_1^2/d_i}{\|x\|^2}\big)\|x\|^2 \\ &\leq \big(1 - \min_{i\in[M]} \frac{\|x_{\mathcal{I}_i}\|_1^2}{d_i\|x_{\mathcal{I}_i}\|^2}\big)\|x\|^2 \leq (1 - \min_{i\in[M]} \frac{1}{d_i})\|x\|^2,\end{aligned}$$

where the last inequality is because l_1 norm is lower bounded by l_2 norm. □

Proposition 3.3 is re-stated here for convenience.

Proposition C.2. *The **heavy-Sign** compressor satisfies Definition 3.1 with $q_{\mathcal{C}}^2 = 1 - \min_{i\in[M]} \frac{k}{d_i}$.*

Proof. Recall Definition 3.2 that $\mathcal{C}_k$ denotes the **TopK** compressor and $\mathcal{C}_s$ is the **Sign** operator. The **heavy-Sign** operator $\mathcal{C}(x) = \mathcal{C}_s\big(\mathcal{C}_k(x)\big)$ admits

$$\begin{aligned}\|\mathcal{C}_{hv}(x) - x\|^2 &= \|\mathcal{C}_s\big(\mathcal{C}_k(x)\big) - \mathcal{C}_k(x) + \mathcal{C}_k(x) - x\|^2 \\ &= \|\mathcal{C}_s\big(\mathcal{C}_k(x)\big) - \mathcal{C}_k(x)\|^2 + \|\mathcal{C}_k(x) - x\|^2,\end{aligned}$$

where the second equality holds because **TopK** zeros out the unpicked coordinates. By Proposition C.1, we continue to obtain

$$\begin{aligned}\|\mathcal{C}_{hv}(x) - x\|^2 &\leq (1 - \min_{i\in[M]} \frac{1}{d_i})\|\mathcal{C}_k(x)\|^2 + \|\mathcal{C}_k(x) - x\|^2 \\ &= \|x\|^2 - \min_{i\in[M]} \frac{1}{d_i}\|\mathcal{C}_k(x)\|^2 \leq (1 - \min_{i\in[M]} \frac{k}{d_i})\|x\|^2,\end{aligned}$$

where M is the number of blocks in **Sign** and we use the fact that $\|\mathcal{C}_k(x)\|^2 + \|\mathcal{C}_k(x) - x\|^2 = \|x\|^2$, and $\|\mathcal{C}(x)\| \geq k\|x\|^2$ by Proposition C.1. □

D. Implementation Details and More Experiment Results

D.1. Implementation and Parameter Tuning

In our experiments, the clients' local data are set to be highly non-iid (heterogeneous), where the local data samples of each client contains at most two classes: we first split the data samples into $2n = 400$ shards each containing samples from only one class; then each client is assigned with two shards uniformly at random.

For **TopK**, in our implementation we also apply it in a "layer-wise" manner similar to **Sign**. Let k denote the proportion of coordinates selected. For each layer with d_i parameters, we pick $\max(1, \lfloor kd_i \rfloor)$ gradient dimensions. The maximum operator avoids the case where a layer is never updated.

We fine-tune the global and local learning rates for the baseline methods and Fed-EF with each hyper-parameter of the compressors (i.e., the compression rate). For the AMSGrad optimizer, we set $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ as the recommended default (Reddi et al., 2018). For each algorithm, we tune η over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 5, 10\}$ and η_l over $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We found that the compressed methods usually have same optimal learning rates as the full-precision training. The best learning rate combinations achieving highest test accuracy are given in Table 2.

	Fed-EF-SGD		Fed-EF-AMS	
	η	η_l	η	η_l
MNIST	10	10^{-3}	10^{-3}	10^{-2}
FMNIST	1	10^{-1}	10^{-2}	10^{-1}
CIFAR-10	1	10^{-1}	10^{-3}	10^{-2}

Table 2. Optimal global (η) and local (η_l) learning rate combinations to attain highest test accuracy.

D.2. Results on CIFAR and ResNet

The CIFAR-10 (Krizhevsky and Hinton, 2009) dataset includes 50000 natural images of size 32×32 each with RGB channels for training and 10000 images for testing. There are 10 classes, e.g., airplanes, cars, cats, etc. We follow a standard strategy for CIFAR-10 dataset to pre-process the training images by a random crop, a random horizontal flip and a normalization of the pixel values to have zero mean and unit variance. For images in the test set, we only apply the normalization step.

We present additional experiments to illustrate that Fed-EF is able to match the full-precision training on larger models, on the task of CIFAR-10 image classification. For this experiment, we train a ResNet-18 (He et al., 2016) network for 200 rounds. The clients local data are distributed in the same way as described above which is highly non-iid.

In Figure 6, we plot the test accuracy of Fed-EF with different compressors. Again, we see that Fed-EF (both variants) is able to attain the same accuracy level as the corresponding full-precision federated learning algorithms. For Fed-EF-SGD, the compression rate is around 32x for **Sign**, 100x for **TopK** and $\sim$300x for **heavy-Sign**. For Fed-EF-AMS, the compression ratio can also be around hundreds. Note that for Fed-EF-AMS, the training curve of **TopK-0.001** is not stable. Though it reaches a high accuracy, we still plot **TopK-0.01** in the third column for comparison.

In Figure 7 we report the results for partial participation with $p = 0.1$. Similarly, for SGD, all three compressors are able to match the full-precision accuracy, with significantly reduced number of communicated bits. For Fed-EF-AMS, similar to the observations on FMNIST, we see that **TopK** outperforms **Sign** and **heavy-Sign**, and achieve the performance of full-precision method with 100x compression ratio. **Sign** also performs reasonably well.

In conclusion, our results on CIFAR-10 and ResNet again confirm that compared with standard full-precision FL algorithms, the proposed Fed-EF scheme can provide significant communication reduction without empirical performance drop in test accuracy, under both data heterogeneity and partial client participation.

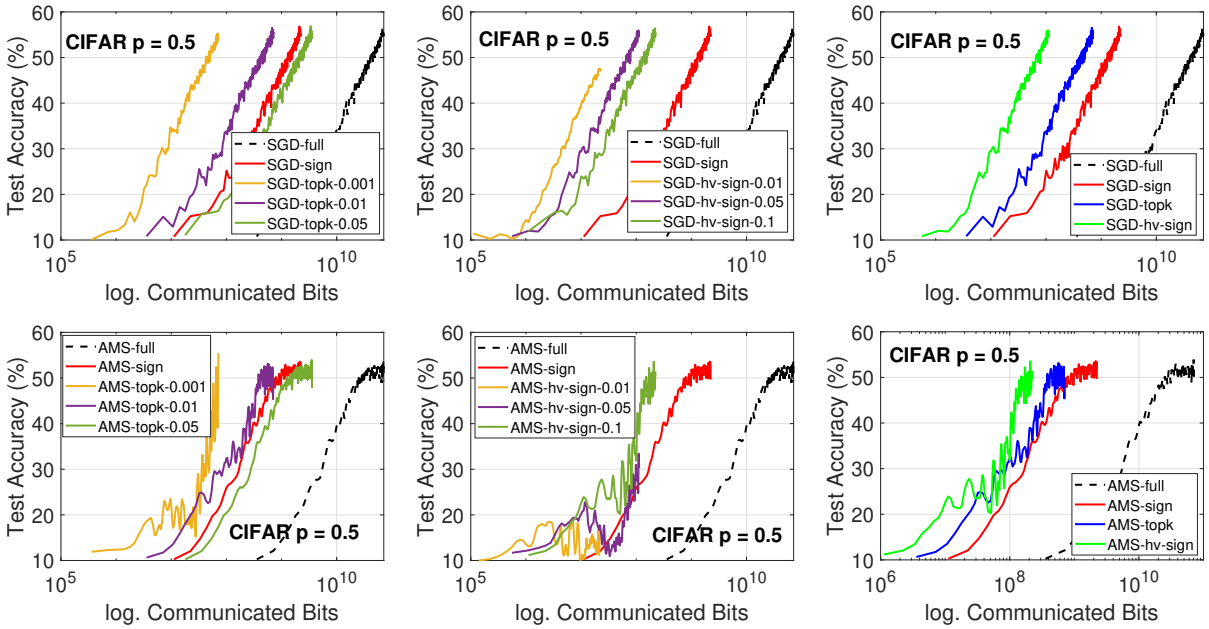

Figure 6. CIFAR-10 dataset trained by ResNet-18. Test accuracy of Fed-EF with **TopK**, **Sign** and **heavy-Sign** compressors. Participation rate $p = 0.5$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS. The last column presents the corresponding curves that achieve the full-precision accuracy using lowest communication.

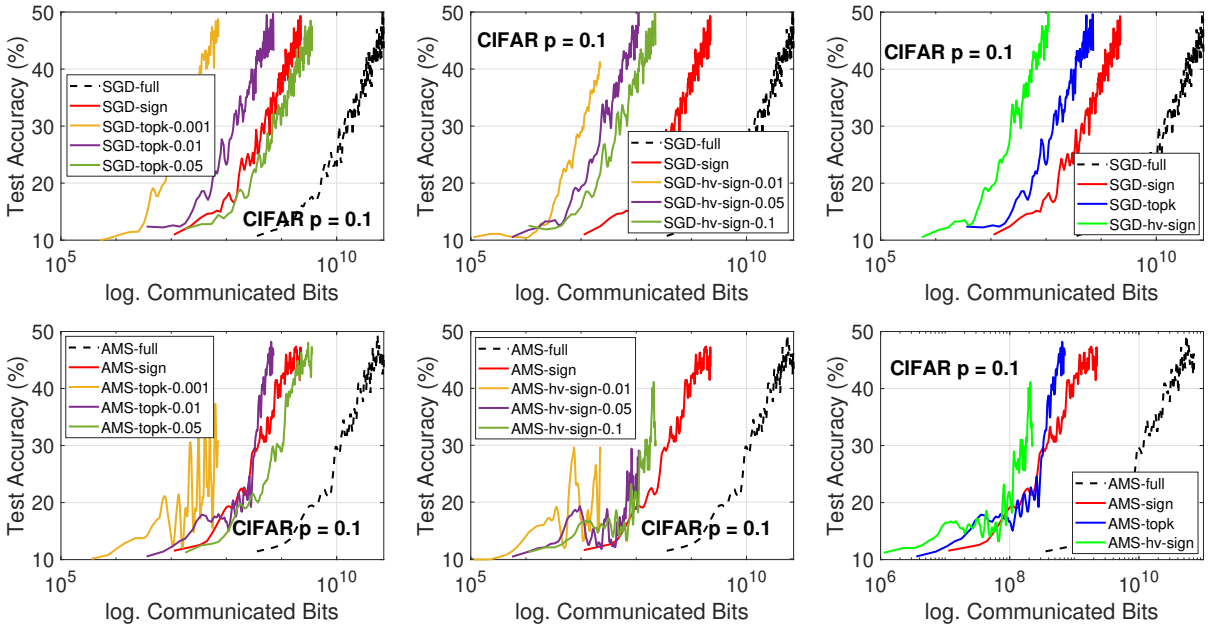

Figure 7. CIFAR-10 dataset trained by ResNet-18. Test accuracy of Fed-EF with **TopK**, **Sign** and **heavy-Sign** compressors. Participation rate $p = 0.1$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS. The last column presents the corresponding curves that achieve the full-precision accuracy using lowest communication.

D.3. More Experiment Results on MNIST and FMNIST

We provide the complete set of experimental results on each method under various compression rates. In Table 3 - Table 6, for completeness we report the average test accuracy at the end of training and the standard deviations, corresponding to the curves (compression parameters) in Figure 2 and Figure 3. Figure 8 to Figure 11 present the training loss and test accuracy under participation rate $p = 0.5$, and Figure 12 to Figure 15 report the loss and accuracy results for $p = 0.1$. All the results suggest that Fed-EF is able to perform on a par with full-precision FL with much less communication cost.

	Fed-EF-SGD			No EF	
	Sign	TopK	Hv-Sign	Stoc	Full-precision
MNIST	90.87 ($\pm$0.84)	91.04 ($\pm$1.05)	91.18 ($\pm$1.10)	90.16 ($\pm$0.96)	90.85 ($\pm$0.89)
FMNIST	71.13 ($\pm$0.68)	71.16 ($\pm$0.77)	71.07 ($\pm$0.83)	71.26 ($\pm$0.87)	71.20 ($\pm$0.71)

Table 3. Test accuracy (%) with client participation rate $p = 0.5$, of Fed-EF-SGD with **Sign**, **TopK** and **heavy-Sign** compressor and **Stoc** (stochastic quantization) without EF. The compression parameters (i.e., k and b) of the compressors are consistent with Figure 2.

	Fed-EF-AMS			No EF	
	Sign	TopK	Hv-Sign	Stoc	Full-precision
MNIST	92.32 ($\pm$0.98)	92.74 ($\pm$0.84)	91.77 ($\pm$1.22)	92.36 ($\pm$0.93)	92.23 ($\pm$0.73)
FMNIST	71.35 ($\pm$0.61)	71.90 ($\pm$0.78)	70.73($\pm$1.03)	71.94 ($\pm$0.95)	71.97 ($\pm$0.86)

Table 4. Test accuracy (%) with client participation rate $p = 0.5$, of Fed-EF-AMS with **Sign**, **TopK** and **heavy-Sign** compressor and **Stoc** (stochastic quantization) without EF. The compression parameters (i.e., k and b) of the compressors are consistent with Figure 2.

	Fed-EF-SGD			No EF	
	Sign	TopK	Hv-Sign	Stoc	Full-precision
MNIST	90.15 ($\pm$1.06)	90.61 ($\pm$0.93)	90.42 ($\pm$1.09)	90.27 ($\pm$1.18)	90.22 ($\pm$0.82)
FMNIST	67.69 ($\pm$0.73)	67.47 ($\pm$0.80)	67.72 ($\pm$0.55)	67.71 ($\pm$0.78)	67.50 ($\pm$0.85)

Table 5. Test accuracy (%) with client participation rate $p = 0.1$, of Fed-EF-SGD with **Sign**, **TopK** and **heavy-Sign** compressor and **Stoc** (stochastic quantization) without EF. The compression parameters (i.e., k and b) of the compressors are consistent with Figure 3.

	Fed-EF-AMS			No EF	
	Sign	TopK	Hv-Sign	Stoc	Full-precision
MNIST	88.67 ($\pm$1.11)	88.97 ($\pm$1.16)	77.49 ($\pm$1.53)	88.76 ($\pm$1.22)	89.05 ($\pm$1.04)
FMNIST	57.60 ($\pm$2.34)	64.09 ($\pm$0.91)	50.77($\pm$2.87)	64.35 ($\pm$1.06)	64.18 ($\pm$0.90)

Table 6. Test accuracy (%) with client participation rate $p = 0.1$, of Fed-EF-AMS with **Sign**, **TopK** and **heavy-Sign** compressor and **Stoc** (stochastic quantization) without EF. The compression parameters (i.e., k and b) of the compressors are consistent with Figure 3.

