Communication-Efficient Adaptive Federated Learning

Yujia Wang [1] **Lu Lin** [2] **Jinghui Chen** [1]

Abstract

Federated learning is a machine learning training paradigm that enables clients to jointly train models without sharing their own localized data. However, the implementation of federated learning in practice still faces numerous challenges, such as the large communication overhead due to the repetitive server-client synchronization and the lack of adaptivity by SGD-based model updates. Despite that various methods have been proposed for reducing the communication cost by gradient compression or quantization, and the federated versions of adaptive optimizers such as FedAdam are proposed to add more adaptivity, the current federated learning framework still cannot solve the aforementioned challenges all at once. In this paper, we propose a novel communication-efficient adaptive federated learning method (FedCAMS) with theoretical convergence guarantees. We show that in the nonconvex stochastic optimization setting, our proposed FedCAMS achieves the same convergence rate of $\mathcal{O}(\frac{1}{\sqrt{TKm}})$ as its non-compressed counterparts. Extensive experiments on various benchmarks verify our theoretical analysis.

1. Introduction

Federated learning (FL) (Konečnỳ et al., 2016; McMahan et al., 2017) has recently become a popular machine learning training paradigm where multiple clients cooperate to jointly learn a machine learning model. In the federated learning setting, training data is distributed across a large number of clients, or edge devices, such as smartphones, personal computers, or IoT devices. These clients own valuable data for training a variety of machine learning models, yet those raw client data is not allowed to share with the server or other clients due to privacy and regulation concerns. Federated Learning (Konečnỳ et al., 2016; McMahan et al., 2017) works by having each client train the ML model locally based on its own data, while having the clients iteratively exchanging and synchronizing their local ML model parameters with each other through a central server. McMahan et al. (2017) proposed FedAvg algorithm, whose global model is updated by averaging multiple steps of local stochastic gradient descent (SGD) updates, and it has become one of the most popular FL methods.

Despite the ability to jointly train the model without directly sharing the data, the implementation of FL in practice still faces several major challenges such as (1) *large communication overhead* due to the repetitive synchronization between the server and the clients; and (2) *lack of adaptivity* as SGD-based update may not be suitable for heavy-tail stochastic gradient noise distributions, which often arise in training large-scale models such as BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), GAN (Goodfellow et al., 2014) or ViT (Dosovitskiy et al., 2021).

Note that various attempts have been made to solve the aforementioned challenges individually but not all of them at once. In terms of reducing communication costs, one can avoid transmitting the complete model updates when synchronizing. Several works, including (Reisizadeh et al., 2020; Jin et al., 2020; Jhunjhunwala et al., 2021; Chen et al., 2021b), have studied the compressed and quantized federated learning optimization method based on FedAvg. Another way is to reduce the number of participating clients such that only part of the clients participate in the model training at each round (Yang et al., 2021; Li et al., 2019b; Nishio & Yonetani, 2019; Li & Wang, 2019). Besides, the network resources allocation also plays an important role in communication-efficient federated learning problems (Li et al., 2019a; Yang et al., 2020). For adaptivity concerns, recently, FedAdam (Reddi et al., 2020) and other variants, such as FedYogi (Reddi et al., 2020) and FedAMSGrad (Tong et al., 2020) were proposed to introduce adaptive gradient methods (Kingma & Ba, 2014; Reddi et al., 2018) into federated learning framework and provided provable convergence guarantees. However, it is still an open problem how to achieve communication efficient adaptive federated optimization while still providing rigorous convergence

[1]College of Information Sciences and Technology, Pennsylvania State University, State College, PA, United States [2]Department of Computer Science, University of Virginia, Charlottesville, VA, United States. Correspondence to: Jinghui Chen <jzc5917@psu.edu>.

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

guarantees.

In this paper, we aim to develop a new compressed federated adaptive gradient optimization method that is communication-efficient while also provably convergences. Specifically, we first propose FedAMS, a variant of FedAdam, with an improved convergence analysis over the original FedAdam. Based on FedAMS, we propose FedCAMS, a **Fed**erated **C**ommunication-compressed **A**MSGrad with **M**ax **S**tablization (FedCAMS), which addresses both the communication and adaptivity challenges within one training framework. We summarize our contributions as follows:

- We provide an improved analysis on the convergence behaviour of FedAMS, a variant of the existing federated adaptive gradient method FedAdam (Reddi et al., 2020), whose analysis is simplified for only considering the case where no momentum is been used. In particular, we prove that FedAMS (with momentum) can achieve the same convergence rate of $\mathcal{O}(\frac{1}{\sqrt{TKm}})$ w.r.t total iterations T, the number of local updates K, and the number of workers m for both full participation and partial participation schemes.
- We propose a new communication-efficient adaptive federated optimization method, FedCAMS, which to the best of our knowledge, for the first time, achieves both communication efficiency and adaptivity in federated learning with one single learning framework. FedCAMS largely reduces the communication cost by error feedback and compression strategy and it is compatible with various commonly-used compressors in practice. We prove that FedCAMS achieves the same convergence rate of $\mathcal{O}(\frac{1}{\sqrt{TKm}})$, as its uncompressed counterpart FedAMS.
- We conduct experiments on various benchmarks and show that our proposed FedAMS and FedCAMS achieve good adaptivity in training real-world machine learning models. Furthermore, we show that FedCAMS effectively reduced the communication cost (number of bits for communication) by orders of magnitude while sacrificing little in terms of prediction accuracy.

Notation: For vectors $\mathsf{x}; \mathsf{y} \in \mathsf{R}^d$, $\sqrt{\mathsf{x}}; \mathsf{x}^2; \mathsf{x}=\mathsf{y}$ denote the element-wise square root, square, and division of the vectors. For vector x and matrix A, $\|\cdot\|$ denotes the ℓ_2 norm of vector/matrix, i.e., $\|\mathsf{x}\| = \|\mathsf{x}\|_2$ and $\|A\| = \|A\|_2$.