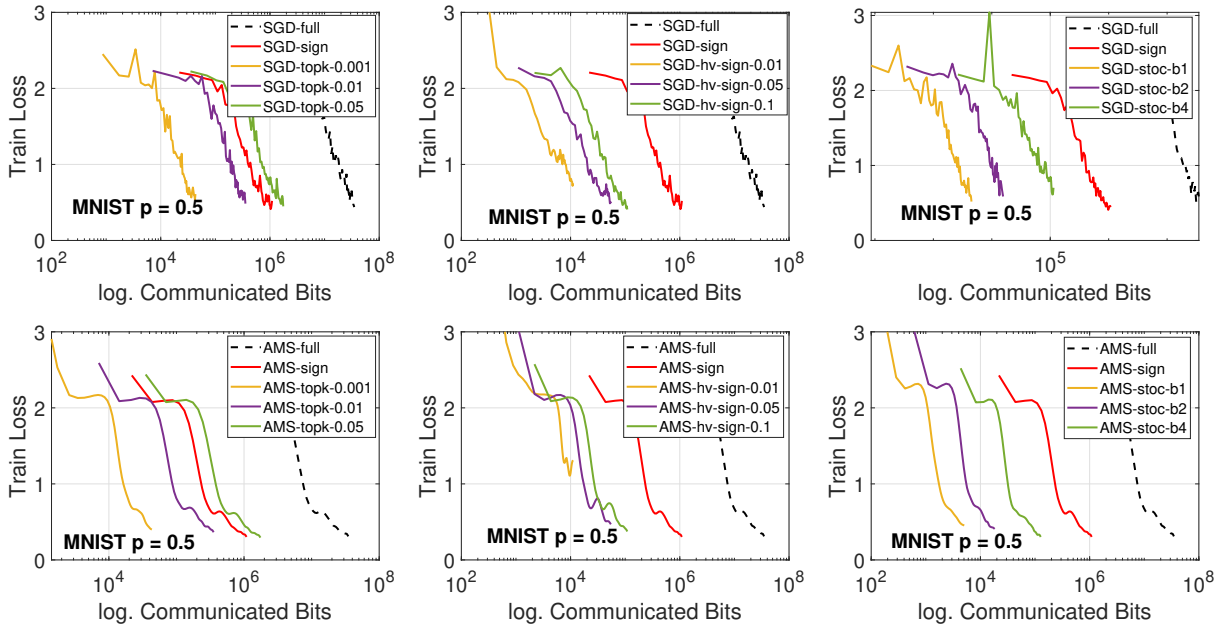

Figure 8. Training loss of Fed-EF on MNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.5$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

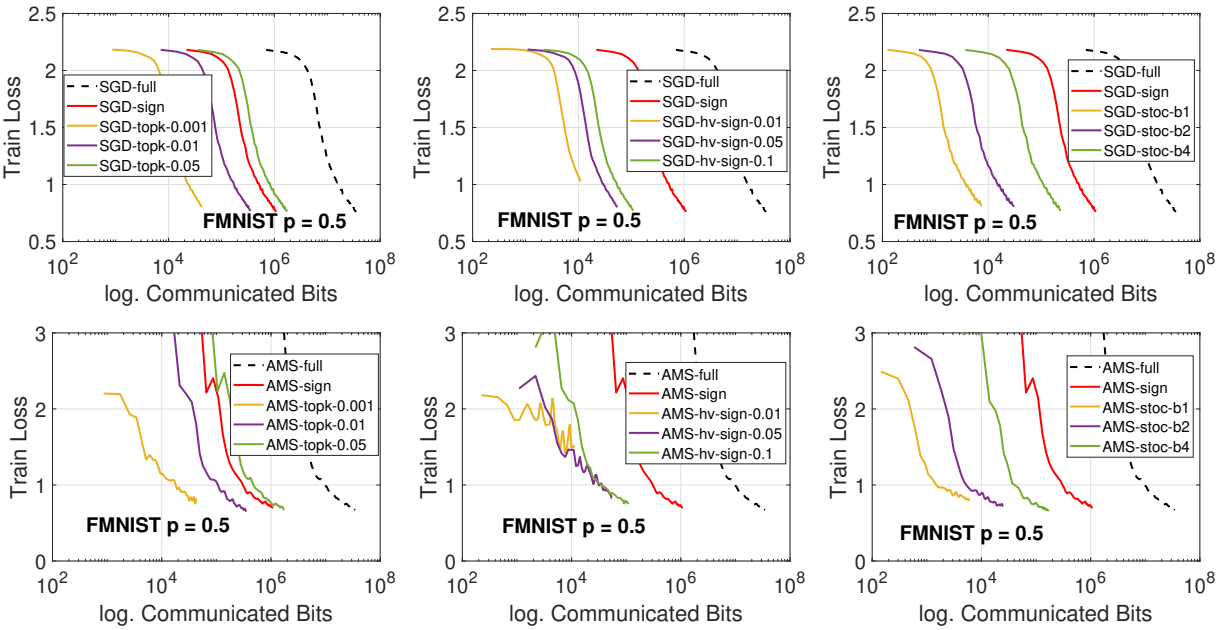

Figure 9. Training loss of Fed-EF on FMNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.5$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

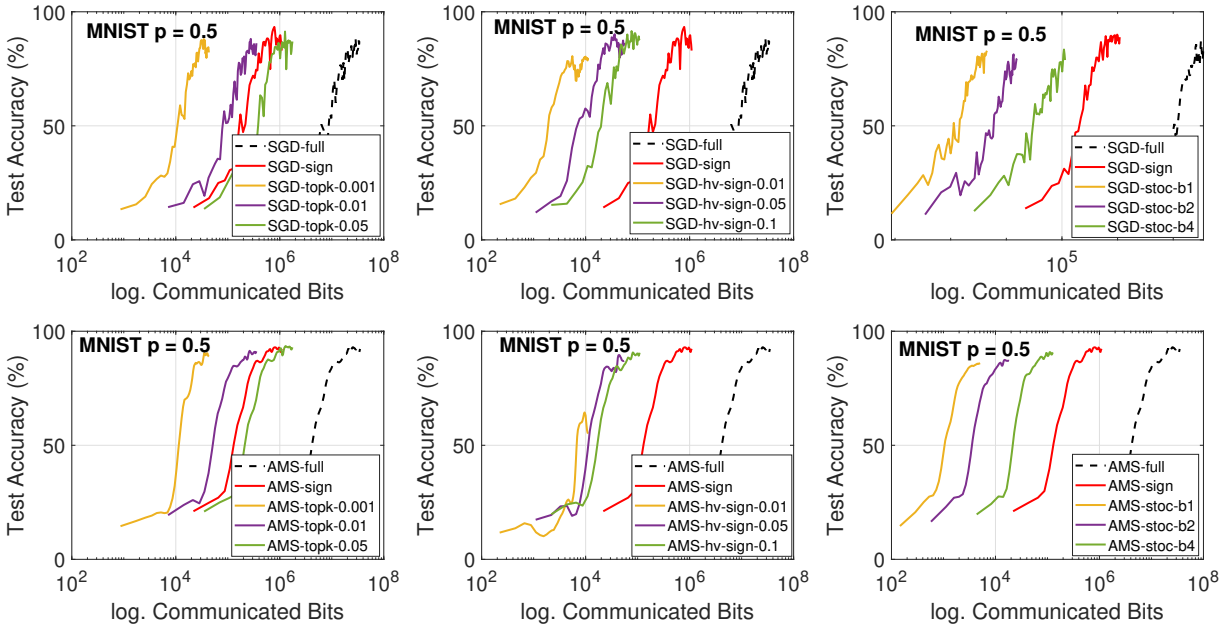

Figure 10. Test accuracy of Fed-EF on MNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.5$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

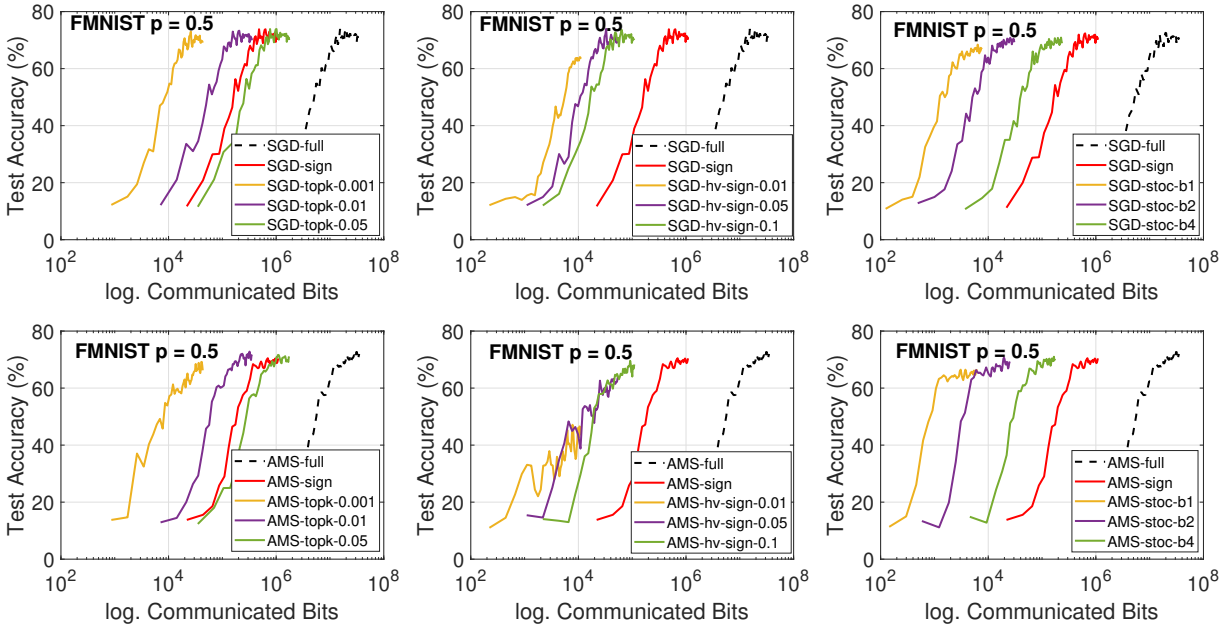

Figure 11. Test accuracy of Fed-EF on FMNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.5$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

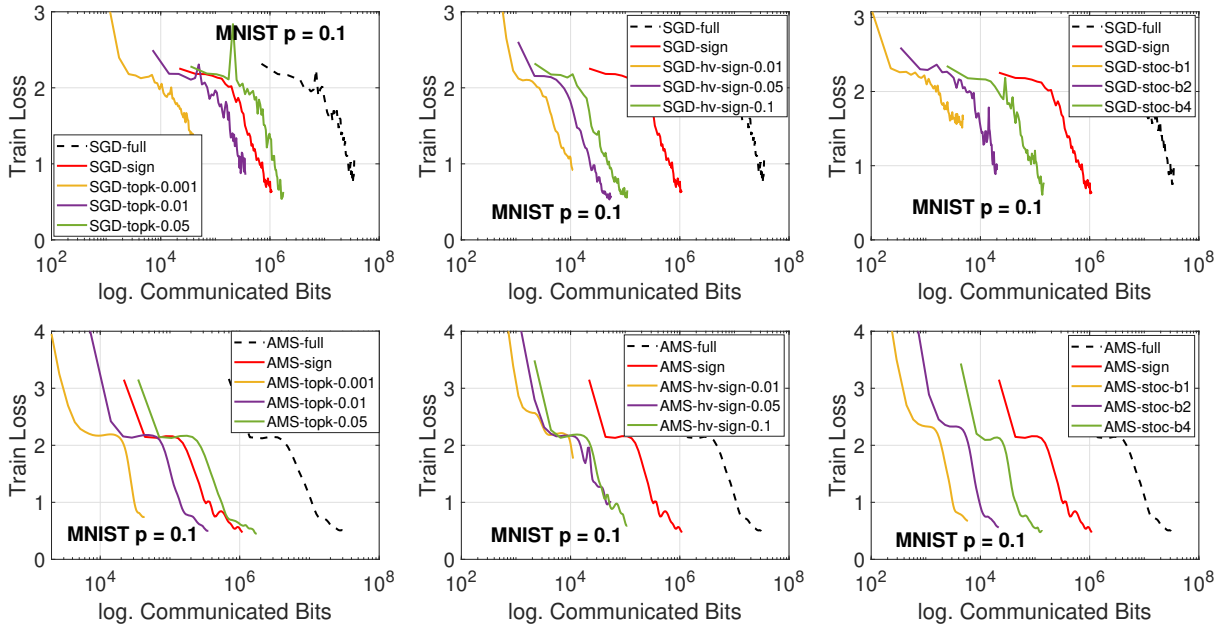

Figure 12. Training loss of Fed-EF on MNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.1$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

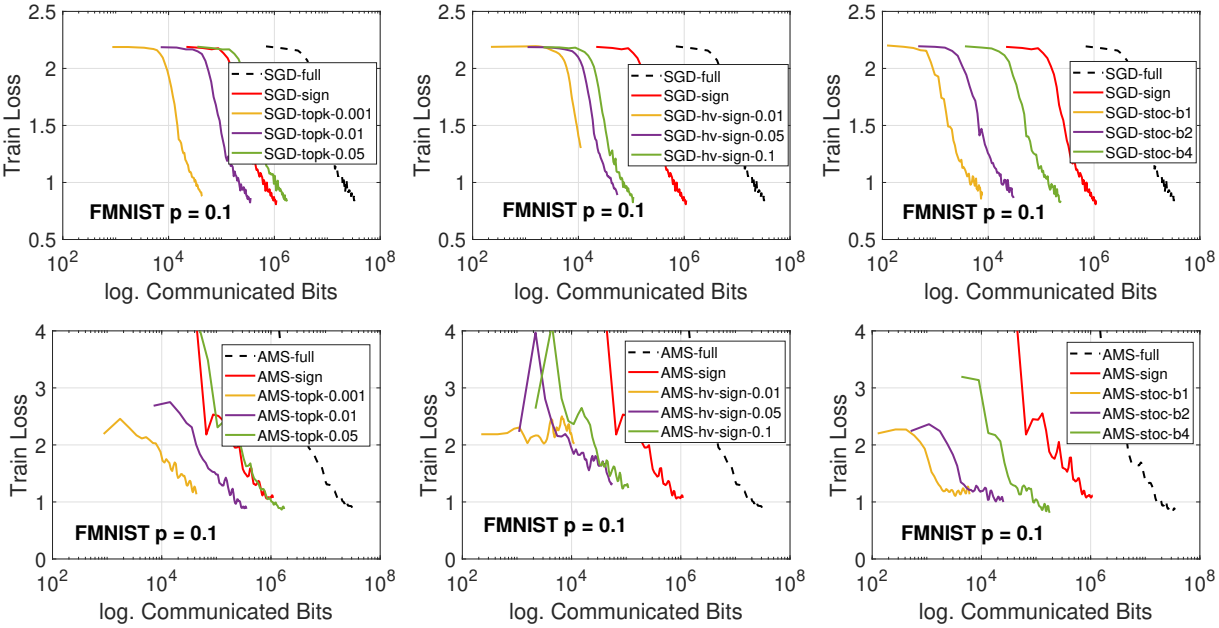

Figure 13. Training loss of Fed-EF on FMNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.1$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

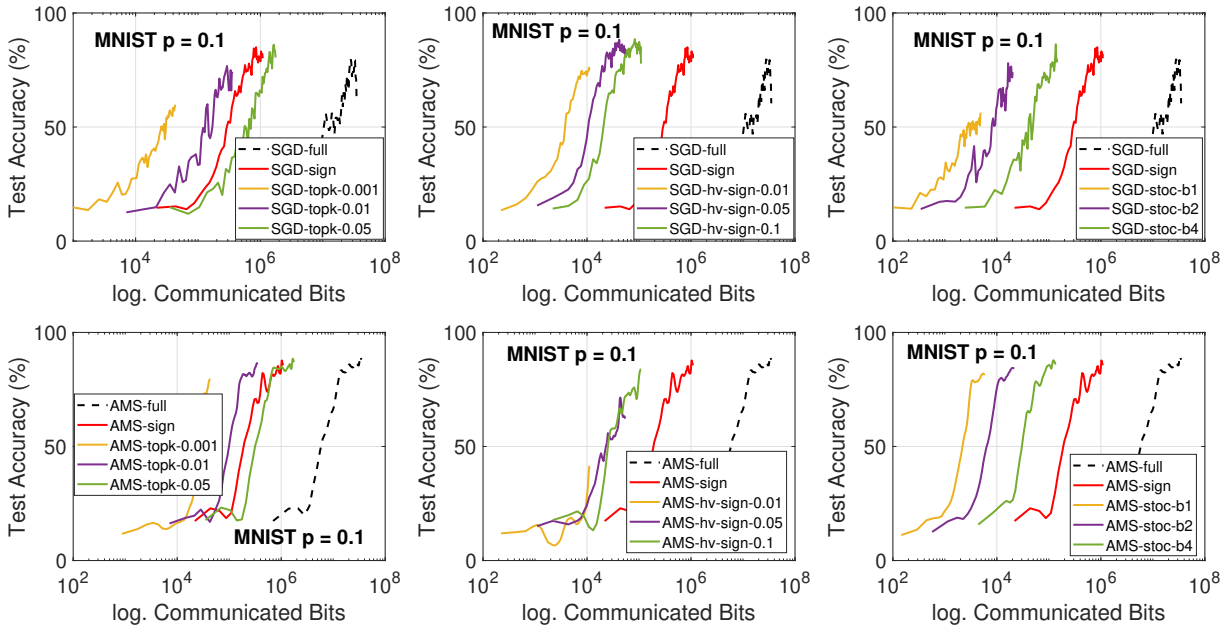

Figure 14. Test accuracy of Fed-EF on MNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.1$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

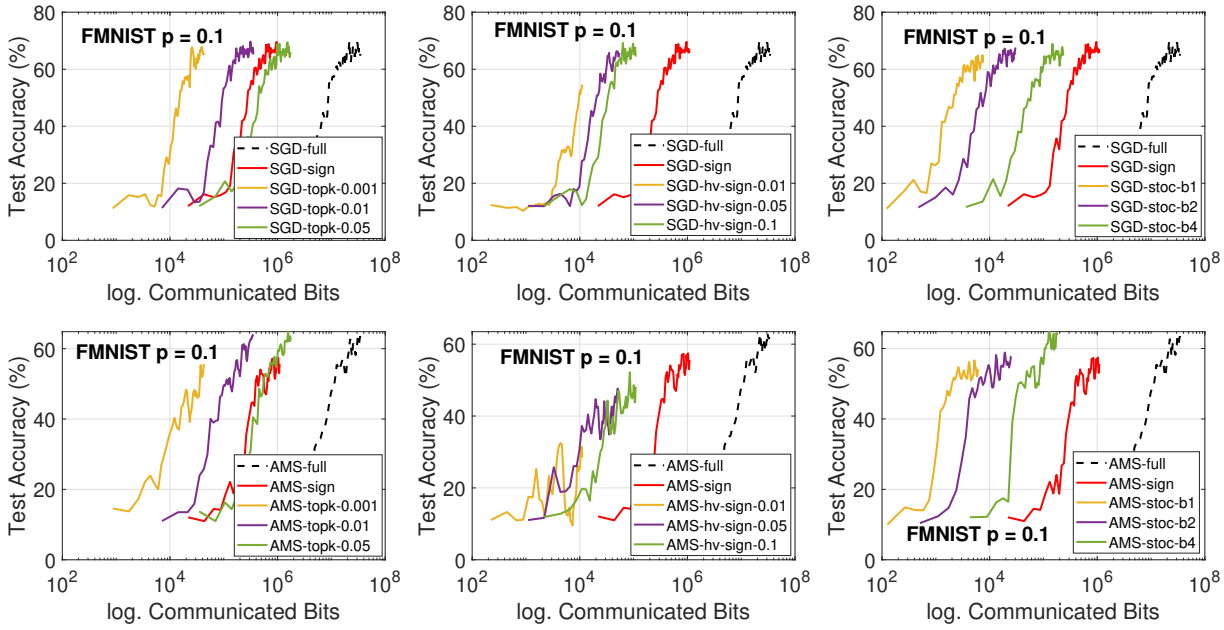

Figure 15. Test accuracy of Fed-EF on FMNIST dataset trained by CNN. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF. Participation rate $p = 0.1$, non-iid data. 1st row: Fed-EF-SGD. 2nd row: Fed-EF-AMS.

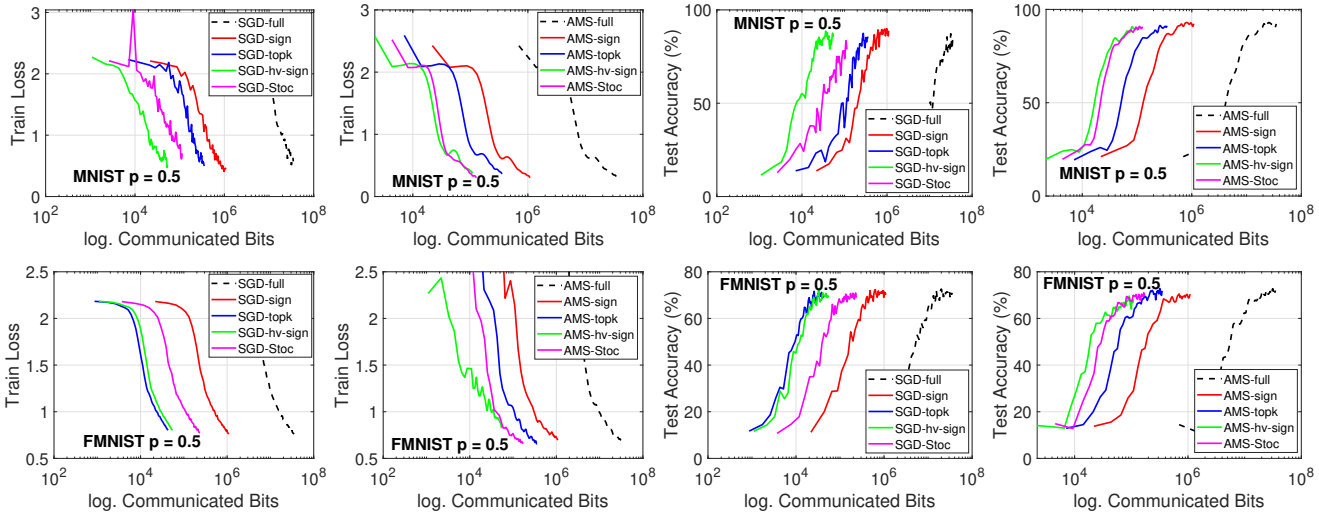

Figure 16. Training loss and test accuracy vs. communicated bits on MNIST and FMNIST datasets, participation rate $p = 0.5$. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF.

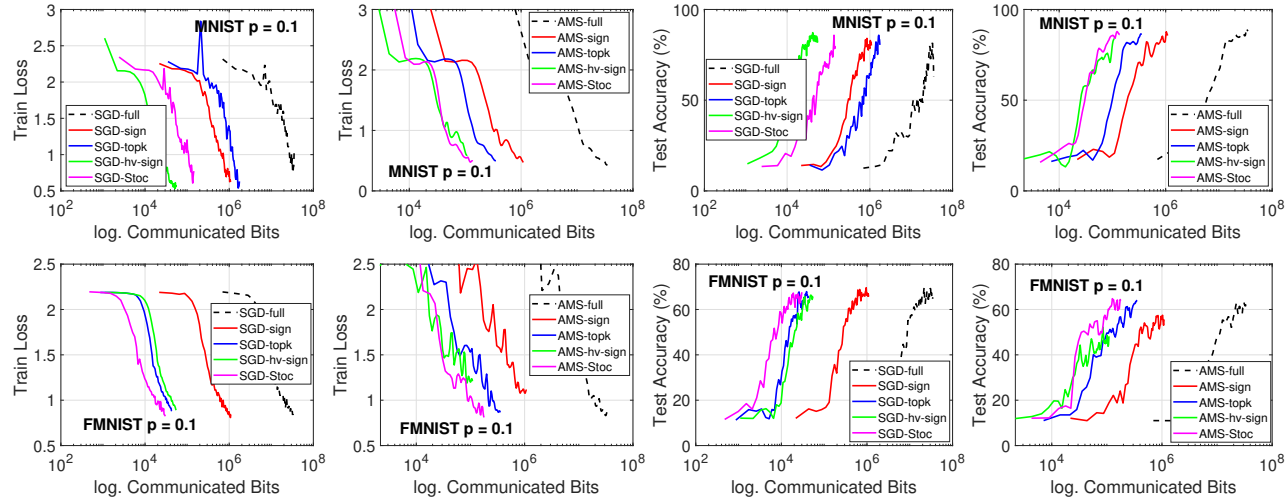

Figure 17. Training loss and test accuracy vs. communicated bits on MNIST and FMNIST datasets, participation rate $p = 0.1$. "sign", "topk" and "hv-sign" are applied with Fed-EF, while "Stoc" is the stochastic quantization without EF.

E. Compression Discrepancy

In our theoretical analysis for Fed-EF, Assumption 4.3 is needed, which states that $\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \mathcal{C}\big(\Delta_{t,i}+e_{t,i}\big) - \frac{1}{n}\sum_{i=1}^n(\Delta_{t,i}+e_{t,i})\|^2] \leq q_{\mathcal{A}}^2\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n(\Delta_{t,i}+e_{t,i})\|^2]$ for some $q_{\mathcal{A}} < 1$ during training. In the following, we justify this assumption to demonstrate how it holds in practice. It is worth mentioning that a similar condition is was assumed in Haddadpour et al. (2021) for the analysis of FL with unbiased compression. To study sparsified SGD, (Alistarh et al., 2018) also used a similar and stronger (uniform bound instead of in expectation) analytical assumption. As a result, our analysis and theoretical results are also valid under their assumption.

E.1. Simulated Data

We first conduct a simulation to investigate how the two compressors, **TopK** and **Sign**, affect $q_{\mathcal{A}}$. In our presented results, for conciseness we use $n = 5$ clients and model dimensionality $d = 1100$. Similar conclusions hold for much larger n and d. We simulate two types of gradients following normal distribution and Laplace distribution (more heavy-tailed), respectively. Examples of the simulated gradients are visualized in Figure 18 and Figure 19. To mimic non-iid data, we assume that each client has some strong signals (large gradients) in some coordinates, and we scale those gradients by a scaling factor $s = 2, 10, 100$. Conceptually, larger s represents higher data heterogeneity.

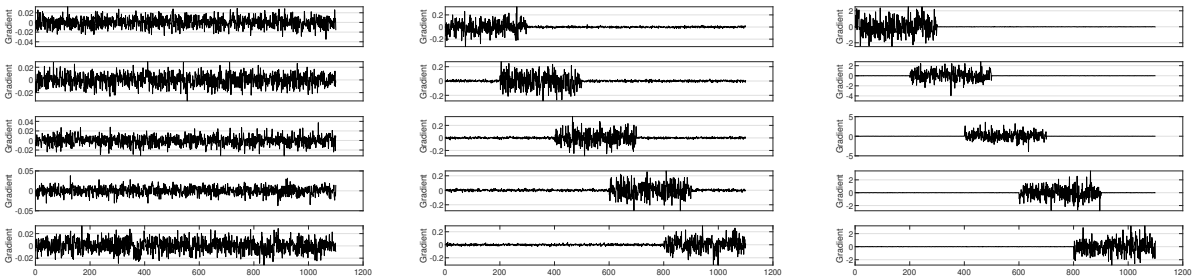

Figure 18. The simulated gradients of 5 heterogeneous clients, from $N(0, \gamma^2)$ with $\gamma = 0.01$. The gradient on each distinct client is scaled by $s = 2, 10, 100$ (left, mid, right), respectively. Larger s implies higher data heterogeneity.

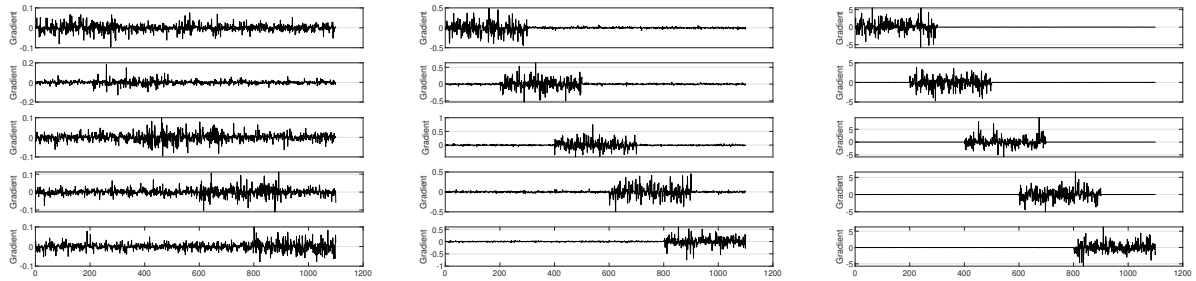

Figure 19. The simulated gradients of 5 heterogeneous clients, from $Lap(0, \lambda)$ with $\lambda = 0.01$. The x-axis is the dimension. The gradient on each distinct client is scaled by $s = 2, 10, 100$ (left, mid, right), respectively. Wee the gradients have heavier tail than normal distribution. Larger s implies higher data heterogeneity.

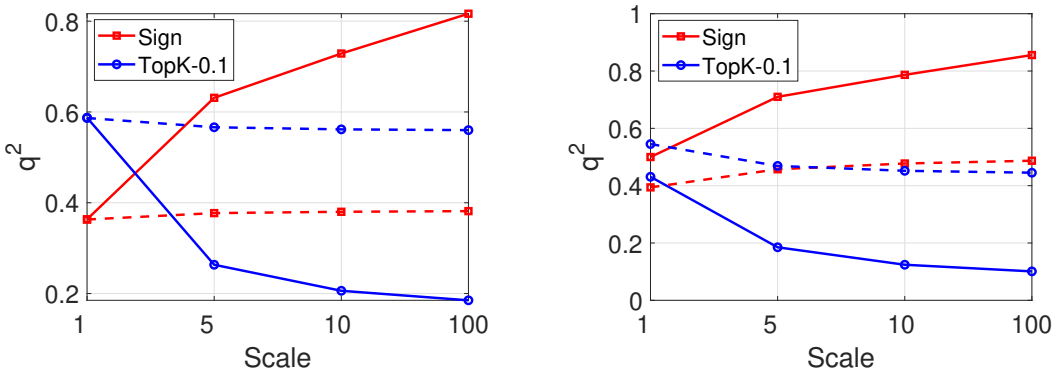

Figure 20. The compression coefficient $q_{\mathcal{A}}$ in Assumption 4.3 on simulated gradients. **TopK** is applied with sparsity $k = 0.1$. Left: Gaussian distribution. Right: Laplace distribution. $q^2_{\mathcal{A}}$ is computed by $q = \frac{\|\delta(x)-x\|^2}{\|x\|^2}$ where $\delta(x) = \frac{1}{n}\sum_{i=1}^n \mathcal{C}(\Delta_{t,i} + e_{t,i})$ and $x = \frac{1}{n}\sum_{i=1}^n(\Delta_{t,i} + e_{t,i})$. The dashed curves are respectively the compression coefficients $q^2_{\mathcal{C}}$ from Definition 3.1, which is calculated by replacing $\delta(x) = \mathcal{C}(\frac{1}{n}\sum_{i=1}^n \Delta_{t,i} + e_{t,i})$. We see that in all cases, $q_{\mathcal{A}} < 1$.

We apply the **TopK-0.1** and **Sign** compressor in Definition 3.1 to the simulated gradients, and compute the averaged $q^2_{\mathcal{A}}$ in Figure 20 over 10^5 independent runs. The dashed curves are respectively the "ideal" compression coefficients $q_{\mathcal{C}}$ such that $\mathbb{E}[\|\mathcal{C}\big(\frac{1}{n}\sum_{i=1}^n \Delta_{t,i} + e_{t,i}\big) - \frac{1}{n}\sum_{i=1}^n(\Delta_{t,i} + e_{t,i})\|^2] \le q^2_{\mathcal{C}}\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n(\Delta_{t,i} + e_{t,i})\|^2]$ from Definition 3.1. We see that in all cases, $q_{\mathcal{A}}$ is indeed less than 1. This still holds even when the data heterogeneity increases to as large as 100.

E.2. Real-world Data

We report the empirical $q_{\mathcal{A}}$ values when training CNN on MNIST and FMNIST datasets. The experimental setup is the same as in Section 5. We present the result in Figure 21 with $\eta = 1$, $\eta_l = 0.01$ under the same heterogeneous setting where client data are highly non-iid. The plots for other learning rate combinations and iid data are similar. In particular, we see for both compressors and both datasets, the empirical $q_{\mathcal{A}}$ is well-bounded below 1 throughout the training process.

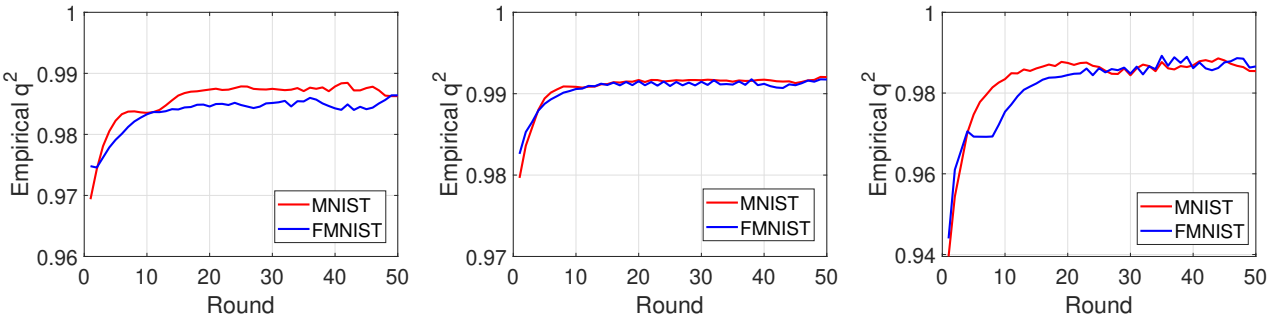

Figure 21. The compression coefficient $q_{\mathcal{A}}$ in Assumption 4.3 in our experiments (Section 5) for CNN trained on MNIST and FMNIST dataset, averaged over multiple runs. $\eta = 1$, $\eta_l = 0.01$, non-iid client data distribution. Left: **Sign** compression. Mid: **TopK** compression with $k = 0.01$. Right: **TopK** compression with $k = 0.1$.

F. Proof of Convergence Results

We first present the proof for the more complicated Fed-EF-AMS in Section F.1, and the proof of Fed-EF-SGD would follow in Section F.2. Section F.3 contains intermediary lemmas, Section F.4 provides the analysis of Fed-EF in partial participation and Section F.5 proves the rate of FL directly using biased compression.

F.1. Proof of Theorem 4.8: Fed-EF-AMS

Proof. We first clarify some notations. At round t, let the full-precision local model update of the i-th worker be $\Delta_{t,i}$, the error accumulator be $e_{t,i}$, and denote $\widetilde{\Delta}_{t,i} = \mathcal{C}(g_{t,i} + e_{t,i})$. Define $\bar{\Delta}_t = \frac{1}{n}\sum_{i=1}^n \Delta_{t,i}$, $\overline{\widetilde{\Delta}}_t = \frac{1}{n}\sum_{i=1}^n \widetilde{\Delta}_{t,i}$ and $\bar{e}_t = \frac{1}{n}\sum_{i=1}^n e_{t,i}$. The second moment computed by the compressed local model updates is denoted as $v_t = \beta_2 v_{t-1} + (1-\beta_2)\overline{\widetilde{\Delta}}_t^2$, and $\hat{v}_t = \max\{\hat{v}_{t-1}, v_t\}$. Also, the first order moving average sequence

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\overline{\widetilde{\Delta}}_t \quad \text{and} \quad m_t' = \beta_1 m_{t-1}' + (1-\beta_1)\bar{\Delta}_t,$$

where m_t' represents the first moment moving average sequence using the uncompressed updates. By construction we have $m_t' = (1-\beta_1)\sum_{\tau=1}^t \beta_1^{t-\tau}\bar{\Delta}_\tau$.