2. Related Work

SGD and Adaptive Gradient Methods: Stochastic gradient descent (SGD) (Robbins & Monro, 1951) has been widely applied in training machine learning models for decades. Although SGD is straightforward to implement, it is known to be sensitive to parameters and relatively slow to converge when facing heavy-tail stochastic gradient noise. Adaptive gradient methods were proposed to overcome these issues of SGD, including AdaGrad (Duchi et al., 2011), RMSProp (Tieleman et al., 2012), AdaDelta (Zeiler, 2012). Adam (Kingma & Ba, 2014) and its variant AMSGrad (Reddi et al., 2018), are tremendously used in training deep neural networks, and other variants (Luo et al., 2019; Loshchilov & Hutter, 2017; Chen et al., 2020a) also play important roles in improving adaptive gradient methods through different aspects.

Federated Learning: As the demand of locally data storing and training models at edge devices, Federated Learning (Konečnỳ et al., 2016; Li et al., 2020) rapidly attracts growing interest in recent years. Federated Averaging method (FedAvg) (McMahan et al., 2017) works by periodically averaging local SGD updates. Stich (2018) provided a concise theoretical convergence guarantee for local SGD. Lin et al. (2018) proposed a variant of local SGD with empirical improvements. There are many works based on FedAvg such as FedProx (Li et al., 2020), FedNova (Wang et al., 2020), SCAFFOLD (Karimireddy et al., 2020), and other work discussed the variants of FedAvg (Yang et al., 2021; Li et al., 2019b; Hsu et al., 2019; Wang et al., 2019). Reddi et al. (2020) recently proposed several adaptive federated optimization methods including FedAdagrad, FedYogi and FedAdam to overcome the existing convergence issues of FedAvg. Chen et al. (2020b) proposed Local AMSGrad and Tong et al. (2020) proposed a family of federated adaptive gradient methods with calibrations. Another line of research focused on addressing data heterogeneity issues or the network resource allocation issues (Ghosh et al., 2019; Li & Wang, 2019; Yang et al., 2020).

Communication-Compressed Federated Learning: Various strategies have been proposed for reducing communication costs in distributed learning for SGD based algorithms (Bernstein et al., 2018; Seide et al., 2014; Alistarh et al., 2017; Basu et al., 2019; Stich et al., 2018; Stich & Karimireddy, 2019; Karimireddy et al., 2019) and also adaptive gradient methods (Tang et al., 2021; Wang et al., 2022). In terms of federated learning, many studies have tried to apply the aforementioned methods to FedAvg and have attracted growing interest recently, e.g., FedPAQ (Reisizadeh et al., 2020), FedCOM (Haddadpour et al., 2021), sign SGD in federated learning (Jin et al., 2020), communication-efficient federated learning (Chen et al., 2021b), AdaQuantFL (Jhunjhunwala et al., 2021). However, there are fewer attempts to develop communication-efficient adaptive gradient methods in federated learning, which is our key focus in this work.

3. Proposed Method

In this paper, we aim to study the following federated learning nonconvex optimization problem:

$$\min_{x\in\mathbb{R}^d} f(\mathsf{x}) = \frac{1}{m}\sum_{i=1}^{m} F_i(\mathsf{x}); \tag{3.1}$$

where m is the total amount of local clients, d denotes the dimension of the model parameters, $F_i(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}_i} F_i(\mathbf{x};\ _i)$ is the local nonconvex loss function on client i associated with a local distribution $\mathcal{D}_i$. In the stochastic setting, we can only obtain the unbiased estimator of $F_i(\mathbf{x})$, i.e., the stochastic gradient $\mathbf{g}_t^i = \nabla F_i(\mathbf{x};\ _i)$. In the non i.i.d setting, distributions $\mathcal{D}_i; \mathcal{D}_j$ can vary from each other, i.e., $\mathcal{D}_i \neq \mathcal{D}_j$, $\forall i \neq j$.

FedAvg (McMahan et al., 2017) is a commonly used optimization approach to solve (3.1). Let $\mathbf{x}_t$ denotes the global model parameters before the t-th iteration. Now at iteration t, the participating client i from the selected subset $\mathcal{S}_t$ (with size n) receives the model $\mathbf{x}_t$ from the server, conducts K steps of local SGD updates with local learning rate $_l$, obtains the local model $\mathbf{x}_{t,K}^i$. Client i then sends the model difference $\Delta_t^i = \mathbf{x}_{t,K}^i - \mathbf{x}_t$ to the server. And the server updates the global model difference Δ_t by simply averaging the local model differences Δ_t^i. The server then updates the global model $\mathbf{x}_{t+1}$ by $\mathbf{x}_{t+1} = \mathbf{x}_t + \Delta_t$, which is the same[1] as directly averaging the local model $\mathbf{x}_{t,K}^i$, i.e., $\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{1}{n}\sum_{i\in\mathcal{S}_t}(\mathbf{x}_{t,K}^i - \mathbf{x}_t) = \frac{1}{n}\sum_{i\in\mathcal{S}_t}\mathbf{x}_{t,K}^i$.

FedAdam was then proposed among several adaptive optimization methods in federated learning (Reddi et al., 2020). FedAdam changes the global update rule of FedAvg from one-step SGD to one-step adaptive gradient optimization. Specifically, after gathering local differences Δ_t^i and averaging to Δ_t, the server updates the global model by Adam optimizer:

$$\mathbf{m}_t = \ _1\mathbf{m}_{t-1} + (1 - \ _1)\Delta_t; \tag{3.2}$$

$$\mathbf{v}_t = \ _2\mathbf{v}_{t-1} + (1 - \ _2)\Delta_t^2; \tag{3.3}$$

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t} + }; \tag{3.4}$$

where Δ_t acts as pseudo gradient, and the global update can be viewed as one step Adam update using Δ_t. Several variants were also proposed will slight changes in the variance term $\mathbf{v}_t$, such as FedAdagrad and FedYogi (Reddi et al., 2020) and FedAMSGrad (Tong et al., 2020). Note that the in (3.4) is used for numerical stabilization purpose as the $\mathbf{v}_t$ term can be quite small and cause unstable optimization behaviours.

3.1. Federated AMSGrad with Max Stabilization

In this section, we propose a general adaptive federated optimization framework, **Fed**erated **A**MSGrad with **M**ax **S**tabilization (FedAMS), where the server conducts one additional max stabilization step before the final update.