Our proof will use the following auxiliary sequences: for round $t = 1, ..., T$,

$$\mathcal{E}_{t+1} := (1-\beta_1)\sum_{\tau=1}^{t+1}\beta_1^{t+1-\tau}\bar{e}_\tau,$$
$$\theta_{t+1}' := \theta_{t+1} - \eta\frac{\mathcal{E}_{t+1}}{\sqrt{\hat{v}_t+\epsilon}}.$$

Then, we can write the evolution of θ_t' as

$$\begin{aligned}
\theta_{t+1}' &= \theta_{t+1} - \eta\frac{\mathcal{E}_{t+1}}{\sqrt{\hat{v}_t+\epsilon}} \\
&= \theta_t - \eta\frac{(1-\beta_1)\sum_{\tau=1}^t\beta_1^{t-\tau}\overline{\widetilde{\Delta}}_\tau + (1-\beta_1)\sum_{\tau=1}^{t+1}\beta_1^{t+1-\tau}\bar{e}_\tau}{\sqrt{\hat{v}_t+\epsilon}} \\
&= \theta_t - \eta\frac{(1-\beta_1)\sum_{\tau=1}^t\beta_1^{t-\tau}(\overline{\widetilde{\Delta}}_\tau + \bar{e}_{\tau+1}) + (1-\beta)\beta_1^t\bar{e}_1}{\sqrt{\hat{v}_t+\epsilon}} \\
&= \theta_t - \eta\frac{(1-\beta_1)\sum_{\tau=1}^t\beta_1^{t-\tau}\bar{e}_\tau}{\sqrt{\hat{v}_t+\epsilon}} - \eta\frac{m_t'}{\sqrt{\hat{v}_t+\epsilon}} \\
&= \theta_t - \eta\frac{\mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}} - \eta\frac{m_t'}{\sqrt{\hat{v}_t+\epsilon}} + \eta(\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}} - \frac{1}{\sqrt{\hat{v}_t+\epsilon}})\mathcal{E}_t \\
&\overset{(a)}{=} \theta_t' - \eta\frac{m_t'}{\sqrt{\hat{v}_t+\epsilon}} + \eta(\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}} - \frac{1}{\sqrt{\hat{v}_t+\epsilon}})\mathcal{E}_t \\
&:= \theta_t' - \eta\frac{m_t'}{\sqrt{\hat{v}_t+\epsilon}} + \eta D_t\mathcal{E}_t,
\end{aligned}$$

where (a) uses the fact of error feedback that for every $i \in [n]$, $\widetilde{\Delta}_{t,i} + e_{t+1,i} = \Delta_{t,i} + e_{t,i}$, and $e_{t,1} = 0$ at initialization. Further define the virtual iterates:

$$x_{t+1} := \theta_{t+1}' - \eta\frac{\beta_1}{1-\beta_1}\frac{m_t'}{\sqrt{\hat{v}_t+\epsilon}},$$

which follows the recurrence:

$$\begin{aligned}
x_{t+1} &= \theta'_{t+1} - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} \\
&= \theta'_t - \eta \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta D_t \mathcal{E}_t \\
&= \theta'_t - \eta \frac{\beta_1 m'_{t-1} + (1-\beta_1)\bar{\Delta}_t + \frac{\beta_1^2}{1-\beta_1} m'_{t-1} + \beta_1 \bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta D_t \mathcal{E}_t \\
&= \theta'_t - \eta \frac{\beta_1}{1-\beta_1} \frac{m'_{t-1}}{\sqrt{\hat{v}_t + \epsilon}} - \eta \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta D_t \mathcal{E}_t \\
&= x_t - \eta \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} + \eta \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + \eta D_t \mathcal{E}_t.
\end{aligned}$$

The general idea is to study the convergence of the sequence x_t, and show that the difference between x_t and θ_t (of interest) is small. First, by the smoothness Assumption 4.1, we have

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2.$$

Taking expectation w.r.t. the randomness at round t and re-arranging terms, we obtain

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - f(x_t) \\
&\le -\eta \mathbb{E}\Big[\big\langle \nabla f(x_t), \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} \big\rangle\Big] + \eta \mathbb{E}\Big[\big\langle \nabla f(x_t), \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t \mathcal{E}_t \big\rangle\Big] \\
&\qquad + \frac{\eta^2 L}{2} \mathbb{E}\Big[\| \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} - \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} - D_t \mathcal{E}_t \|^2\Big] \\
&= \underbrace{-\eta \mathbb{E}\Big[\big\langle \nabla f(\theta_t), \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} \big\rangle\Big]}_{I} + \underbrace{\eta \mathbb{E}\Big[\big\langle \nabla f(x_t), \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t \mathcal{E}_t \big\rangle\Big]}_{II} \\
&\qquad + \underbrace{\frac{\eta^2 L}{2} \mathbb{E}\Big[\| \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} - \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} - D_t \mathcal{E}_t \|^2\Big]}_{III} + \underbrace{\eta \mathbb{E}\Big[\big\langle \nabla f(\theta_t) - \nabla f(x_t), \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t + \epsilon}} \big\rangle\Big]}_{IV},
\end{aligned} \tag{3}$$

Bounding term I. We have

$$\begin{aligned}
\boldsymbol{I} &= -\eta \mathbb{E}\Big[\big\langle \nabla f(\theta_t), \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_{t-1} + \epsilon}} \big\rangle\Big] - \eta \mathbb{E}\Big[\big\langle \nabla f(\theta_t), (\frac{1}{\sqrt{\hat{v}_t + \epsilon}} - \frac{1}{\sqrt{\hat{v}_{t-1} + \epsilon}}) \bar{\Delta}_t \big\rangle\Big] \\
&\le -\eta \mathbb{E}\Big[\big\langle \nabla f(\theta_t), \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_{t-1} + \epsilon}} \big\rangle\Big] + \eta \eta_l K G^2 \mathbb{E}[\|D_t\|_1],
\end{aligned} \tag{4}$$

where we use Assumption 4.7 on the stochastic gradient magnitude. The last inequality holds by simply bounding the aggregated local model update by

$$\|\bar{\Delta}_t\| \le \frac{1}{n} \sum_{i=1}^{n} \|\eta_l \sum_{k=1}^{K} g_{t,i}^{(k)}\| \le \eta_l K G,$$

and the fact that for any vector in $\mathbb{R}^d$, the l_2 norm is upper bounded by the l_1 norm.

Regarding the first term in (4), we have

$$
\begin{aligned}
&-\eta\mathbb{E}\Big[\langle\nabla f(\theta_t),\frac{\bar{\Delta}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\Big]=-\eta\mathbb{E}\Big[\langle\frac{\nabla f(\theta_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}},\bar{\Delta}_t-\eta_l K\nabla f(\theta_t)+\eta_l K\nabla f(\theta_t)\rangle\Big]\\
&=-\eta\eta_l K\mathbb{E}\Big[\frac{\|\nabla f(\theta_t)\|^2}{\sqrt{\hat{v}_{t-1}+\epsilon}}\Big]+\eta\mathbb{E}\Big[\langle\frac{\nabla f(\theta_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}},-\bar{\Delta}_t+\eta_l K\nabla f(\theta_t)\rangle\Big]\\
&\overset{(a)}{\leq}-\frac{\eta\eta_l K}{\sqrt{\frac{4\eta_l^2(1+q^2)^3K^2}{(1-q^2)^2}G^2+\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\eta\big\langle\frac{\nabla f(\theta_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}},\mathbb{E}\big[-\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K\eta_l g_{t,i}^{(k)}+\eta_l K\nabla f(\theta_t)\big]\big\rangle\\
&\overset{(b)}{=}-\frac{\eta\eta_l K}{\sqrt{\frac{4\eta_l^2(1+q^2)^3K^2}{(1-q^2)^2}G^2+\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]\\
&\qquad\qquad+\underbrace{\eta\,\big\langle\frac{\sqrt{\eta_l}\nabla f(\theta_t)}{(\hat{v}_{t-1}+\epsilon)^{1/4}},\mathbb{E}\Big[\frac{\sqrt{\eta_l}}{n(\hat{v}_{t-1}+\epsilon)^{1/4}}(-\sum_{i=1}^n\sum_{k=1}^K\nabla f_i(\theta_{t,i}^{(k)})+K\nabla f(\theta_t))\Big]\big\rangle}_{V},
\end{aligned}
$$

where (a) uses Lemma F.7 and (b) is due to Assumption 4.2 that $g_{t,i}^{(k)}$ is an unbiased estimator of $\nabla f_i(\theta_{t,i}^{(k)})$. To bound term **V**, we use the similar proof structure as in the proof of Lemma F.3. Specifically, we have

$$
\begin{aligned}
\boldsymbol{V}&\leq\frac{\eta_l K}{2\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{\eta_l}{2K\sqrt{\epsilon}}\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K(\nabla f_i(\theta_{t,i}^{(k)})-\nabla f_i(\theta_t))\|^2\big]\\
&\leq\frac{\eta_l K}{2\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{\eta_l}{2nK\sqrt{\epsilon}}\mathbb{E}\big[\sum_{i=1}^n\|\sum_{k=1}^K(\nabla f_i(\theta_{t,i}^{(k)})-\nabla f_i(\theta_t))\|^2\big]\\
&\leq\frac{\eta_l K}{2\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{\eta_l}{2n\sqrt{\epsilon}}\mathbb{E}\big[\sum_{i=1}^n\sum_{k=1}^K\|\nabla f_i(\theta_{t,i}^{(k)})-\nabla f_i(\theta_t)\|^2\big]\\
&\leq\frac{\eta_l K}{2\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{\eta_l L^2}{2n\sqrt{\epsilon}}\mathbb{E}\big[\sum_{i=1}^n\sum_{k=1}^K\|\theta_{t,i}^{(k)}-\theta_t\|^2\big],
\end{aligned}
$$

where the last inequality is a result of the L-smoothness assumption on the loss function $f_i(x)$. Applying Lemma F.1 to the consensus error, we can further bound term V by

$$
\begin{aligned}
\boldsymbol{V}&\leq\frac{\eta_l K}{2\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{\eta_l KL^2}{2\sqrt{\epsilon}}\big[5\eta_l^2K(\sigma^2+6K\sigma_g^2)+30\eta_l^2K^2\mathbb{E}[\|\nabla f(\theta_t)\|^2]\big]\\
&\leq\frac{47\eta_l K}{64\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{5\eta_l^3K^2L^2}{2\sqrt{\epsilon}}(\sigma^2+6K\sigma_g^2),
\end{aligned}
$$

when we choose $\eta_l\leq\frac{1}{8KL}$. Further, if we set $\eta_l\leq\frac{\sqrt{15}(1-q^2)\sqrt{\epsilon}}{14(1+q^2)^{1.5}KG}$, we have

$$
\frac{4\eta_l^2(1+q^2)^3K^2}{(1-q^2)^2}G^2+\epsilon\leq\frac{60}{196}\epsilon+\epsilon=\frac{64}{49}\epsilon.
$$

Hence, as $\frac{47}{64}<\frac{3}{4}$, we can establish from (4) that

$$
\boldsymbol{I}\leq-\frac{\eta\eta_l K}{8\sqrt{\epsilon}}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+\frac{5\eta\eta_l^3K^2L^2}{2\sqrt{\epsilon}}(\sigma^2+6K\sigma_g^2)+\eta\eta_l KG^2\mathbb{E}[\|D_t\|_1].\tag{5}
$$

Bounding term II. By Lemma F.6, we know that $\|\mathcal{E}_t\| \leq \frac{2\eta_l q K G}{1-q^2}$, and by Lemma F.4, $\|m'_t\| \leq \eta_l K G$. Thus, we have

$$\begin{aligned}
\boldsymbol{II} &\leq \eta\Big(\mathbb{E}\big[\langle \nabla f(\theta_t), \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t\mathcal{E}_t\rangle\big] + \mathbb{E}\big[\langle \nabla f(x_t) - \nabla f(\theta_t), \frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t\mathcal{E}_t\rangle\big]\Big) \\
&\leq \eta\mathbb{E}\big[\|\nabla f(\theta_t)\|\|\frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t\mathcal{E}_t\|\big] \\
&\qquad + \eta^2\, L\mathbb{E}\big[\|\frac{\frac{\beta_1}{1-\beta_1} m'_{t-1} + \mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\|\|\frac{\beta_1}{1-\beta_1} D_t m'_{t-1} + D_t\mathcal{E}_t\|\big] \\
&\leq \eta\eta_l C_1 K G^2 \mathbb{E}[\|D_t\|_1] + \frac{\eta^2\eta_l^2 C_1^2 L K^2 G^2}{\sqrt{\epsilon}}\mathbb{E}[\|D_t\|_1],
\end{aligned} \tag{6}$$

where $C_1 := \frac{\beta_1}{1-\beta_1} + \frac{2q}{1-q^2}$, and the second inequality is due to the smoothness of $f(\theta)$.

Bounding term III. This term can be bounded as follows:

$$\begin{aligned}
\boldsymbol{III} &\leq \eta^2 L\mathbb{E}\big[\|\frac{\bar{\Delta}_t}{\sqrt{\hat{v}_t+\epsilon}}\|^2\big] + \eta^2 L\mathbb{E}\big[\|\frac{\beta_1}{1-\beta_1} D_t m'_{t-1} - D_t\mathcal{E}_t\|^2\big] \\
&\leq \frac{\eta^2 L}{\epsilon}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] + \eta^2 L\mathbb{E}\big[\|D_t(\frac{\beta_1}{1-\beta_1} m'_{t-1} - \mathcal{E}_t)\|^2\big] \\
&\leq \frac{\eta^2 L(2\eta_l^2 K^2 + 120\eta_l^4 K^4 L^2)}{\epsilon}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{4\eta^2\eta_l^2 K L}{n\epsilon}\sigma^2 \\
&\qquad + \frac{20\eta^2\eta_l^4 K^3 L^3}{\epsilon}(\sigma^2 + 6K\sigma_g^2) + \eta^2\eta_l^2 C_1^2 L K^2 G^2 \mathbb{E}[\|D_t\|^2],
\end{aligned} \tag{7}$$

where we apply Lemma F.2 and use similar argument as in bounding term II.

Bounding term IV. Lastly, for term IV, we have for some $\rho > 0$,

$$\begin{aligned}
\boldsymbol{IV} &= \eta\mathbb{E}\Big[\big\langle \nabla f(\theta_t) - \nabla f(x_t), \frac{\bar{\Delta}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\big\rangle\Big] + \eta\mathbb{E}\Big[\big\langle \nabla f(\theta_t) - \nabla f(x_t), (\frac{1}{\sqrt{\hat{v}_t+\epsilon}} - \frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}})\bar{\Delta}_t\big\rangle\Big] \\
&\overset{(a)}{\leq} \frac{\eta\rho}{2\epsilon}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] + \frac{\eta}{2\rho}\mathbb{E}\big[\|\nabla f(\theta_t) - \nabla f(x_t)\|^2\big] + \eta^2 L\mathbb{E}\Big[\|\frac{\frac{\beta_1}{1-\beta_1} m'_{t-1} + \mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\|\|D_t\Delta_t\|\Big] \\
&\overset{(b)}{\leq} \frac{\rho\eta(\eta_l^2 K^2 + 60\eta_l^4 K^4 L^2)}{\epsilon}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2\rho\eta\eta_l^2 K}{\epsilon n}\sigma^2 + \frac{10\rho\eta\eta_l^4 K^3 L^2}{\epsilon}(\sigma^2 + 6K\sigma_g^2) \\
&\qquad + \frac{\eta^3 L^2}{2\rho}\mathbb{E}\Big[\|\frac{\frac{\beta_1}{1-\beta_1} m'_{t-1} + \mathcal{E}_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\|^2\Big] + \frac{\eta^2\eta_l^2 C_1 L K^2 G^2}{\sqrt{\epsilon}}\mathbb{E}[\|D_t\|] \\
&\leq \frac{\rho\eta\eta_l^2 K^2(60\eta_l^2 K^2 L^2 + 1)}{\epsilon}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2\rho\eta\eta_l^2 K}{\epsilon n}\sigma^2 + \frac{10\rho\eta\eta_l^4 K^3 L^2}{\epsilon}(\sigma^2 + 6K\sigma_g^2) \\
&\qquad + \frac{\eta^3 L^2}{\rho\epsilon}\Big[\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}\big[\|m'_t\|^2\big] + \mathbb{E}\big[\|\mathcal{E}_t\|^2\big]\Big] + \frac{\eta^2\eta_l^2 C_1 L K^2 G^2}{\sqrt{\epsilon}}\mathbb{E}[\|D_t\|_1],
\end{aligned} \tag{8}$$

where (a) is a consequence of Young's inequality (ρ will be specified later) and the smoothness Assumption 4.1, and (b) is based on Lemma F.2.

After we have bounded all four terms in (3), the next step is to gather the ingredients by taking the telescope sum over $t = 1, ..., T$. For the ease of presentation, we first do this for the third term in (8). According to Lemma F.4 and Lemma F.6,

summing over $t = 1, ..., T$, we conclude

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\eta^3 L^2}{\rho\epsilon}\Big[\frac{\beta_1^2}{(1-\beta_1)^2}\mathbb{E}[\|m_t'\|^2] + \mathbb{E}[\|\mathcal{E}_t\|^2]\Big] \\
&\leq \frac{\eta^3\beta_1^2 L^2}{\rho(1-\beta_1)^2\epsilon}\Big[2\eta_l^2K^2(60\eta_l^2K^2L^2+1)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + 4\frac{T\eta_l^2K}{n}\sigma^2 + 20T\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2)\Big] \\
&+ \frac{\eta^3 q^2 L^2}{\rho(1-q^2)^2\epsilon}\Big[8\eta_l^2K^2(60\eta_l^2K^2L^2+1)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big] + \frac{16T\eta_l^2K}{n}\sigma^2 + 80T\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2)\Big] \\
&\leq \frac{2\eta^3\eta_l^2C_2K^2L^2}{\rho\epsilon}(60\eta_l^2K^2L^2+1)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \\
&\qquad + \frac{4T\eta^3\eta_l^2C_2KL^2}{\rho n\epsilon}\sigma^2 + \frac{20T\eta^3\eta_l^4C_2K^3L^4}{\rho\epsilon}(\sigma^2+6K\sigma_g^2), \qquad (9)
\end{aligned}$$

with $C_2 := \frac{\beta_1^2}{(1-\beta_1)^2} + \frac{4q^2}{(1-q^2)^2}$.

Putting together. We are in the position to combine pieces together to get our final result by integrating (5), (6), (7), (8) and (9) into (3) and taking the telescoping sum over $t = 1, ..., T$. After re-arranging terms, when $\eta_l \leq \min\big\{\frac{1}{8KL}, \frac{(1-q^2)\sqrt{\epsilon}}{4(1+q^2)^{1.5}KG}\big\}$, we have

$$\begin{aligned}
&\mathbb{E}[f(x_{T+1}) - f(x_1)] \\
&\leq -\frac{\eta\eta_l K}{8\sqrt{\epsilon}}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5T\eta\eta_l^3K^2L^2}{2\sqrt{\epsilon}}(\sigma^2+6K\sigma_g^2) + \eta\eta_l KG^2\sum_{t=1}^{T}\mathbb{E}[\|D_t\|_1] \\
&+ \eta\eta_l C_1KG^2\sum_{t=1}^{T}\mathbb{E}[\|D_t\|_1] + \frac{\eta^2\eta_l^2C_1^2LK^2G^2}{\sqrt{\epsilon}}\sum_{t=1}^{T}\mathbb{E}[\|D_t\|_1] \\
&+ \frac{\eta^2L(2\eta_l^2K^2+120\eta_l^4K^4L^2)}{\epsilon}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{4T\eta^2\eta_l^2KL}{n\epsilon}\sigma^2 \\
&\qquad + \frac{20T\eta^2\eta_l^4K^3L^3}{\epsilon}(\sigma^2+6K\sigma_g^2) + \eta^2\eta_l^2C_1^2LK^2G^2\sum_{t=1}^{T}\mathbb{E}[\|D_t\|^2] \\
&+ \frac{\rho\eta\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{\epsilon}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2T\rho\eta\eta_l^2K}{\epsilon n}\sigma^2 + \frac{10T\rho\eta\eta_l^4K^3L^2}{\epsilon}(\sigma^2+6K\sigma_g^2) \\
&+ \frac{2\eta^3\eta_l^2C_2K^2L^2}{\rho\epsilon}(60\eta_l^2K^2L^2+1)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{4T\eta^3\eta_l^2C_2KL^2}{\rho n\epsilon}\sigma^2 \\
&\qquad + \frac{20T\eta^3\eta_l^4C_2K^3L^4}{\rho\epsilon}(\sigma^2+6K\sigma_g^2) + \frac{\eta^2\eta_l^2C_1LK^2G^2}{\sqrt{\epsilon}}\sum_{t=1}^{T}\mathbb{E}[\|D_t\|_1] \\
&= \Upsilon_1\cdot\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \Upsilon_2\cdot(\sigma^2+6K\sigma_g^2) + \Upsilon_3\cdot\sigma^2 \\
&\qquad + \Upsilon_4\cdot\sum_{t=1}^{T}\mathbb{E}[\|D_t\|_1] + \eta^2\eta_l^2C_1^2LK^2G^2\sum_{t=1}^{T}\mathbb{E}[\|D_t\|^2], \qquad (10)
\end{aligned}$$

where

$$\begin{aligned}\Upsilon_1 &= -\frac{\eta\eta_l K}{8\sqrt{\epsilon}} + \frac{\eta^2 L(2\eta_l^2 K^2 + 120\eta_l^4 K^4 L^2)}{\epsilon} \\ &\qquad + \frac{\rho\eta\eta_l^2 K^2(60\eta_l^2 K^2 L^2 + 1)}{\epsilon} + \frac{2\eta^3\eta_l^2 C_2 K^2 L^2}{\rho\epsilon}(60\eta_l^2 K^2 L^2 + 1) \\ &\le -\frac{\eta\eta_l K}{8\sqrt{\epsilon}} + \frac{2\eta^2\eta_l^2 K^2 L}{\epsilon} + \frac{120\eta^2\eta_l^4 K^4 L^3}{\epsilon} + \frac{2\rho\eta\eta_l^2 K^2}{\epsilon} + \frac{4\eta^3\eta_l^2 C_2 K^2 L^2}{\rho\epsilon}, \qquad (11) \\ \Upsilon_2 &= \frac{5T\eta\eta_l^3 K^2 L^2}{2\sqrt{\epsilon}} + \frac{20T\eta^2\eta_l^4 K^3 L^3}{\epsilon} + \frac{10T\rho\eta\eta_l^4 K^3 L^2}{\epsilon} + \frac{20T\eta^3\eta_l^4 C_2 K^3 L^4}{\rho\epsilon}, \\ \Upsilon_3 &= \frac{4T\eta^2\eta_l^2 KL}{n\epsilon} + \frac{2T\rho\eta\eta_l^2 K}{n\epsilon} + \frac{4T\eta^3\eta_l^2 C_2 KL^2}{\rho n\epsilon}, \\ \Upsilon_4 &= \eta\eta_l(C_1+1)KG^2 + \frac{\eta^2\eta_l^2 C_1^2 LK^2G^2}{\sqrt{\epsilon}} + \frac{\eta^2\eta_l^2 C_1 LK^2G^2}{\sqrt{\epsilon}},\end{aligned}$$

where to bound Υ_1 we use the fact that $\eta_l \le \frac{1}{8KL}$. We now look at the upper bound (11) of Υ_1 which contains 5 terms. In the following, we choose $\rho \equiv L\eta$ in (8) and (9). Suppose $\epsilon < 1$. Then, when the local learning rate satisfies

$$\begin{aligned}\eta_l &\le \frac{1}{K}\min\Big\{\frac{1}{8L}, \frac{(1-q^2)\sqrt{\epsilon}}{4(1+q^2)^{1.5}G}, \frac{\sqrt{\epsilon}}{128\eta L}, \frac{\sqrt{\epsilon}}{256C_2\eta L}, \big(\frac{\sqrt{\epsilon}}{7680\eta}\big)^{1/3}\frac{1}{L}\Big\} \\ &\le \frac{\sqrt{\epsilon}}{8KL}\min\Big\{\frac{1}{\sqrt{\epsilon}}, \frac{2(1-q^2)L}{(1+q^2)^{1.5}G}, \frac{1}{\max\{16, 32C_2\}\eta}, \frac{1}{3\eta^{1/3}}\Big\},\end{aligned}$$

each of the last four terms can be bounded by $\frac{\eta\eta_l K}{64\sqrt{\epsilon}}$. Thus, under this learning rate setting,

$$\Upsilon_1 \le -\frac{\eta\eta_l K}{16\sqrt{\epsilon}}.$$

Taking the above into (10), we arrive at

$$\begin{aligned}\frac{\eta\eta_l K}{16\sqrt{\epsilon}}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] &\le f(x_1) - \mathbb{E}[f(x_{T+1})] + \Upsilon_2\cdot(\sigma^2 + 6K\sigma_g^2) \\ &\qquad + \Upsilon_3\cdot\sigma^2 + \Upsilon_4\cdot\frac{d}{\sqrt{\epsilon}} + \frac{\eta^2\eta_l^2 C_1^2 LK^2G^2 d}{\epsilon},\end{aligned}$$

where Lemma F.8 on the difference sequence D_t is applied. Consequently, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] &\lesssim \frac{f(x_1) - \mathbb{E}[f(x_{T+1})]}{\eta\eta_l TK} + \widetilde{\Upsilon}_2\cdot(\sigma^2 + 6K\sigma_g^2) + \widetilde{\Upsilon}_3\cdot\sigma^2 \\ &\qquad + \frac{(C_1+1)G^2 d}{T\sqrt{\epsilon}} + \frac{2\eta\eta_l C_1^2 LKG^2 d}{T\epsilon} + \frac{\eta\eta_l C_1^2 LKG^2 d}{T\epsilon} \\ &\le \frac{f(x_1) - \mathbb{E}[f(x_{T+1})]}{\eta\eta_l TK} + \widetilde{\Upsilon}_2\cdot(\sigma^2 + 6K\sigma_g^2) + \widetilde{\Upsilon}_3\cdot\sigma^2 \\ &\qquad + \frac{(C_1+1)G^2 d}{T\sqrt{\epsilon}} + \frac{3\eta\eta_l C_1^2 LKG^2 d}{T\epsilon},\end{aligned}$$

where we make simplification at the second inequality using the fact that $C_1 \le C_1^2$ since $C_1 \ge 1$. Moreover, $\widetilde{\Upsilon}_2$ and $\widetilde{\Upsilon}_3$ is defined as (recall that we have chosen $\rho \equiv L\eta$)

$$\begin{aligned}\widetilde{\Upsilon}_2 &= \frac{5\eta_l^2 KL^2}{2\sqrt{\epsilon}} + \frac{20\eta\eta_l^3 K^2 L^3}{\epsilon} + \frac{10\eta\eta_l^3 K^2 L^3}{\epsilon} + \frac{20\eta\eta_l^3 C_2 K^2 L^3}{\epsilon} \\ &\le \frac{5\eta_l^2 KL^2}{2\sqrt{\epsilon}} + \frac{\eta\eta_l^3(30 + 20C_2)K^2 L^3}{\epsilon}, \\ \widetilde{\Upsilon}_3 &= \frac{4\eta\eta_l L}{n\epsilon} + \frac{2\eta\eta_l L}{n\epsilon} + \frac{4\eta\eta_l C_2 L}{n\epsilon} \le \frac{\eta\eta_l L(6 + 4C_2)}{n\epsilon}.\end{aligned}$$

Finally, to connect the virtual iterates x_t with the actual iterates θ_t, note that $x_1 = \theta_1$, and $f(x_{T+1}) \geq f(\theta^*)$ since $\theta^* = \arg\min_\theta f(\theta)$. Replacing $\widetilde{\Upsilon}_2$ and $\widetilde{\Upsilon}_3$ with above upper bounds, this eventually leads to the bound

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \lesssim \frac{f(\theta_1)-f(\theta^*)}{\eta\eta_l TK} + \Big[\frac{5\eta_l^2 KL^2}{2\sqrt{\epsilon}} + \frac{\eta\eta_l^3(30+20C_1^2)K^2L^3}{\epsilon}\Big](\sigma^2+6K\sigma_g^2) + \frac{\eta\eta_l L(6+4C_1^2)}{n\epsilon}\sigma^2 + \frac{(C_1+1)G^2 d}{T\sqrt{\epsilon}} + \frac{3\eta\eta_l C_1^2 LKG^2 d}{T\epsilon},$$

which gives the desired result. Here we use the fact that $C_2 \leq C_1^2$. This completes the proof. □

F.2. Proof of Theorem 4.6: Fed-EF-SGD

Proof. Now, we analyze the variant of Fed-EF with SGD as the central server update rule. The proof follows the same routine as the one for Fed-EF-AMS, but is simpler since there are no moving average terms that need to be handled. Note that for this algorithm, we do not need Assumption 4.7 that the stochastic gradients are uniformly bounded.

For Fed-EF-SGD, consider the virtual sequence

$$\begin{aligned} x_{t+1} &= \theta_{t+1} - \eta\bar{e}_{t+1} \\ &= \theta_t - \eta\bar{\tilde{\Delta}}_t - \eta\bar{e}_{t+1} \\ &= \theta_t - \frac{\eta}{n}\sum_{i=1}^{n}(\tilde{\Delta}_{t,i} + e_{t+1,i}) \\ &= \theta_t - \eta\bar{\Delta}_t - \eta\bar{e}_t \\ &= x_t - \eta\bar{\Delta}_t, \end{aligned} \tag{12}$$

where the second last equality follows from the update rule that $\tilde{\Delta}_{t,i} + e_{t+1,i} = \Delta_{t,i} + e_{t,i}$ for all $i \in [n]$ and $t \in [T]$.

By the smoothness Assumption 4.1, we have

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

Taking expectation w.r.t. the randomness at round t gives

$$\begin{aligned} &\mathbb{E}[f(x_{t+1})] - f(x_t) \\ &\leq -\eta\mathbb{E}\big[\langle \nabla f(x_t), \bar{\Delta}_t\rangle\big] + \frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] \\ &= -\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_t\rangle\big] + \frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] + \eta\mathbb{E}\big[\langle \nabla f(\theta_t) - \nabla f(x_t), \bar{\Delta}_t\rangle\big]. \end{aligned} \tag{13}$$

We can bound the first term in (13) using similar technique as bounding term **I** in the proof of Fed-EF-AMS. Specifically, we have

$$\begin{aligned} -\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_t\rangle\big] &= -\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_t - \eta_l K\nabla f(\theta_t) + \eta_l K\nabla f(\theta_t)\rangle\big] \\ &= -\eta\eta_l K\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \eta\mathbb{E}\big[\langle \nabla f(\theta_t), -\bar{\Delta}_t + \eta_l K\nabla f(\theta_t)\rangle\big]. \end{aligned}$$

With $\eta_l \leq \frac{1}{8KL}$, applying Lemma F.3, we have

$$\begin{aligned} -\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_t\rangle\big] &\leq -\eta\eta_l K\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{3\eta\eta_l K}{4}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5\eta\eta_l^3 K^2L^2}{2}(\sigma^2+6K\sigma_g^2) \\ &= -\frac{\eta\eta_l K}{4}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5\eta\eta_l^3 K^2L^2}{2}(\sigma^2+6K\sigma_g^2). \end{aligned}$$

The second term in (13) can be bounded using Lemma F.2 as

$$\frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] \le \eta^2\eta_l^2 K^2 L(60\eta_l^2K^2L^2+1)\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2\eta^2\eta_l^2 KL}{n}\sigma^2 + 10\eta^2\eta_l^4K^3L^3(\sigma^2+6K\sigma_g^2).$$

The last term in (13) can be bounded similarly as **VI** in Fed-EF-AMS by

$$\begin{aligned}
&\eta\mathbb{E}\big[\langle\nabla f(\theta_t)-\nabla f(x_t),\bar{\Delta}_t\rangle\big]\\
&\le \frac{\eta\rho}{2}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big]+\frac{\eta}{2\rho}\mathbb{E}\big[\|\nabla f(\theta_t)-\nabla f(x_t)\|^2\big]\\
&\overset{(a)}{\le} \frac{\eta^2}{2}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big]+\frac{\eta^2L^2}{2}\mathbb{E}\big[\|\bar{e}_t\|^2\big]\\
&\overset{(b)}{\le} \frac{\eta^2L}{2}\Big[2\eta_l^2K^2(60\eta_l^2K^2L^2+1)\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+4\frac{\eta_l^2K}{n}\sigma^2+20\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2)\Big]\\
&+\frac{\eta^2L}{2}\Big[\frac{4q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big]\\
&\qquad+\frac{16\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2+\frac{80\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2)\Big],
\end{aligned}\tag{14}$$

where (a) uses Young's inequality and (b) uses Lemma F.2 and Lemma F.5. When taking telescoping sum of this term over $t=1,...,T$, again using the geometric summation trick, we further obtain

$$\begin{aligned}
&\eta\sum_{t=1}^{T}\mathbb{E}\big[\langle\nabla f(\theta_t)-\nabla f(x_t),\bar{\Delta}_t\rangle\big]\\
&\le \eta^2\eta_l^2C_1K^2L(60\eta_l^2K^2L^2+1)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]\\
&\qquad+\frac{2T\eta^2\eta_l^2C_1KL}{n}\sigma^2+10T\eta^2\eta_l^4C_1K^3L^3(\sigma^2+6K\sigma_g^2),
\end{aligned}$$

where $C_1=1+\frac{4q^2}{(1-q^2)^2}$. Now, taking the summation over all terms in (13), we get

$$\begin{aligned}
\mathbb{E}[f(x_{t+1})]-f(x_1) \le &\Big(-\frac{\eta\eta_lK}{4}+\eta^2\eta_l^2(C_1+1)K^2L(60\eta_l^2K^2L^2+1)\Big)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]\\
&+\frac{2T\eta^2\eta_l^2C_1KL}{n}\sigma^2+\frac{2T\eta^2\eta_l^2(C_1+1)KL}{n}\sigma^2+10T\eta^2\eta_l^4(C_1+1)K^3L^3(\sigma^2+6K\sigma_g^2).
\end{aligned}$$

Since $\eta_l\le\frac{1}{8KL}$, we know that $60\eta_l^2K^2L^2+1<2$. Therefore, provided that the local learning rate is such that

$$\eta_l\le\frac{1}{2KL\cdot\max\{4,\eta(C_1+1)\}},$$

we have

$$\begin{aligned}
\frac{\eta\eta_lK}{8}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \le f(x_1)-\mathbb{E}[f(x_{t+1})]+\frac{2T\eta^2\eta_l^2(C_1+1)KL}{n}\sigma^2\\
+10T\eta^2\eta_l^4(C_1+1)K^3L^3(\sigma^2+6K\sigma_g^2),
\end{aligned}$$

leading to

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] &\lesssim \frac{f(x_1)-\mathbb{E}[f(x_{t+1})]}{\eta\eta_l TK}+\frac{2\eta\eta_l(C_1+1)L}{n}\sigma^2 \\ &\quad +10\eta\eta_l^3(C_1+1)K^2L^3(\sigma^2+6K\sigma_g^2) \\ &\le \frac{f(\theta_1)-f(\theta^*)}{\eta\eta_l TK}+\frac{2\eta\eta_l(C_1+1)L}{n}\sigma^2 \\ &\quad +10\eta\eta_l^3(C_1+1)K^2L^3(\sigma^2+6K\sigma_g^2),\end{aligned}$$

which concludes the proof. □

F.3. Intermediate Lemmas

In our analysis, we will make use of the following lemma on the consensus error. Note that this is a general result holding for algorithms (both Fed-EF-SGD and Fed-EF-AMS) with local SGD steps.