Algorithm 1 summarize the details of general FedAMS framework. At the beginning of global round t, we first select a subset of clients $\mathcal{S}_t$, each participating client $i \in \mathcal{S}_t$ obtains the local model $\mathbf{x}_{t,K}^i$ after K steps of local SGD updates with learning rate $_l$. The model difference Δ_t^i is the difference between the local updated model $\mathbf{x}_{t,K}^i$ and the current global model $\mathbf{x}_t$, i.e., $\Delta_t^i = \mathbf{x}_{t,K}^i - \mathbf{x}_t$. The server aggregates Δ_t^i and gets the global difference Δ_t, this Δ_t acts as a pseudo gradient to calculate momentum $\mathbf{m}_t$ and variance $\mathbf{v}_t$ following (3.2) and (3.3). Now for updating $\mathbf{x}_{t+1}$, our general FedAMS framework provides two options for max stabilization:

$$\text{Option 1: } \hat{\mathbf{v}}_t = \max(\hat{\mathbf{v}}_{t-1}; \mathbf{v}_t;\ ); \mathbf{x}_{t+1} = \mathbf{x}_t + \frac{\mathbf{m}_t}{\sqrt{\hat{\mathbf{v}}_t}};$$

$$\text{Option 2: } \hat{\mathbf{v}}_t = \max(\hat{\mathbf{v}}_{t-1}; \mathbf{v}_t); \mathbf{x}_{t+1} = \mathbf{x}_t + \frac{\mathbf{m}_t}{\sqrt{\hat{\mathbf{v}}_t + }};$$

Note that Option 2 is the same as the AMSGrad (Reddi et al., 2018) update rule, which brings a non-decreasing $\mathbf{v}_t$ to solve a non-convergence issue in Adam (Kingma & Ba, 2014). For Option 1, FedAMS directly adopts $\sqrt{\hat{\mathbf{v}}_t}$ as the denominator where is token as the part of the max operation in $\hat{\mathbf{v}}_t$. Intuitively, the unstable behaviour of the denominator (small value in the $\mathbf{v}_t$) usually only happens for a small set of dimensions. Therefore, the max stabilization strategy in Option 1 only affects those dimensions with small $\mathbf{v}_t$ values, while the traditional adding strategy as in Option 2 will affect the accuracy on all dimensions.

Moreover, we want to emphasize that although the theoretical analysis in Reddi et al. (2020) assumes $_1 = 0$ and only considers the impact of variance $\mathbf{v}_t$, thus the non-decreasing variance is not necessary for the analysis in Reddi et al. (2020). While the non-decreasing variance is indeed necessary for us to obtain the complete proof with a positive $_1$ (see Appendix for details).

3.2. Federated Communication-Compressed AMSGrad

In order to reduce the communication costs between synchronization, we propose **Fed**erated **C**ommunication-compressed **A**MSGrad with **M**ax **S**tabilization (FedCAMS), which is summarized in Algorithm 2. The main difference lies in that after the client i obtains the model differences Δ_t^i via local SGD, FedCAMS will compress Δ_t^i to $\hat{\Delta}_t^i$ via error feedback compression strategy, and then send $\hat{\Delta}_t^i$ to the central server. In details, at round t, the client i will apply the compressor on the summation of model differences Δ_t^i together with the cumulative compression error $\mathbf{e}_t^i$ to obtain $\hat{\Delta}_t^i$. After that, the client will update term $\mathbf{e}_{t+1}^i$ by calculating the new cumulative compression error, i.e., $\mathbf{e}_{t+1}^i = \Delta_t + \mathbf{e}_t^i - \hat{\Delta}_t^i$, which will be useful for next round's computation. The rest part of FedCAMS is similar to FedAMS: the server aggregates $\hat{\Delta}_t^i$ and obtains $\hat{\Delta}_t$, which will participate in the global update.

To summarize, FedCAMS is indeed a communication-

[1] The global update of FedAvg is equivalent to perform one step SGD update with the pseudo gradient $_t$ and learning rate $\eta = 1$.

Algorithm 1 FedAMS

Input: initial point x_1, local step size η, global stepsize η, $\beta_1, \beta_2, \epsilon$.

1: $m_0 = 0, v_0 = 0$
2: **for** $t = 1$ to T **do**
3: Random sample a subset S_t of clients
4: Server sends x_t to the subset S_t of clients
5: $x^i_{t,0} = x_t$
6: **for** each client $i \in S_t$ in parallel **do**
7: **for** $k = 0, \ldots, K-1$ **do**
8: Compute local stochastic gradient $g^i_{t,k} = \nabla F_i(x^i_{t,k}; \xi^i_{t,k})$
9: $x^i_{t,k+1} = x^i_{t,k} - \eta g^i_{t,k}$
10: **end for**
11: $\Delta^i_t = x^i_{t,K} - x_t$
12: **end for**
13: Server aggregates local updates: $\Delta_t = \frac{1}{|S_t|} \sum_{i \in S_t} \Delta^i_t$
14: Update $m_t = \beta_1 m_{t-1} + (1-\beta_1)\Delta_t$
15: Update $v_t = \beta_2 v_{t-1} + (1-\beta_2)\Delta_t^2$
//Option 1:
16: $\hat{v}_t = \max(\hat{v}_{t-1}, v_t, \epsilon)$, update $x_{t+1} = x_t + \eta \frac{m_t}{\sqrt{\hat{v}_t}}$
//Option 2:
17: $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$, update $x_{t+1} = x_t + \eta \frac{m_t}{\sqrt{\hat{v}_t + \epsilon}}$
18: **end for**

ef cient with the following features.

Error Feedback Compression: Although error-feedback strategy (Karimireddy et al., 2019; Stich et al., 2018; Stich & Karimireddy, 2019) has been widely used in various distributed learning settings, there is much less use of error feedback in the federated settings, especially for adaptive federated optimization. Note that combining error-feedback with adaptive federated optimization is not a trivial task at all, instead, it is actually quite complicated. Speci cally, the theoretical analysis of the adaptive gradient method in the typical nonconvex setting relies on the construction of the Lyapunov function. Compared to directly analyzing the model parameter x, this causes extra dif culty as applying the error feedback strategy on the smoothness-expanded terms from the Lyapunov function will result in an accumulation of the compression error which leads to divergence.[2] In our theoretical analysis, we have to modify the original construction of the Lyapunov function and introduce a new auxiliary sequence about the compression error which eliminates the accumulation of compression error. Unlike those direct compression strategies such as simple quantization or direct compression strategies (Haddadpour et al., 2021; Reisizadeh et al., 2020; Jin et al., 2020) which usually require an unbiased compressor to work, error feedback allows for various biased compressors such as commonly used scaled sign compressor or top-k compressors. Furthermore, the design of the error feedback strategy is well-known for reducing unnecessary compression error, which leads to a more precise model update.