Lemma F.1 ((Reddi et al., 2021)). *For $\eta_l \le \frac{1}{8LK}$, for any round t, local step $k \in [K]$ and client $i \in [n]$, under Assumption 4.1 to Assumption 4.2, it holds that*

$$\mathbb{E}\big[\|\theta_{t,i}^{(k)}-\theta_t\|^2\big] \le 5\eta_l^2K(\sigma^2+6K\sigma_g^2)+30\eta_l^2K^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big].$$

We then state some results that bound several key ingredients in our analysis.

Lemma F.2. *Recall $\bar{\Delta}_t=\frac{1}{n}\sum_{i=1}^n\Delta_{t,i}$. Under Assumption 4.1 to Assumption 4.2, for $\forall t$, the following bounds hold:*

1. Bound by local gradients:

$$\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] \le \frac{\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^n\sum_{k=1}^K\nabla f_i(\theta_{t,i}^{(k)})\|^2\big]+\frac{\eta_l^2K}{n}\sigma^2.$$

2. Bound by global gradient:

$$\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] \le (2\eta_l^2K^2+120\eta_l^4K^4L^2)\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]+4\frac{\eta_l^2K}{n}\sigma^2+20\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2).$$

Proof. By definition, we have

$$\begin{aligned}\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] &= \mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K\eta_l g_{t,i}^{(k)}\|^2\big] \\ &\le \frac{\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^n\sum_{k=1}^K(g_{t,i}^{(k)}-\nabla f_i(\theta_{t,i}^{(k)})\|^2\big]+\frac{\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^n\sum_{k=1}^K\nabla f_i(\theta_{t,i}^{(k)})\|^2\big] \\ &\le \frac{\eta_l^2K}{n}\sigma^2+\frac{\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^n\sum_{k=1}^K\nabla f_i(\theta_{t,i}^{(k)})\|^2\big],\end{aligned}$$

where the second line is due to the variance decomposition, and the last inequality uses Assumption 4.2 on independent and

unbiased stochastic gradients. This proves the first part. For the second part,

$$\begin{aligned}
\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] &= \mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\eta_l g_{t,i}^{(k)} - K\eta_l\nabla f(\theta_t) + K\eta_l\nabla f(\theta_t)\|^2\big] \\
&\le 2\eta_l^2K^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + 2\eta_l^2\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}g_{t,i}^{(k)} - \frac{K}{n}\sum_{i=1}^{n}\nabla f_i(\theta_t)\|^2\big] \\
&= 2\eta_l^2K^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}(g_{t,i}^{(k)} - \nabla f_i(\theta_t))\|^2\big] \\
&\le 2\eta_l^2K^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}(g_{t,i}^{(k)} - \nabla f_i(\theta_{t,i}^{(k)}) + \nabla f_i(\theta_{t,i}^{(k)}) - \nabla f_i(\theta_t))\|^2\big] \\
&\le 2\eta_l^2K^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{2\eta_l^2}{n^2}\underbrace{\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}(g_{t,i}^{(k)} - \nabla f_i(\theta_{t,i}^{(k)}) + \nabla f_i(\theta_{t,i}^{(k)}) - \nabla f_i(\theta_t))\|^2\big]}_{A}.
\end{aligned}$$

The expectation A can be further bounded as

$$\begin{aligned}
A &\le 2\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}(g_{t,i}^{(k)} - \nabla f_i(\theta_{t,i}^{(k)}))\|^2\big] + 2\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}(\nabla f_i(\theta_{t,i}^{(k)}) - \nabla f_i(\theta_t))\|^2\big] \\
&\overset{(a)}{\le} 2nK\sigma^2 + 2nK\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\big[\|\nabla f_i(\theta_{t,i}^{(k)}) - \nabla f_i(\theta_t)\|^2\big] \\
&\overset{(b)}{\le} 2nK\sigma^2 + 2nKL^2\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\big[\|\theta_{t,i}^{(k)} - \theta_t\|^2\big] \\
&\overset{(c)}{\le} 60\eta_l^2n^2K^4L^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + 2nK\sigma^2 + 10\eta_l^2n^2K^3L^2(\sigma^2 + 6K\sigma_g^2),
\end{aligned}$$

where (a) is implied by Assumption 4.2 that each local stochastic gradient $g_{t,i}^{(k)}$ can be written as $g_{t,i}^{(k)} = \nabla f_i(\theta_{t,i}^{(k)}) + \xi_{t,i}^{(k)}$, where $\xi_{t,i}^{k}$ is a zero-mean random noise with bounded variance σ^2, and all the noises for $t \in [T], i \in [n], k \in [K]$ are independent. The inequality (b) is due to the smoothness Assumption 4.1, and (c) follows from Lemma F.1. Therefore, we obtain

$$\mathbb{E}[\|\bar{\Delta}_t\|^2] \le (2\eta_l^2K^2 + 120\eta_l^4K^4L^2)\mathbb{E}[\|\nabla f(\theta_t)\|^2] + 4\frac{\eta_l^2K}{n}\sigma^2 + 20\eta_l^4K^3L^2(\sigma^2 + 6K\sigma_g^2),$$

which completes the proof of the second claim. □

Lemma F.3. *Under Assumption 4.1 and Assumption 4.2, when $\eta_l \le \frac{1}{8KL}$, Fed-EF-SGD admits*

$$\mathbb{E}\big[\langle\nabla f(\theta_t), -\bar{\Delta}_t + \eta_lK\nabla f(\theta_t)\rangle\big] \le \frac{3\eta_lK}{4}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5\eta_l^3K^2L^2}{2}(\sigma^2 + 6K\sigma_g^2).$$

Proof. It holds that

$$
\begin{aligned}
&\mathbb{E}\big[\langle \nabla f(\theta_t), -\bar{\Delta}_t + \eta_l K \nabla f(\theta_t)\rangle\big] \\
&= \big\langle \nabla f(\theta_t), \mathbb{E}\big[-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \eta_l g_{t,i}^{(k)} + \eta_l K \nabla f(\theta_t)\big]\big\rangle \\
&= \big\langle \sqrt{\eta_l}\nabla f(\theta_t), \sqrt{\eta_l}\mathbb{E}\big[-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \nabla f_i(\theta_t^{(k)}) + K\nabla f(\theta_t)\big]\big\rangle \\
&\overset{(a)}{\leq} \frac{\eta_l K}{2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{\eta_l}{2K}\mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(\nabla f_i(\theta_{t,i}^{(k)}) - \nabla f_i(\theta_t))\|^2\big] \\
&\leq \frac{\eta_l K}{2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{\eta_l}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\big[\|\nabla f_i(\theta_{t,i}^{(k)}) - \nabla f_i(\theta_t)\|^2\big] \\
&\overset{(b)}{\leq} \frac{\eta_l K}{2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{\eta_l L^2}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\big[\|\theta_{t,i}^{(k)} - \theta_t\|^2\big] \\
&\overset{(c)}{\leq} \frac{\eta_l K}{2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{\eta_l K L^2}{2}\Big[5\eta_l^2 K(\sigma^2 + 6K\sigma_g^2) + 30\eta_l^2 K^2 \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\Big]
\end{aligned}
$$

where (a) is due to $\langle a, b\rangle \leq \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2$ for any $a, b \in \mathbb{R}$ and $\alpha > 0$, , (b) is a consequence of Assumption 4.1, and (c) is due to Lemma F.1. If $\eta_l \leq \frac{1}{8KL}$, we have that $\eta_l^2 \leq \frac{1}{64K^2L^2}$, bounding the last term by $\frac{15}{64}\eta_l K \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]$. Hence, we obtain

$$
\mathbb{E}\big[\langle \nabla f(\theta_t), -\bar{\Delta}_t + \eta_l K \nabla f(\theta_t)\rangle\big] \leq \frac{47\eta_l K}{64}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5\eta_l^3 K^2 L^2}{2}(\sigma^2 + 6K\sigma_g^2),
$$

where the proof is completed since $\frac{47}{64} < \frac{3}{4}$. □

Lemma F.4. *Under Assumption 4.1 to Assumption 4.2 we have:*

$$
\begin{aligned}
&\|m_t'\| \leq \eta_l K G, \quad \textit{for } \forall t, \\
&\sum_{t=1}^{T}\mathbb{E}\big[\|m_t'\|^2\big] \leq (2\eta_l^2 K^2 + 120\eta_l^4 K^4 L^2)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \\
&\qquad\qquad + 4\frac{T\eta_l^2 K}{n}\sigma^2 + 20T\eta_l^4 K^3 L^2(\sigma^2 + 6K\sigma_g^2).
\end{aligned}
$$

Proof. For the first part, by Assumption 4.7 we know that

$$
\begin{aligned}
\|m_t'\| &= (1-\beta_1)\|\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{\Delta}_t\| \\
&= (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\frac{\eta_l}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\|g_{t,i}^{(k)}\| \\
&\leq \eta_l K G.
\end{aligned}
$$

For the second claim, by Lemma F.2 we know that

$$
\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] \leq (2\eta_l^2 K^2 + 120\eta_l^4 K^4 L^2)\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + 4\frac{\eta_l^2 K}{n}\sigma^2 + 20\eta_l^4 K^3 L^2(\sigma^2 + 6K\sigma_g^2).
$$

Let $\bar{\Delta}_{t,j}$ denote the j-th coordinate of $\bar{\Delta}_t$. By the updating rule of Fed-EF, we have

$$\begin{aligned}
\mathbb{E}\big[\|m_t'\|^2\big] &= \mathbb{E}\big[\|(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{\Delta}_\tau\|^2\big] \\
&\le (1-\beta_1)^2\sum_{j=1}^{d}\mathbb{E}\big[(\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{\Delta}_{\tau,j})^2\big] \\
&\overset{(a)}{\le} (1-\beta_1)^2\sum_{j=1}^{d}\mathbb{E}\big[(\sum_{\tau=1}^{t}\beta_1^{t-\tau})(\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{\Delta}_{\tau,j}^2)\big] \\
&\le (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\mathbb{E}\big[\|\bar{\Delta}_\tau\|^2\big] \\
&\le (2\eta_l^2K^2+120\eta_l^4K^4L^2)(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \\
&\qquad + 4\frac{\eta_l^2K}{n}\sigma^2 + 20\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2),
\end{aligned}$$

where (a) is due to Cauchy-Schwartz inequality. Summing over $t=1,...,T$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\big[\|m_t'\|^2\big] &\le (2\eta_l^2K^2+120\eta_l^4K^4L^2)(1-\beta)\sum_{t=1}^{T}\sum_{\tau=1}^{t}\beta_1^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \\
&\qquad + 4\frac{T\eta_l^2K}{n}\sigma^2 + 20T\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2) \\
&\le (2\eta_l^2K^2+120\eta_l^4K^4L^2)\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \\
&\qquad + 4\frac{T\eta_l^2K}{n}\sigma^2 + 20T\eta_l^4K^3L^2(\sigma^2+6K\sigma_g^2),
\end{aligned}$$

which concludes the proof. □

Lemma F.5. *Under Assumption 4.2, we have for $\forall t$ and each local worker $\forall i \in [n]$,*

$$\begin{aligned}
&\|e_{t,i}\|^2 \le \frac{4\eta_l^2q^2K^2G^2}{(1-q^2)^2},\ \forall t, \\
&\mathbb{E}[\|\bar{e}_{t+1}\|^2] \le \frac{4q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big] \\
&\qquad\qquad + \frac{16\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2).
\end{aligned}$$

Proof. To prove the second claim, we start by using Assumption 4.3 and Young's inequality to get

$$\begin{aligned}
\|\bar{e}_{t+1}\|^2 &= \|\bar{\Delta}_t+\bar{e}_t-\frac{1}{n}\sum_{i=1}^{n}\mathcal{C}(\Delta_{t,i}+e_{t,i})\|^2 \\
&\le q^2\|\bar{\Delta}_t+\bar{e}_t\|^2 \\
&\le q^2(1+\rho)\|\bar{e}_t\|^2 + q^2(1+\frac{1}{\rho})\|\bar{\Delta}_{t,i}\|^2 \\
&\le \frac{1+q^2}{2}\|\bar{e}_t\|^2 + \frac{2q^2}{1-q^2}\|\bar{\Delta}_t\|^2, \qquad (15)
\end{aligned}$$

where (15) is derived by choosing $\rho = \frac{1-q^2}{2q^2}$ and the fact that $q < 1$. Now by recursion and the initialization $e_{1,i} = 0, \forall i$, we have

$$\begin{aligned}\mathbb{E}\big[\|\bar{e}_{t+1}\|^2\big] &\leq \frac{2q^2}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\bar{\Delta}_\tau\|^2\big] \\ &\leq \frac{4q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big] \\ &\quad + \frac{16\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2),\end{aligned}$$

which proves the second argument, where we use Lemma F.1 to bound the local model update. In addition, we know that $\|\Delta_t\| \leq \eta_l KG$ by Assumption 4.7 for any t.

The absolute bound $\|e_{t,i}\|^2 \leq \frac{4\eta_l^2q_{\mathcal{C}}^2K^2G^2}{(1-q_{\mathcal{C}}^2)^2}$ follows from (15) by a similar recursion argument used on local error $e_{t,i}$, and the fact that $q_{\mathcal{C}} \leq \max\{q_{\mathcal{C}}, q_{\mathcal{A}}\} = q$. □

Lemma F.6. *For the moving average error sequence $\mathcal{E}_t$, it holds that*

$$\begin{aligned}&\|\mathcal{E}_t\|^2 \leq \frac{4\eta_l^2q^2K^2G^2}{(1-q^2)^2}, \quad \textit{for } \forall t, \\ &\sum_{t=1}^{T}\mathbb{E}\big[\|\mathcal{E}_t\|^2\big] \leq \frac{8q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{(1-q^2)^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big] \\ &\qquad + \frac{16T\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80T\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2).\end{aligned}$$

Proof. The first argument can be easily deduced by the definition of $\mathcal{E}_t$ that

$$\begin{aligned}\|\mathcal{E}_t\| &= (1-\beta_1)\|\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{e}_t\| \\ &\leq \|e_{t,i}\| \leq \frac{2\eta_l qKG}{1-q^2}.\end{aligned}$$

Denote the quantity

$$K_t := \sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big].$$

By the same technique as in the proof of Lemma F.4, denoting $\bar{e}_{t,j}$ as the j-th coordinate of $\bar{e}_t$, we can bound the accumulated

error sequence by

$$\begin{aligned}
\mathbb{E}\big[\|\mathcal{E}_t\|^2\big] &= \mathbb{E}\big[\|(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{e}_\tau\|^2\big] \\
&\le (1-\beta_1)^2\sum_{j=1}^{d}\mathbb{E}\big[(\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{e}_{\tau,j})^2\big] \\
&\overset{(a)}{\le} (1-\beta_1)^2\sum_{j=1}^{d}\mathbb{E}\big[(\sum_{\tau=1}^{t}\beta_1^{t-\tau})(\sum_{\tau=1}^{t}\beta_1^{t-\tau}\bar{e}_{\tau,j}^2)\big] \\
&\le (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\mathbb{E}\big[\|\bar{e}_\tau\|^2\big] \\
&\overset{(b)}{\le} \frac{16\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2) \\
&\qquad + \frac{4(1-\beta_1)q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{1-q^2}\sum_{\tau=1}^{t}\beta_1^{t-\tau}K_\tau,
\end{aligned}$$

where (a) is due to Cauchy-Schwartz inequality and (b) is a result of Lemma F.5. Summing over $t = 1, ..., T$ and using the technique of geometric series summation leads to

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\big[\|\mathcal{E}_t\|^2\big] &\le \frac{16T\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80T\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2) \\
&\qquad + \frac{4(1-\beta_1)q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{1-q^2}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\beta_1^{t-\tau}K_\tau \\
&\le \frac{16T\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80T\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2) \\
&\qquad + \frac{4q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{1-q^2}\sum_{t=1}^{T}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big] \\
&\le \frac{16T\eta_l^2q^2K}{(1-q^2)^2n}\sigma^2 + \frac{80T\eta_l^4q^2K^3L^2}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2) \\
&\qquad + \frac{8q^2\eta_l^2K^2(60\eta_l^2K^2L^2+1)}{(1-q^2)^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big].
\end{aligned}$$

The desired result is obtained. □

Lemma F.7. *It holds that* $\forall t \in [T]$, $\forall i \in [d]$, $\hat{v}_{t,i} \le \frac{4\eta_l^2(1+q^2)^3K^2}{(1-q^2)^2}G^2$.