Support for Partial Participation: Here we also make error feedback compatible with partial participation settings by keeping the stale cumulative compression error for clients who were not selected for the current round training (see Lines 14-16 in Algorithm 2). Such design makes FedCAMS more practical and communication ef cient. Note that the default client sampling strategy in FedCAMS is to randomly select the participating clients (without replacement) in each round, i.e., $p_i = \Pr\{i \in S_t\} = n/m$. This can be easily extended to the weighted sampling strategy with probability $p_i = w_i$, even with varying numbers of participating workers n.

Algorithm 2 FedCAMS

Input: initial point x_1, local step size η, global stepsize η, $\beta_1, \beta_2, \epsilon$, compressor $C(\cdot)$.

1: $m_0 = 0, v_0 = 0, e^i_1 = 0$
2: **for** $t = 1$ to T **do**
3: Random sample a subset S_t of clients
4: Server sends x_t to the subset S_t of clients
5: $x^i_{t,0} = x_t$
6: **for** each client $i \in S_t$ in parallel **do**
7: **for** $k = 0, \ldots, K-1$ **do**
8: Compute local stochastic gradient $g^i_{t,k} = \nabla F_i(x^i_{t,k}; \xi^i_{t,k})$
9: $x^i_{t,k+1} = x^i_{t,k} - \eta g^i_{t,k}$
10: **end for**
11: $\Delta^i_t = x^i_{t,K} - x_t$
12: Compress $b^i_t = C(\Delta^i_t + e^i_t)$, send b^i_t to the server and update $e^i_{t+1} = \Delta^i_t + e^i_t - b^i_t$
13: **end for**
14: **for** each client $j \notin S_t$ in parallel **do**
15: client j maintains the stale compression error $e^j_{t+1} = e^j_t$
16: **end for**
17: Server aggregates local updates: $b_t = \frac{1}{|S_t|} \sum_{i \in S_t} b^i_t$
18: Server updates x_{t+1} using b_t in the same way as in Algorithm 1 (Line 14-17)
19: **end for**

4. Convergence Analysis

In this section, we present the theoretical convergence results of our proposed FedAMS and FedCAMS in Algorithm 1 and 2. We rst introduce some assumptions needed for the proof.

[2] Similar divergence issue has also been discussed in Tang et al. (2021); Wang et al. (2022) in the distributed setting.

Assumption 4.1(Smoothness)Each loss function on the i-th worker$F_i(x)$ is L-smooth, i.e.,$\forall x, y \in \mathbb{R}^d$,

$$F_i(x) \quad F_i(y) \quad hr \quad F_i(y); x \quad yi \quad \frac{L}{2}kx \quad yk^2:$$

This also implies that$L$-gradient Lipschitz condition, i.e.,$kr \ F_i(x) \ r \ F_i(y)k \quad Lkx \quad yk$. Assumption 4.1 is a standard assumption in nonconvex optimization problems, which has been also adopted in Kingma & Ba (2014); Reddi et al. (2018); Li et al. (2019b); Yang et al. (2021).

Assumption 4.2(Bounded Gradient)Each loss function on the$i$-th worker$F_i(x)$ has$G$-bounded stochastic gradient onℓ_2, i.e., for all , we have$kr \ f_i(x; \ )k \quad G$.

The assumption of bounded gradient is usually adopted in adaptive gradient methods (Kingma & Ba, 2014; Reddi et al., 2018; Zhou et al., 2018; Chen et al., 2020a)

Assumption 4.3(Bounded Variance)Each stochastic gradient on the$i$-th worker has a bounded local variance, i.e., for all $x; i \in [m]$,we have$\mathbb{E} \ kr \ f_i(x; \ ) \ r \ F_i(x)k^2 \quad \sigma_l^2$, and the loss function on each worker has a global variance bound,$\frac{1}{m} \sum_{i=1}^{m} kr \ F_i(x) \ r \ f(x)k^2 \quad \sigma_g^2$.

Assumption 4.3 is widely used in federated optimization problems (Li et al., 2019b; Reddi et al., 2020; Yang et al., 2021). The bounded local variance represents the randomness of stochastic gradients, and the bounded global variance represents data heterogeneity between clients. Note that $\sigma_g = 0$ corresponds to the i.i.d setting, in which datasets from each client have the same distribution.

In the following, we will show the convergence results of FedAMS[3] and FedCAMS.

4.1. Convergence Analysis for FedAMS

Full Participation: For the full participation scheme, all workers participate in the communication rounds and model update, i.e.,$|S_t| = m; \forall t \in [t]$.

Theorem 4.4. Under Assumptions 4.1-4.3, if the local learning rate η satisfies the following condition: $\eta \quad \min\left\{\frac{1}{8KL}; \frac{\sqrt{\quad}}{K\sqrt{\quad^2K^2G^2 + [(3+C_1^2)L^2 + 2\sqrt{2(1 \quad \quad^2)}G]}}\right\}$, then the iterates of FedAMS in Algorithm 1 under full participation scheme satisfy

$$\min_{t \in [T]} \mathbb{E} \ kr \ f(x_t)k^2 \quad 4\sqrt{\quad^2\eta^2K^2G^2 + \quad} \left(\frac{f_0 \quad f^*}{\eta KT} + \frac{}{T} + \quad\right); \qquad (4.1)$$

where $= \frac{C_1\sqrt{G^2d}}{\sqrt{\quad}} + \frac{2C_1^2 \quad \eta KLG^2d}{}$, $= \frac{5\sqrt{\quad}\eta^2KL^2}{2}(\quad^2 + \quad 6K^2_g) + [(3+C_1^2)L^2 + 2\sqrt{2(1 \quad \quad^2)}G]\frac{\eta^2}{2m}\quad^2$ and $C_1 = \frac{1}{1 \quad \quad_1}$.