Proof. For any t, by Lemma F.5 and Assumption 4.7 we have

$$\begin{aligned}
\|\widetilde{\Delta}_t\|^2 &= \|\mathcal{C}(\Delta_t+e_t)\|^2 \\
&\le \|\mathcal{C}(\Delta_t+e_t)-(\Delta_t+e_t)+(\Delta_t+e_t)\|^2 \\
&\le 2(q^2+1)\|\Delta_t+e_t\|^2 \\
&\le 4(q^2+1)(\eta_l^2K^2G^2+\frac{4\eta_l^2q^2K^2G^2}{(1-q^2)^2}) \\
&= \frac{4\eta_l^2(1+q^2)^3K^2}{(1-q^2)^2}G^2.
\end{aligned}$$

Consider the updating rule of $\hat{v}_t = \max\{v_t, \hat{v}_{t-1}\}$. We know that there exists a $j \in [t]$ such that $\hat{v}_t = v_j$. Thus, we have

$$\hat{v}_{t,i} = (1-\beta_2)\sum_{\tau=1}^{j}\beta_2^{j-\tau}\tilde{g}_{t,i}^2 \le \frac{4\eta_l^2(1+q^2)^3K^2}{(1-q^2)^2}G^2,$$

which proves the claim. □

The next Lemma is analogue to Lemma 5 in Li et al. (2022b).

Lemma F.8. *Let* $D_t := \frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}} - \frac{1}{\sqrt{\hat{v}_t+\epsilon}}$ *be defined as above. Then,*

$$\sum_{t=1}^{T}\|D_t\|_1 \le \frac{d}{\sqrt{\epsilon}}, \quad \sum_{t=1}^{T}\|D_t\|^2 \le \frac{d}{\epsilon}.$$

Proof. By the updating rule of Fed-EF-AMS, $\hat{v}_{t-1} \le \hat{v}_t$ for $\forall t$. Therefore, by the initialization $\hat{v}_0 = 0$, we have

$$\begin{aligned}\sum_{t=1}^{T}\|D_t\|_1 &= \sum_{t=1}^{T}\sum_{i=1}^{d}\Big(\frac{1}{\sqrt{\hat{v}_{t-1,i}+\epsilon}} - \frac{1}{\sqrt{\hat{v}_{t,i}+\epsilon}}\Big)\\ &= \sum_{i=1}^{d}\Big(\frac{1}{\sqrt{\hat{v}_{0,i}+\epsilon}} - \frac{1}{\sqrt{\hat{v}_{T,i}+\epsilon}}\Big)\\ &\le \frac{d}{\sqrt{\epsilon}}.\end{aligned}$$

For the sum of squared l_2 norm, note the fact that for $a \ge b > 0$, it holds that

$$(a-b)^2 \le (a-b)(a+b) = a^2 - b^2.$$

Thus,

$$\begin{aligned}\sum_{t=1}^{T}\|D_t\|^2 &= \sum_{t=1}^{T}\sum_{i=1}^{d}\Big(\frac{1}{\sqrt{\hat{v}_{t-1,i}+\epsilon}} - \frac{1}{\sqrt{\hat{v}_{t,i}+\epsilon}}\Big)^2\\ &\le \sum_{t=1}^{T}\sum_{i=1}^{d}\Big(\frac{1}{\hat{v}_{t-1,i}+\epsilon} - \frac{1}{\hat{v}_{t,i}+\epsilon}\Big)\\ &\le \frac{d}{\epsilon},\end{aligned}$$

which gives the desired result. □

F.4. Proof of Theorem 4.10: Partial Participation

Proof. We can use a similar proof structure as previous analysis for full participation Fed-EF-SGD as in Section F.2. Like before, we first define the following virtual iterates:

$$\begin{aligned}x_{t+1} = \theta_{t+1} - \eta\frac{1}{m}\sum_{i=1}^{n}e_{t+1,i} &= \theta_t - \eta\overline{\tilde{\Delta}}_{t,\mathcal{M}_t} - \eta\frac{1}{m}\sum_{i\in\mathcal{M}_t}e_{t+1,i} - \eta\frac{1}{m}\sum_{i\notin\mathcal{M}_t}e_{t+1,i}\\ &= \theta_t - \eta\bar{\Delta}_{t,\mathcal{M}_t} - \eta\frac{1}{m}\sum_{i\in\mathcal{M}_t}e_{t,i} - \eta\frac{1}{m}\sum_{i\notin\mathcal{M}_t}e_{t,i} \qquad (16)\\ &= \theta_t - \eta\bar{\Delta}_{t,\mathcal{M}_t} - \eta\frac{1}{m}\sum_{i=1}^{n}e_{t,i}\\ &= x_t - \eta\bar{\Delta}_{t,\mathcal{M}_t}.\end{aligned}$$

Here, (16) follows from the partial participation setup where there is no error accumulation for inactive clients. The smoothness of loss functions implies

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

Taking expectation w.r.t. the randomness at round t, we have

$$\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - f(x_t) \\
&\leq -\eta \mathbb{E}\big[\langle \nabla f(x_t), \bar{\Delta}_{t,\mathcal{M}_t} \rangle\big] + \frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_{t,\mathcal{M}_t}\|^2\big] \\
&= \underbrace{-\eta \mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_{t,\mathcal{M}_t} \rangle\big]}_{I} + + \underbrace{\frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_{t,\mathcal{M}_t}\|^2\big]}_{II} + \underbrace{\eta \mathbb{E}\big[\langle \nabla f(x_t) - \nabla f(\theta_t), \bar{\Delta}_{t,\mathcal{M}_t} \rangle\big]}_{III}.
\end{aligned} \tag{17}$$

Note that the expectation is also with respect to the randomness in the client sampling procedure. For the first term, we may adopt the similar idea of the proof of Lemma F.3. Since the client sampling is random, we have $\mathbb{E}[\bar{\Delta}_{t,\mathcal{M}_t}] = \mathbb{E}[\bar{\Delta}_t]$. Thus,

$$\begin{aligned}
\boldsymbol{I} &= -\eta \mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_{t,\mathcal{M}_t} \rangle\big] \\
&= -\eta \mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_t - \eta_l K \nabla f(\theta_t) + \eta_l K \nabla f(\theta_t) \rangle\big] \\
&= -\eta\eta_l K \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \eta\eta_l \big\langle \sqrt{K}\nabla f(\theta_t), -\frac{1}{n\sqrt{K}} \sum_{i=1}^{n}\sum_{k=1}^{K} \big(\nabla f_i(\theta_{t,i}^{(k)}) + \nabla f(\theta_t)\big) \big\rangle \\
&\overset{(a)}{=} -\eta\eta_l K \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \eta\eta_l \mathbb{E}\Big[\frac{K}{2}\|\nabla f(\theta_t)\|^2 \\
&\qquad + \frac{1}{2Kn^2}\|\sum_{i=1}^{n}\sum_{k=1}^{K}(\nabla f_i(\theta_{t,i}^{(k)}) + \nabla f(\theta_t))\|^2 - \frac{1}{2Kn^2}\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|^2\Big] \\
&\overset{(b)}{\leq} -\eta\eta_l K \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{\eta\eta_l K}{2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \\
&\quad + \frac{\eta\eta_l K L^2}{2}\Big[5\eta_l^2 K(\sigma^2 + 6K\sigma_g^2) + 30\eta_l^2 K^2 \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]\Big] - \frac{\eta\eta_l}{2Kn^2}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|^2\big] \\
&\leq -\frac{\eta\eta_l K}{4}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5\eta\eta_l^3 K^2 L^2}{2}(\sigma^2 + 6K\sigma_g^2) - \frac{\eta\eta_l}{2Kn^2}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|^2\big],
\end{aligned}$$

when $\eta_l \leq \frac{1}{8KL}$, where (a) is a result of the fact that $\langle z_1, z_2 \rangle = \|z_1\|^2 + \|z_2\|^2 - \|z_1 - z_2\|^2$, (b) is because of Lemma F.1.

For term II, we have

$$\begin{aligned}
\boldsymbol{II} &\leq \frac{\eta^2\eta_l^2 KL}{2m}\sigma^2 + \frac{\eta^2\eta_l^2 L}{2n(n-1)}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|\big]^2 \\
&\quad + C'\Big[\frac{3\eta^2\eta_l^2 K^2 L(30\eta_l^2 K^2 L^2 + 1)}{2m}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{15\eta^2\eta_l^4 K^3 L^3}{2m}(\sigma^2 + 6K\sigma_g^2) + \frac{3\eta^2\eta_l^2 K^2 L}{2m}\sigma_g^2\Big],
\end{aligned}$$

with $C' = \frac{n-m}{n-1}$. Furthermore, we have that

$$\boldsymbol{III} \leq 2\eta^2 L \mathbb{E}\big[\|\frac{1}{m}\sum_{i=1}^{n} e_{t,i}\|^2\big] + \frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_{t,\mathcal{M}_t}\|^2\big].$$

The second term is the same as term II. We now bound the first term. Denote $\tilde{e}_{t,i} = e_{t,i} + \Delta_{t,i} - \widetilde{\Delta}_{t,i}$, we have

$$\mathbb{E}\big[\|\frac{1}{m}\sum_{i=1}^n e_{t+1,i}\|^2\big] = \frac{1}{m^2}\mathbb{E}_{\boldsymbol{e}_t}\Big[\underbrace{\mathbb{E}_{\mathcal{M}_t}\big[\|\sum_{i=1}^n \mathbb{1}\{i\in\mathcal{M}_t\}\tilde{e}_{t,i} + \sum_{i=1}^n \mathbb{1}\{i\notin\mathcal{M}_t\}e_{t,i}\|^2|\boldsymbol{e}_t\big]}_{A}\Big].$$

By the updating rule of $e_{t,i}$, the inner expectation, conditional on $\boldsymbol{e}_t = (e_{t,1},...,e_{t,n})^T$, can be computed as

$$\begin{aligned}
A =& \frac{m}{n}\sum_{i=1}^n \|\tilde{e}_{i,t}\|^2 + \frac{m(m-1)}{n(n-1)}\sum_{i\neq j}^n \tilde{e}_{t,i}\tilde{e}_{t,j} + \frac{n-m}{n}\sum_{i=1}^n \|e_{i,t}\|^2 \\
&+ \frac{(n-m)(n-m-1)}{n(n-1)}\sum_{i\neq j}^n e_{t,i}e_{t,j} + \frac{m(n-m)}{n(n-1)}\sum_{i\neq j}^n \tilde{e}_{t,i}e_{t,j} \\
=& \frac{m}{n}\|\sum_{i=1}^n \tilde{e}_{t,i}\|^2 - \frac{m(n-m)}{n(n-1)}\sum_{i\neq j}^n \tilde{e}_{t,i}\tilde{e}_{t,j} + \frac{n-m}{n}\|\sum_{i=1}^n e_{i,t}\|^2 \\
&- \frac{m(n-m)}{n(n-1)}\sum_{i\neq j}^n e_{t,i}e_{t,j} + \frac{m(n-m)}{n(n-1)}\sum_{i\neq j}^n \tilde{e}_{t,i}e_{t,j} \\
=& \frac{m}{n}\|\sum_{i=1}^n \tilde{e}_{t,i}\|^2 + \frac{n-m}{n}\|\sum_{i=1}^n e_{i,t}\|^2 - \frac{m(n-m)}{n(n-1)}\sum_{i\neq j}^n (\tilde{e}_{t,i}\tilde{e}_{t,j} + e_{t,i}e_{t,j} - \tilde{e}_{t,i}e_{t,j}) \\
=& \frac{m}{n}\|\sum_{i=1}^n \tilde{e}_{t,i}\|^2 + \frac{n-m}{n}\|\sum_{i=1}^n e_{i,t}\|^2 \\
&- \frac{m(n-m)}{n(n-1)}\|\sum_{i=1}^n (\tilde{e}_{t,i} - e_{t,i})\|^2 + \frac{m(n-m)}{n(n-1)}\sum_{i=1}^n (\|\tilde{e}_{t,i}\|^2 + \|e_{t,i}\|^2) \\
\leq& \frac{m}{n}\|\sum_{i=1}^n \tilde{e}_{t,i}\|^2 + \frac{n-m}{n}\|\sum_{i=1}^n e_{t,i}\|^2 + \frac{m(n-m)}{n(n-1)}\sum_{i=1}^n (\|\tilde{e}_{t,i}\|^2 + \|e_{t,i}\|^2).
\end{aligned}$$

Therefore, by Definition 3.1 and Assumption 4.3 we obtain

$$\begin{aligned}
&\mathbb{E}\big[\|\frac{1}{m}\sum_{i=1}^n e_{t+1,i}\|^2\big] \\
&\leq \frac{mq^2}{n}\mathbb{E}[\|\frac{1}{m}\sum_{i\in\mathcal{G}}(e_{t,i}+\Delta_{t,i})\|^2] + \frac{n-m}{n}\mathbb{E}[\|\frac{1}{m}\sum_{i=1}^n e_{t,i}\|^2] \\
&\qquad + \frac{(n-m)}{mn(n-1)}\sum_{i=1}^n ((2q^2+1)\|e_{t,i}\|^2 + 2q^2\|\Delta_{t,i}\|^2) \\
&\leq \frac{m(1+\gamma)q^2 + (n-m)}{n}\mathbb{E}[\|\frac{1}{m}\sum_{i=1}^n e_{t,i}\|^2 + \frac{m(1+1/\gamma)q^2}{n}\mathbb{E}[\|\frac{1}{m}\sum_{i=1}^n \Delta_{t,i}\|^2] \\
&\qquad + \frac{(n-m)}{mn(n-1)}\sum_{i=1}^n ((2q^2+1)\|e_{t,i}\|^2 + 2q^2\|\Delta_{t,i}\|^2).
\end{aligned}$$

We have by Lemma F.1 that

$$\begin{aligned}
\mathbb{E}[\|e_{t,i}\|^2] &\leq \frac{20q^2\eta_l^2 K}{(1-q^2)^2}(\sigma^2 + 6K\sigma_g^2) + \frac{60\eta_l^2 q^2 K^2}{1-q^2}\sum_{\tau=1}^t (\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big], \\
\mathbb{E}[\|\Delta_{t,i}\|^2] &\leq 5\eta_l^2 K(\sigma^2 + 6K\sigma_g^2) + 30\eta_l^2 K^2\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big],
\end{aligned}$$

which implies

$$\begin{aligned}
&\frac{(n-m)}{mn(n-1)}\sum_{i=1}^{n}((2q^2+1)\|e_{t,i}\|^2+2q^2\|\Delta_{t,i}\|^2)\\
&\le \frac{(n-m)}{m(n-1)}\Big[\frac{70q^2\eta_l^2K}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2)+\frac{180\eta_l^2q^2K^2}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big]\\
&\qquad+\frac{60\eta_l^2q^2K^2}{(1-q^2)^2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]\Big].
\end{aligned}$$

Recall $q=\max\{q_{\mathcal{A}},q_{\mathcal{C}}\}$. Let $\gamma=(1-q^2)/2q^2$. We have

$$\begin{aligned}
&\frac{m(1+\gamma)q^2+(n-m)}{n}=1-\frac{(1-q^2)m}{2n}<1,\\
&\frac{m(1+1/\gamma)q^2}{n}=\frac{m(1+q^2)q^2}{n(1-q^2)}\le\frac{2mq^2}{n(1-q^2)}.
\end{aligned}$$

By the recursion argument used before, applying Lemma F.2 (adjusted by an n^2/m^2 factor) we obtain

$$\begin{aligned}
&\mathbb{E}\big[\|\frac{1}{m}\sum_{i=1}^{n}e_{t+1,i}\|^2\big]\\
&\le\frac{2mq^2}{n(1-q^2)}\sum_{\tau=1}^{t}\big(1-\frac{(1-q^2)m}{2n}\big)^{t-\tau}\frac{n^2}{m^2}\big[\frac{\eta_l^2}{n^2}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{\tau,i}^{(k)})\|^2\big]+\frac{\eta_l^2K}{n}\sigma^2\big]\\
&+\frac{2n(n-m)}{(1-q^2)m^2(n-1)}\Big[\frac{70q^2\eta_l^2K}{(1-q^2)^2}(\sigma^2+6K\sigma_g^2)+\frac{180\eta_l^2q^2K^2}{1-q^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big]\\
&\qquad+\frac{60\eta_l^2q^2K^2}{(1-q^2)^2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big]\Big]\\
&\le\frac{2\eta_l^2q^2}{(1-q^2)mn}\sum_{\tau=1}^{t}\big(1-\frac{(1-q^2)m}{2n}\big)^{t-\tau}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{\tau,i}^{(k)})\|^2\big]+\frac{4\eta_l^2q^2Kn}{(1-q^2)^2m^2}\sigma^2\\
&\quad+\frac{280\eta_l^2q^2(n-m)K}{(1-q^2)^3m^2}(\sigma^2+6K\sigma_g^2)+\frac{720\eta_l^2q^2(n-m)K^2}{(1-q^2)^2m^2}\sum_{\tau=1}^{t}(\frac{1+q^2}{2})^{t-\tau}\mathbb{E}\big[\|\nabla f(\theta_\tau)\|^2\big]\\
&\qquad+\frac{240\eta_l^2q^2(n-m)K^2}{(1-q^2)^3m^2}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big].
\end{aligned}$$

Summing over $t=1,...,T$ gives

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\big[\|\frac{1}{m}\sum_{i=1}^{n}e_{t+1,i}\|^2\big]\\
&\le\frac{4\eta_l^2q^2}{(1-q^2)^2m^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|^2\big]+\frac{4T\eta_l^2q^2Kn}{(1-q^2)^2m^2}\sigma^2\\
&\quad+\frac{280T\eta_l^2q^2(n-m)K}{(1-q^2)^3m^2}(\sigma^2+6K\sigma_g^2)+\frac{1680\eta_l^2q^2(n-m)K^2}{(1-q^2)^3m^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big].
\end{aligned}$$