Remark 4.5. The upper bound for$\min_{t \in [T]} \mathbb{E}[kr \ f(x_t)k^2]$ contains three parts: the first two items that directly related to the total number of steps$T$ are vanishing as$T \to 1$. The last term in(4.1) relates to the local stochastic varianceσ_l and global varianceσ_g. In the i.i.d setting where each worker has the same data distribution, we have zero global variance, i.e.,$\sigma_g = 0$, and the variance term will be smaller and less dependent on the number of local steps$K$.

Corollary 4.6. Suppose we choose the global learning rate $\eta = (\sqrt{Km})$ and local learning rate $\eta = (\frac{1}{\sqrt{TK}})$, when$T$ is sufficient large, i.e.,$T \quad Km$, the convergence rate for FedAMS in Algorithm 1 under full participation scheme satisfies

$$\min_{t \in [T]} \mathbb{E} \ kr \ f(x_t)k^2 \quad = O\left(\frac{1}{\sqrt{TKm}}\right): \qquad (4.2)$$

Remark 4.7. Corollary 4.6 suggests that with sufficient large$T$, FedAMS achieves a convergence rate of$O(\frac{1}{\sqrt{TKm}})$, which matches the result for general federated non-convex optimization methods such as SCAFFOLD (Karimireddy et al., 2020), FedAdam (Reddi et al., 2020).

Remark 4.8. Note that compared with FedAdam (Reddi et al., 2020), our theoretical analysis on FedAMS makes improvements in completing the proof. The analysis in Reddi et al. (2020) can only consider the case when$\beta_1 = 0$ which largely simplifies the proof in their theoretical analysis, while we provide a full analysis on FedAMS with non-zero momentum term.

Partial Participation: In the partial participation scheme, we assume that only$n$ of$m$ workers participate the local update and communicate with the central server on each step$t$, i.e.,$|S_t| = n; \forall t \in [1; T]$. The partial participation includes the randomness of sampling, and the coefficient varies for different sampling methods. Here we consider the random sampling without replacement. At the$t$-th iteration, we randomly sample a subsetS_t contains$n$ workers for local updating, for any two workers$i; j \in S_t$, the probability of being sampled to participate in model update are$\mathbb{P}\{i \in S_t\} = \frac{n}{m}$ and$\mathbb{P}\{i; j \in S_t\} = \frac{n(n \quad 1)}{m(m \quad 1)}$.

Theorem 4.9. Under Assumption 4.1-4.3, if the local learning rate η satisfies: $\eta \quad \min\left\{\frac{1}{8KL}; \frac{\sqrt{\quad}}{24m(n \quad 1)K\sqrt{\quad^2K^2G^2 + [3L^2 + C_1^2L^2 + 2\sqrt{2(1 \quad \quad^2)}G]}}\right\}$, then the iterates of FedAMS in Algorithm 1 under partial participation scheme satisfy

$$\min_{t \in [T]} \mathbb{E} \ kr \ f(x_t)k^2 \quad 8\sqrt{\quad^2\eta^2K^2G^2 + \quad} \left(\frac{f_0 \quad f^*}{\eta KT} + \frac{}{T} + \quad\right); \qquad (4.3)$$

[3]For simplicity, we will only present the convergence guarantee with Option 1. Note that the theoretical analysis can be easily extended to Option 2 with constant-only changes.

where $= \frac{C_1 G^2 d}{p} + \frac{2C_1^2 \, _l KLG^2 d}{p}$ and $= \frac{5 p^2 KL^2}{2}(\, _l^2 + 6K \, _g^2) + [(3 + C_1^2) L + 2\sqrt{2(1 \quad _2)}G]\frac{_l}{2n} \, _l^2 + [(3 + C_1^2) L + 2\sqrt{2(1 \quad _2)}G]\frac{_l(m \quad n)}{2n(m \quad 1)}[15K^2 L^2 \, ^2(\, _l^2 + 6K \, _g^2) + 3K \, _g^2]$ and $C_1 = \frac{1}{1 \quad _1}$.

Remark 4.10. The upper bound for $\min_{t2[T]} E[kr \, f(x_t)k^2]$ of partial participation is similar to full participation case but with a larger variance term. This is due to the fact that random sampling of participating workers introduces an additional variance during sampling. In the i.i.d setting where we have zero global variance, i.e. $_g = 0$, the variance term will get smaller and less dependent on the number of local steps K as well.

Corollary 4.11. Suppose we choose the global learning rate $= (\sqrt{Kn})$ and local learning rate $_l = (p\frac{1}{TK})$, the convergence rate for FedAMS in Algorithm 1 under partial participation scheme without replacement sampling is

$$\min_{t2[T]} E \, kr \, f(x_t)k^2 \quad = O \quad p\frac{K}{Tn} \quad : \qquad (4.4)$$

Remark 4.12. Note that Corollary 4.11 suggests that the dominant term in (4.3) is $O(p\frac{K}{Tn})$, which directly relates to the global variance ^2_g. Such convergence rate is consistent with the partial participation result of FedAvg in the non i.i.d case in (Yang et al., 2021). It shows that the global variance has more impact on convergence behaviour in partial participation cases, especially in highly non i.i.d. cases where $_g$ is large. Corollary 4.11 also suggests that larger number of participating clients n would accelerate the convergence. Note that although FedAdam (Reddi et al., 2020) did not provide explicit conclusions on the partial participation setting, its appendix introduced the necessary steps for analyzing such setting, which cannot imply such desired relationship with n.

Remark 4.13. The impact of the number of local updates K is complicated. In partial participation settings, it shows that larger K slows down the convergence while full participation suggests the opposite. Similar slow-down result has also been mentioned in Li et al. (2019b), while some others (Stich, 2018; McMahan et al., 2017) showed that larger K would increase the convergence rate. We will leave this as future work.

4.2. Convergence Analysis for FedCAMS

Let us first introduce the assumption for the compressor.

Assumption 4.14 (Biased Compressor). Consider a biased operator $C : R^d \ ! \ R^d$: for $8x \ 2 \ R^d$, there exists constant $0 \quad q \quad 1$ such that

$$E[kC(x) \quad xk] \quad qkxk; 8x \ 2 \ R^d:$$

Note that $q = 0$ leads to $C(x) = x$ which means no compression to x. Assumption 4.14 is a standard assumption for biased compressors (Karimireddy et al., 2019; Stich et al., 2018; Tang et al., 2021). There are several widely used compressors satisfying 4.14 such as scaled-sign compressor and topk compressor.

Top-k (Stich et al., 2018): For $1 \quad k \quad d$ and $8x \ 2 \ R^d$, the coordinate of x is ordered by the magnitude $jx_{(1)}j \quad jx_{(2)}j \quad \cdots \quad jx_{(d)}j$. Denote $_1; \ _2; \ldots; \ _d$ as standard unit basis vectors in R^d. The compressor $C_{top} : R^d \ ! \ R^d$ is defined as: $C_{top}(x) = \sum_{i=d \quad k+1}^{d} x_{(i)} \, _{(i)}$.

Remark 4.15. Let us define the compression ratio as $r = k=d$. It can be shown that $kC_{top}(x) \quad xk^2 \quad p\frac{(1 \quad r)}{k=d}kxk^2 = (1 \quad r)kxk^2$, and thus we have $q = p\frac{1 \quad r}{1}$.

Scaled sign (Karimireddy et al., 2019): For $1 \quad k \quad d$ and $8x \ 2 \ R^d$, the compressor $C_{sign} : R^d \ ! \ R^d$ is defined as

$$C_{sign}(x) = kxk_1 \quad sign(x)=d:$$

Remark 4.16. For scaled sign compressor, we have $kC_{sign}(x) \quad xk^2 = (1 \quad k \, xk_1^2=dkxk^2)kxk^2$, thus we have $q = p\frac{1 \quad k \, xk_1^2=dkxk^2}{1}$.

Next, we show the convergence analysis for FedCAMS. Due to the space limit, we only show the full participation setting and leave the partial participation setting in Appendix B.3.

Theorem 4.17. Under Assumptions 4.1-4.3 and 4.14, if the local learning rate $_l$ satisfies the following condition: $_l \quad \min\{\frac{1}{8KL}; \frac{p}{KC_{;q}[3L \, ^2 + 2C_2 L \, ^2 + 2\sqrt{2(1 \quad _2)}G]}\}$, where $C_{;q} = p\frac{}{4 \, _2(1+q^2)^3(1 \quad q^2)^2K^2G^2 +}$, then the iterates of FedCAMS in Algorithm 2 satisfy

$$\min_{t2[T]} E \, kr \, f(x_t)k^2 \quad s\left(4 \quad 2\frac{4(1+q^2)^3}{(1 \quad q^2)^2} \, _l^2K^2G^2 + \right)\frac{f_0 \quad f}{_lKT} + \frac{}{T} + \quad ; \qquad (4.5)$$

where $= \frac{C_1 G^2 d}{p} + \frac{2C_1^2 \, _l KLG^2 d}{p}$ and $= \frac{5 p^2 KL^2}{2}(\, _l^2 + 6K \, _g^2) + [(3 + 2 \, C_2) L + 2\sqrt{2(1 \quad _2)}G]\frac{_l}{2m} \, _l^2$, $C_1 = \frac{1}{1 \quad _1} + \frac{2q}{1 \quad q^2}$ and $C_2 = \frac{^2_1}{(1 \quad _1)^2} + \frac{4q^2}{(1 \quad q^2)^2}$.

Remark 4.18. The convergence rate in Theorem 4.17 contains three parts as well, the first two parts are related to total iterates T, and they vanish as T increases. The last term shows no direct dependency on T, but on local and global variances. In the i.i.d case where $_g = 0$, the variance will decrease and show less dependency on the number of local steps K.

Corollary 4.19. Suppose we choose the global learning rate $= (\sqrt{Km})$ and local learning rate $_l = (p\frac{1}{TK})$, when T is sufficient large, i.e., $T \quad Km$, the convergence rate for FedCAMS in Algorithm 2 under full participation

scheme satis es

$$\min_{t\in[T]} \mathbb{E}\,\|\nabla f(x_t)\|^2 = O\Big(\frac{1}{\sqrt{TKm}}\Big). \tag{4.6}$$

Remark 4.20. Corollary 4.19 suggests that with suf cient large T, FedCAMS achieves the desired $O(\frac{1}{\sqrt{TKm}})$ convergence rate which matches the result for its uncompressed counterpart, FedAMS. This suggests that FedCAMS can indeed achieve better communication ef ciency without sacri cing much on the accuracy.

Remark 4.21. The constants C_1 and C_2 in Theorem 4.17 are related to the compression constant q. Speci cally, if we track the dependency on q in the convergence rate, we have $O(1/(1-q)\sqrt{TKm})$ under the full participation scheme. A larger q ($q \to 1$) corresponds to a stronger compression we applied, leading to worse convergence due to heavier information losses. Note that this q-dependency is common for the adaptive gradient method since the convergence proof of adaptive gradient methods heavily relies on the bounded gradient assumption and thus the compressed gradient bound is related to q. Similar type of q-dependency also occurs in other communication compressed distributed Adam methods such as Chen et al. (2021a).[4]

5. Experiments

In this section, we present empirical validations toward the effectiveness of our proposed algorithms. Firstly, we provide comparisons between FedAMS and other first-order federated optimization baselines. Secondly, we provide experimental results of our proposed communication-ef cient adaptive federated learning method, FedCAMS, to show its effectiveness in achieving communication-ef cient adaptive federated learning.

Experimental Setup: We test all federated learning baselines, including ours on CIFAR10 and CIFAR100 datasets (Krizhevsky et al., 2009) using the following two models: (1) ResNet-18 (He et al., 2016), a widely used convolutional neural network model which is commonly trained by SGD; and (2) ConvMixer model (Trockman & Kolter, 2022), which shares similar ideas to vision transformer (Dosovitskiy et al., 2021) to use patch embeddings to preserve locality and similarly is trained via adaptive gradient methods by default. We set in total 100 clients for all federated training experiments. We set the partial participation ratio as 0.1, i.e., in each round, the server picks 10 out of 100 clients to participate in the communication and model update. In each round, the client will perform 3 local epochs of local training with batch size 20. We search for the best training hyper-parameters for each baseline, including ours. Due to the space limit, we leave all the hyper-parameter details as well as the CIFAR-100 experiments in the Appendix.