Now we turn back to (17). By taking the telescoping sum, we have

$$
\begin{aligned}
&\mathbb{E}[f(x_{t+1})] - f(x_1) \\
&\leq -\frac{\eta\eta_l K}{4}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5T\eta\eta_l^3K^2L^2}{2}(\sigma^2+6K\sigma_g^2) - \frac{\eta\eta_l}{2Kn^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|^2\big] \\
&\qquad + \frac{T\eta^2\eta_l^2KL}{m}\sigma^2 + \frac{\eta^2\eta_l^2L}{n(n-1)}\sum_{t=1}^{T}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|\big]^2 \\
&+ \frac{3\eta^2\eta_l^2C'K^2L(30\eta_l^2K^2L^2+1)}{m}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{15T\eta^2\eta_l^4C'K^3L^3}{m}(\sigma^2+6K\sigma_g^2) + \frac{3T\eta^2\eta_l^2C'K^2L}{m}\sigma_g^2 \\
&+ \frac{8\eta^2\eta_l^2q^2L}{(1-q^2)^2m^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\sum_{i=1}^{n}\sum_{k=1}^{K}\nabla f_i(\theta_{t,i}^{(k)})\|^2\big] + \frac{8T\eta^2\eta_l^2q^2KLn}{(1-q^2)^3m^2}\sigma^2 \\
&\qquad + \frac{560T\eta^2\eta_l^2q^2(n-m)KL}{(1-q^2)^3m^2}(\sigma^2+6K\sigma_g^2) + \frac{3360\eta^2\eta_l^2q^2(n-m)K^2L}{(1-q^2)^3m^2}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big].
\end{aligned}
$$

When the learning rate is chosen such that

$$
\eta_l \leq \min\Big\{\frac{1}{6}, \frac{m}{96C'\eta}, \frac{m^2}{53760(n-m)C_1\eta}, \frac{1}{4\eta}, \frac{1}{32C_1\eta}\Big\}\frac{1}{KL},
$$

we can get

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \lesssim \frac{f(\theta_1)-f(\theta^*)}{\eta\eta_l TK} + \Big[\frac{\eta\eta_l L}{m} + \frac{8\eta\eta_l C_1 Ln}{m^2}\Big]\sigma^2 + \frac{3\eta\eta_l C'KL}{m}\sigma_g^2 \\
+ \Big[\frac{5\eta_l^2KL^2}{2} + \frac{15\eta\eta_l^3C'K^2L^3}{m} + \frac{560\eta\eta_l C_1(n-m)L}{m^2}\Big](\sigma^2+6K\sigma_g^2),
\end{aligned}
$$

where $C_1 = q^2/(1-q^2)^3$. Denote $B = n/m$. When choosing $\eta = \Theta(\sqrt{Km})$, $\eta_l = \Theta(\frac{1}{K\sqrt{TB}})$, the rate can be further bounded by

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] = \mathcal{O}\Big(\frac{\sqrt{B}(f(\theta_1)-f(\theta^*))}{\sqrt{TKm}} + (\frac{1}{\sqrt{TKmB}} + \frac{\sqrt{B}}{\sqrt{TKm}})\sigma^2 + \frac{\sqrt{K}}{\sqrt{TmB}}\sigma_g^2 \\
+ (\frac{1}{TKB} + \frac{1}{T^{3/2}B^{3/2}\sqrt{Km}} + \frac{\sqrt{B}}{\sqrt{TKm}})(\sigma^2+6K\sigma_g^2)\Big),
\end{aligned}
$$

which can be further simplified by ignoring smaller terms as

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] = \mathcal{O}\Big(\frac{\sqrt{n}}{\sqrt{m}}\big(\frac{f(\theta_1)-f(\theta^*)}{\sqrt{TKm}} + \frac{1}{\sqrt{TKm}}\sigma^2 + \frac{\sqrt{K}}{\sqrt{Tm}}\sigma_g^2\big)\Big).
$$

This completes the proof. □

F.5. Proof of Theorem 4.4: Directly Using Biased Compressors Without EF

Proof. We first clarify some (slightly modified) notations. Since we use the biased compression directly, there are no error compensation terms as in previous analysis. The update rule is simply

$$
\theta_{t+1} = \theta_t - \eta\overline{\widetilde{\Delta}}_t,
$$

where $\overline{\widetilde{\Delta}}_t = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\Delta}_{t,i} := \frac{1}{n}\sum_{i=1}^{n}\mathcal{C}(\Delta_{t,i})$ is the average of compressed local model updates. Denote $b_{t,i} = \widetilde{\Delta}_{t,i} - \Delta_{t,i}$ as the difference (bias) between the true local model update of client i in round t, and $\bar{b} = \frac{1}{n}\sum_{i=1}^{n}b_{t,i}$. By Assumption 4.3,

we have

$$\mathbb{E}[\|\bar{b}_t\|^2] = \mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n \widetilde{\Delta}_{t,i} - \frac{1}{n}\sum_{i=1}^n \Delta_{t,i}\|^2\big] \le q^2\mathbb{E}[\|\bar{\Delta}_t\|^2]. \tag{18}$$

Our analysis starts with the smoothness Assumption 4.1, where

$$f(\theta_{t+1}) \le f(\theta_t) + \langle \nabla f(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{L}{2}\|\theta_{t+1} - \theta_t\|^2.$$

Taking expectation w.r.t. the randomness at round t gives

$$\mathbb{E}[f(\theta_{t+1})] - f(\theta_t) \le -\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \overline{\widetilde{\Delta}}_t\rangle\big] + \frac{\eta^2 L}{2}\mathbb{E}\big[\|\overline{\widetilde{\Delta}}_t\|^2\big]. \tag{19}$$

The second term in (19) admits the following:

$$\begin{aligned}
\frac{\eta^2 L}{2}\mathbb{E}\big[\|\overline{\widetilde{\Delta}}_t\|^2\big] &= \frac{\eta^2 L}{2}\mathbb{E}\big[\|\bar{\Delta}_t + \bar{b}_t\|^2\big] \\
&\le (1+q^2)\eta^2 L\mathbb{E}[\|\bar{\Delta}_t\|^2] \\
&\le (1+q^2)\eta^2 L\Big[(2\eta_l^2K^2 + 120\eta_l^4K^4L^2)\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \\
&\qquad + 4\frac{\eta_l^2 K}{n}\sigma^2 + 20\eta_l^4K^3L^2(\sigma^2 + 6K\sigma_g^2)\Big],
\end{aligned}$$

where the last inequality uses Lemma F.2. We can bound the first term in (19) by

$$\begin{aligned}
&-\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \overline{\widetilde{\Delta}}_t\rangle\big] \\
&= -\eta\mathbb{E}\big[\langle \nabla f(\theta_t), \bar{\Delta}_t + \bar{b}_t - \eta_l K\nabla f(\theta_t) + \eta_l K\nabla f(\theta_t)\rangle\big] \\
&= -\eta\eta_l K\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \eta\underbrace{\mathbb{E}\big[\langle \nabla f(\theta_t), -\bar{\Delta}_t + \eta_l K\nabla f(\theta_t)\rangle\big]}_{\textbf{VI}} + \eta\underbrace{\mathbb{E}\big[\langle \nabla f(\theta_t), -\bar{b}_t\rangle\big]}_{\textbf{VII}}.
\end{aligned}$$

For the second term in the above, by Lemma F.3, with $\eta_l \le \frac{1}{8KL}$, we have

$$\textbf{VI} \le \frac{3\eta_l K}{4}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{5\eta_l^3K^2L^2}{2}(\sigma^2 + 6K\sigma_g^2).$$

Regarding term **VII**, by Assumption 4.3, Young's inequality and (18), we obtain

$$\begin{aligned}
\textbf{VII} &\le \frac{\eta_l K}{16}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + \frac{16q^2}{\eta_l K}\mathbb{E}[\|\bar{\Delta}_t\|^2] \\
&\le \frac{\eta_l K}{16}\mathbb{E}[\|\nabla f(\theta_t)\|^2] + 16q^2\Big[(2\eta_l K + 120\eta_l^3K^3L^2)\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] \\
&\qquad + 4\frac{\eta_l}{n}\sigma^2 + 20\eta_l^3K^2L^2(\sigma^2 + 6K\sigma_g^2)\Big].
\end{aligned}$$

When the learning rates admit $\eta_l \le \frac{1}{8KL\max\{1, 8(1+q^2)\eta\}}$ and $q \le \frac{1}{32}$, taking the summation over all terms in (19) we get

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})] - f(\theta_t) \\
&\le -\frac{\eta\eta_l K}{8}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] + \frac{4\eta^2\eta_l^2(1+q^2)KL}{n}\sigma^2 + 20\eta^2\eta_l^4(1+q^2)K^3L^3(\sigma^2 + 6K\sigma_g^2) \\
&\qquad + \frac{64\eta\eta_l q^2}{n}\sigma^2 + (320q^2+3)\eta\eta_l^3K^2L^2(\sigma^2 + 6K\sigma_g^2).
\end{aligned}$$

We now take the telescope summation from round 1 to T and re-organize terms to obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] &\lesssim \frac{f(\theta_1) - \mathbb{E}[f(\theta_{t+1})]}{\eta\eta_l TK} + \frac{4\eta\eta_l(1+q^2)L}{n}\sigma^2 + \frac{64q^2}{Kn}\sigma^2 \\
&\quad + 20\eta\eta_l^3(1+q^2)K^2L^3(\sigma^2 + 6K\sigma_g^2) + (320q^2+3)\eta_l^2KL^2(\sigma^2 + 6K\sigma_g^2).
\end{aligned}$$

Set $\eta_l = \Theta(\frac{1}{K\sqrt{T}})$ and $\eta = \Theta(\sqrt{Kn})$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(\theta_t)\|^2\big] = \mathcal{O}\Big(\frac{1+q^2}{\sqrt{TKn}} + \frac{1+q^2}{TK}(\sigma^2 + K\sigma_g^2) + \frac{q^2\sigma^2}{Kn}\Big).$$

This completes the proof. □

Note that if we consider unbiased compressors, i.e., $\mathbb{E}[b_{t,i}|\Delta_{t,i}] = 0, \forall t, i$, then term **VII** equals zero which eliminates the bias term in the final convergence rate. The resultant convergence rate would match the analysis of Haddadpour et al. (2021).

Algorithm 3 Fed-EF Scheme with Two-Way Compression

Input: learning rates η, η_l, hyper-parameters β_1, β_2, ϵ
Initialize: central server parameter $\theta_1 \in \mathbb{R}^d \subseteq \mathbb{R}^d$; $e_{1,i} = \mathbf{0}$ the accumulator for each worker; $m_0 = \mathbf{0}$, $v_0 = \mathbf{0}$, $\hat{v}_0 = \mathbf{0}$; $\phi_1 = \tilde{H}_t = \mathbf{0}$; $\theta_{0,i}^{(1)} = \theta_1$ for all $i \in [n]$
for $t = 1, \ldots, T$ **do**
  parallel for worker $i \in [n]$ **do**:
    Receive $\tilde{H}_t$ from the server and set $\theta_{t,i}^{(1)} = \theta_{t-1,i}^{(1)} + \tilde{H}_t$ { Download compression }
    for $k = 1, \ldots, K$ **do**
      Compute stochastic gradient $g_{t,i}^{(k)}$ at $\theta_{t,i}^{(k)}$
      Local update $\theta_{t,i}^{(k+1)} = \theta_{t,i}^{(k)} - \eta_l g_{t,i}^{(k)}$
    end for
    Compute the local model update $\Delta_{t,i} = \theta_{t,i}^{(K+1)} - \theta_t$
    Send compressed adjusted local update $\widetilde{\Delta}_{t,i} = \mathcal{C}(\Delta_{t,i} + e_{t,i})$ to central server
    Update the error $e_{t+1,i} = e_{t,i} + \Delta_{t,i} - \widetilde{\Delta}_{t,i}$
  end parallel
  Central server do:
  Global aggregation $\overline{\widetilde{\Delta}}_t = \frac{1}{n}\sum_{i=1}^n \widetilde{\Delta}_{t,i}$
  Compress $\tilde{H}_t = \mathcal{C}(\overline{\widetilde{\Delta}}_t + \phi_t)$ { Fed-EF-SGD }
  $m_t = \beta_1 m_{t-1} + (1-\beta_1)\overline{\widetilde{\Delta}}_t$ { Fed-EF-AMS }
  $v_t = \beta_2 v_{t-1} + (1-\beta_2)\overline{\widetilde{\Delta}}_t^2, \quad \hat{v}_t = \max(v_t, \hat{v}_{t-1})$
  Compress $\tilde{H}_t = \mathcal{C}(\frac{m_t}{\sqrt{\hat{v}_t + \epsilon}} + \phi_t)$
  Update server error accumulator $\phi_{t+1} = \phi_t + (\theta_{t+1} - \theta_t) - \tilde{H}_t$
  Update global model $\theta_{t+1} = \theta_t - \eta \tilde{H}_t$ and broadcast $\tilde{H}_t$ to clients
end for

G. Two-Way Compression in Fed-EF

As discussed in Section 3, our Fed-EF scheme can also extend to two-way compression, for both uploading (clients-to-server) and downloading (server-to-clients) channels. This can lead to even more communication reduction in practice. The steps can be found in Algorithm 3. The general approach is: 1) the clients transmit $\widetilde{\Delta}_{t,i}$ to the server which are compressed; 2) the server again compresses the aggregated update $\overline{\widetilde{\Delta}}_t$ and broadcast the compressed $\tilde{H}_t$ to the clients, also using an error feedback at the central node. Note that this approach requires the clients to additionally store the model at the beginning of each round.

Next, we briefly demonstrate that Algorithm 3 has the same order of convergence rate as Algorithm 1. For simplicity, we will focus on two-way compressed Fed-EF-SGD here, while same arguments hold for Fed-EF-AMS. Assume the same conditions as in Theorem 4.6. To study the convergence of Algorithm 3, we consider a series of virtual iterates as

$$\begin{aligned}\tilde{\theta}_{t+1} = \theta_{t+1} - \eta\phi_{t+1} &= \theta_t - \eta(\tilde{H}_t + \phi_{t+1}) \\ &= \theta_t - \frac{1}{n}\sum_{i=1}^n \widetilde{\Delta}_{t,i} - \phi_t \\ &= \tilde{\theta}_t - \overline{\widetilde{\Delta}}_t,\end{aligned}$$

where we use the fact of EF that $\phi_{t+1} + \tilde{H}_t = \phi_t + \overline{\widetilde{\Delta}}_t$. Then we can construct a similar sequence x_t as in (12) associated

with $\tilde{\theta}_t$ by

$$x_{t+1} = \tilde{\theta}_{t+1} - \eta \bar{e}_{t+1} = x_t - \eta \bar{\Delta}_t.$$

We can then apply same analysis to derive the convergence bound as in Section F.2. The only difference is in (14), where the second term becomes

$$\frac{\eta^2 L^2}{2} \mathbb{E}\big[\|\bar{e}_t + \phi_t\|^2\big] \leq \eta^2 L^2 \mathbb{E}\big[\|\bar{e}_t\|^2\big] + \eta^2 L^2 \mathbb{E}\big[\|\phi_t\|^2\big]. \tag{20}$$

The first term can be bounded in the same way as in (14). Regarding the second term, we can use a similar trick as Lemma F.5 that under Assumption 4.3,

$$\begin{aligned}
\|\phi_{t+1}\|^2 &= \|\phi_t + \overline{\widetilde{\Delta}}_t - \mathcal{C}(\phi_t + \overline{\widetilde{\Delta}}_t)\|^2 \\
&\leq q_{\mathcal{C}}^2 \|\phi_t + \overline{\widetilde{\Delta}}_t\|^2 \\
&\leq \frac{1 + q_{\mathcal{C}}^2}{2} \|\phi_t\|^2 + \frac{2 q_{\mathcal{C}}^2}{1 - q_{\mathcal{C}}^2} \|\overline{\widetilde{\Delta}}_t\|^2.
\end{aligned}$$

Then, by recursion and the geometric sum, $\|\phi_{t+1}\|^2$ can be bounded by the second term in above up to a constant. We can write

$$\begin{aligned}
\mathbb{E}\big[\|\overline{\widetilde{\Delta}}_t\|^2\big] &\leq \mathbb{E}\big[\|\bar{\Delta}_t + \bar{e}_t - \bar{e}_{t+1}\|^2\big] \\
&\leq 3(\mathbb{E}\big[\|\bar{\Delta}_t\|^2\big] + \mathbb{E}\big[\|\bar{e}_t\|^2\big] + \mathbb{E}\big[\|\bar{e}_{t+1}\|^2\big]).
\end{aligned}$$

As a result, it holds that $\mathbb{E}\big[\|\phi_t\|^2\big] = \Theta(\mathbb{E}\big[\|\bar{e}_t\|\big]^2)$ since $\mathbb{E}\big[\|\bar{e}_t\|^2\big] = \Theta(\mathbb{E}\big[\|\bar{\Delta}_t\|\big]^2)$ by Lemma F.2 and Lemma F.5. Therefore, (20) has same order as (14). Since other parts of the proof are the same, we conclude that two-way compression does not change the asymptotic convergence rate of Fed-EF.