5.1. FedAMS and Adaptive Federated Optimization

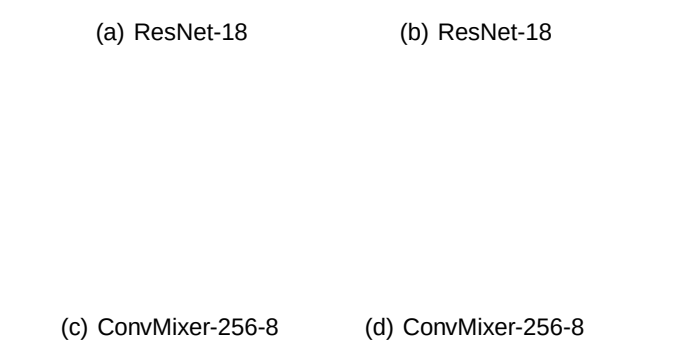
We compare two options of the FedAMS framework with several state-of-the-art adaptive federated learning optimization methods, including: (1) FedAdam (Reddi et al., 2020) (2) FedYogi (Reddi et al., 2020) as well as standard federated baselines: (3) FedAvg (McMahan et al., 2017). Note that the Option 2 for FedAMS is same as FedAMSGrad (Tong et al., 2020). Thus in this section, we denote FedAMS for Option 1 and FedAMSGrad for Option 2 in the general FedAMS framework.

Figure 1 shows the convergence result of FedAMS and other federated learning baselines on training CIFAR-10 dataset with ResNet-18 model and ConvMixer-256-8 model. We compare the training loss and test accuracy against global rounds for each model. For the ResNet-18 model, FedAMS and FedYogi achieve quite similar performances, which are signi cantly better than the other three baselines. In particular, FedAMS performs the best in terms of the nal training loss and test accuracy. On the other hand, FedAMSGrad and FedAdam obtain quite similar results on test accuracy and training loss. FedAvg achieves a slightly better training loss to FedAdam and FedAMSGrad but much higher test accuracy which is close to FedYogi and FedAMS. For the ConvMixer-256-8 model, which is typically trained via adaptive gradient methods, we observe that all adaptive federated optimization methods (FedAdam, FedYogi, FedAMSGrad and FedAMS) achieve much better performance in terms of both training loss and test accuracy than FedAvg. In detail, FedAMS again achieves a signi cantly better result than other baselines. Other adaptive methods, including FedAdam, FedYogi, and FedAMSGrad, have similar convergence behaviour when training the ConvMixer-256-8 model. Such results empirically show the effectiveness of our proposed FedAMS method with max stabilization.

Figure 2 shows the effect of parameter m on the convergence rate by choosing different number of m from $\{5, 10, 20\}$. From Figure 2 we can observe that a larger number of participating clients m in general achieves a faster convergence rate. This veri ed our theoretical results in Section 4.1.

Figure 3 shows the ablation study with different local epochs by choosing different number of local epochs E from $\{3, 10, 30, 100\}$ when training CIFAR-10 data on ResNet-18 with FedAMS optimizer. We observe that large E leads to faster convergence, but larger E does not show a signi cant advantage in achieving a higher test accuracy. We follow FedAvg (McMahan et al., 2017) and FedAdam (Reddi et al., 2020) and set local epoch $E = 3$ by default unless otherwise speci ed.

[4] Note that the q-dependency in 1-bit Adam (Tang et al., 2021) is actually different due to the use of variance-freezed Adam update, i.e., freeze the variance term v_t of Adam update as a constant after a few epochs, which make it resembles momentum SGD.



(a) ResNet-18 (b) ResNet-18

(c) ConvMixer-256-8 (d) ConvMixer-256-8

Figure 1.The learning curves for FedAMS and other federated learning baselines on training CIFAR-10 data (a)(b) show the results for the ResNet-18 model and (c)(d) show the results for the ConvMixer-256-8 model.

(a) ResNet-18 (b) ConvMixer-256-8

Figure 2.The learning curves for FedAMS with different participating number of clients n in training CIFAR-10 data on the ResNet-18 and the ConvMixer-256-8 models.

(a) Training Loss (b) Test Accuracy

Figure 3.The learning curves for FedAMS with different number of local epochs E in training CIFAR-10 data on the ResNet-18 model.

5.2. Communication-Ef cient FedCAMS

Figure 4 shows the convergence results of FedAMS and FedCAMS[5] with different compression strategies on training CIFAR-10 dataset with the ResNet-18 model. It includes comparisons between scaled sign compressor and top-k compressor with compression ratio 2 $f 1=64; 1=128; 1=256g$. We compare the training loss and test accuracy against global rounds and the (pseudo) gradient communication bits[6]. FedAMS, who does not conduct any communication compression, performs the best in terms of training loss yet requires a large volume of communication costs. For our FedCAMS compression methods, sign compressor and top-k compressor with ratio $r = 1=64$ achieve similar performance in terms of test accuracy against the training rounds and obtain the best trade-off between communication ef ciency and model accuracy. Figure 4 (b)(d) show the direct comparison against the communication bits of training ResNet-18 on CIFAR-10. In particular, we can observe that the top-k compressor with a smaller$r$ (i.e., a heavier compression with more information lost), obtains better communication ef ciency but a slower convergence rate. Note that for a d dimensional vector, the overall cost of a scaled sign compressor is $32 + d$ bits, and it is roughly the same communication costs as a top-k compressor[7] with a ratio$r = 1=64$. This veri es our theoretical results in Section 4.2.

Figure 5 shows the convergence results of FedAMS and FedCAMS with the same compression strategies as in Figure 4 on training CIFAR-10 dataset with the ConvMixer-256-8 model. We notice that FedCAMS with the scaled sign compressor achieves roughly the same training loss and test accuracy as FedAMS but with a few orders of magnitude less in communication costs, while other top-k compression models have signi cantly worse performance. Among the top-k compressor trained models, the one with compression ratio $r = 1=64$ still obtains better training loss and test accuracy but higher communication costs. These results suggest that our proposed FedCAMS is communication-ef cient while maintaining high accuracy.

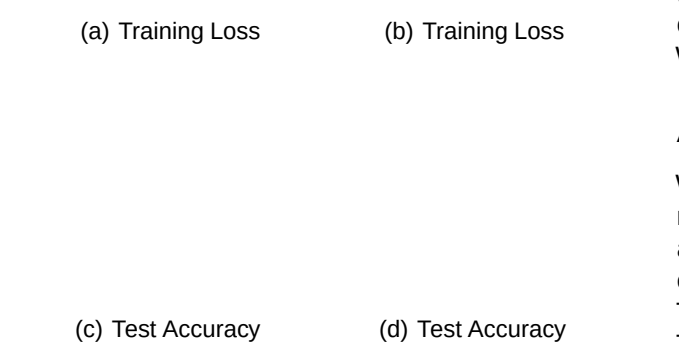
6. Conclusions and Future Work

In this paper, we propose a communication-ef cient compressed federated adaptive gradient optimization framework, FedCAMS, which largely reduces the communication overhead and addresses the adaptivity issue in federated optimization methods. FedCAMS is based on our proposed gen-

[5]Here FedCAMS adopt Option 1 for the nal update step for fairness comparisons (same as FedAMS).

[6]Note that here we only count the client-to-server one-way communications compression.

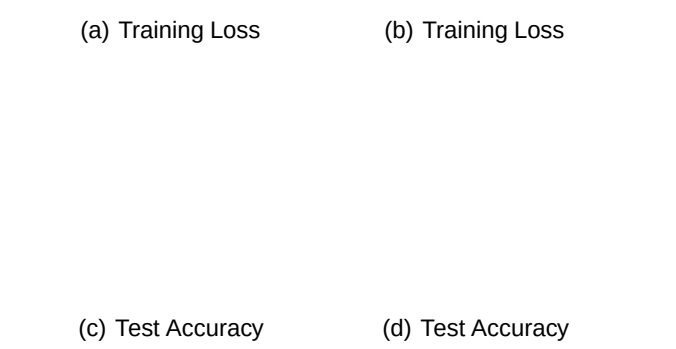
[7]Top-k compressor also needs to communicate about the chosen$k$ locations which roughly double the costs.



(a) Training Loss (b) Training Loss

(c) Test Accuracy (d) Test Accuracy

Figure 4.The learning curves for FedCAMS and uncompressed FedAMS on training CIFAR-10 data on the ResNet-18 model.



(a) Training Loss (b) Training Loss

(c) Test Accuracy (d) Test Accuracy

Figure 5.The learning curves for FedCAMS and uncompressed FedAMS on training CIFAR-10 data on ConvMixer-256-8 model.

eral adaptive federated optimization framework, FedAMS, which contains variants of FedAdam feature max stabilization mechanisms. We present an improved theoretical convergence analysis of adaptive federated optimization, based on which we prove that in the nonconvex stochastic optimization setting, our proposed FedCAMS achieves the same convergence rate as its uncompressed counterpart FedAMS with a few orders of magnitude less communication cost. Experiments on various benchmarks veri ed our theoretical results.

Our current analysis is limited to one-way communication compression from clients to the central server. However, extending our current analysis to two-way communication compression is highly non-trivial as it can be hard to guarantee the distributed global model from server to clients to stay synchronized due to error feedback and the biased compressor, especially in the partial participation setting. We leave it as future work.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by a Seed Grant award from the Institute for Computational and Data Sciences at the Pennsylvania State University as well as Dell Technology AI Infrastructure Level Technologies Grant. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

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

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-sgd: Distributed sgd with quantization, sparsi cation, and local computations. arXiv preprint arXiv:1906.02367, 2019.

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Chen, C., Shen, L., Huang, H., and Liu, W. Quantized adam with error feedback. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–26, 2021a.

Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *Proceedings of the International Joint Conference on Arti cial Intelligence (IJCAI)*, 2020a.

Chen, M., Shlezinger, N., Poor, H. V., Eldar, Y. C., and Cui, S. Communication-ef cient federated learning. *Pro-*

ceedings of the National Academy of Sciences, 118(17), 2021b.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. arXiv preprint arXiv:1808.02941, 2018.

Chen, X., Li, X., and Li, P. Toward communication ef cient adaptive gradient method. In Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, pp. 119–128, 2020b.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research, 12(7), 2011.

Ghosh, A., Hong, J., Yin, D., and Ramchandran, K. Robust federated learning in a heterogeneous environment. arXiv preprint arXiv:1906.06629, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. Federated learning with compression: Uni ed analysis and sharp guarantees. In International Conference on Arti cial Intelligence and Statistics, pp. 2350–2358. PMLR, 2021.

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

Hsu, T.-M. H., Qi, H., and Brown, M. Measuring the effects of non-identical data distribution for federated visual classi cation. arXiv preprint arXiv:1909.06335, 2019.

Jhunjhunwala, D., Gadhikar, A., Joshi, G., and Eldar, Y. C. Adaptive quantization of model updates for communication-ef cient federated learning. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3110–3114. IEEE, 2021.

Jin, R., Huang, Y., He, X., Dai, H., and Wu, T. Stochastic-sign sgd for federated learning with theoretical guarantees. arXiv preprint arXiv:2002.10940, 2020.

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

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In International Conference on Machine Learning, pp. 5132–5143. PMLR, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Konecny, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication ef ciency. arXiv preprint arXiv:1610.05492, 2016.

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

Li, D. and Wang, J. Fedmd: Heterogenous federated learning via model distillation. arXiv preprint arXiv:1910.03581, 2019.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. arXiv preprint arXiv:1905.10497, 2019a.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3):50–60, 2020.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. arXiv preprint arXiv:1907.02189, 2019b.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. Don't use large mini-batches, use local sgd. arXiv preprint arXiv:1808.07217, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Luo, L., Xiong, Y., Liu, Y., and Sun, X. Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint arXiv:1902.09843, 2019.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-ef cient learning of deep networks from decentralized data. Arti cial intelligence and statistics, pp. 1273–1282. PMLR, 2017.

Nishio, T. and Yonetani, R. Client selection for federated learning with heterogeneous resources in mobile edge. In ICC 2019-2019 IEEE International Conference on Communications (ICC), pp. 1–7. IEEE, 2019.

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konecny, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. arXiv preprint arXiv:2003.00295, 2020.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In International Conference on Learning Representations, 2018.

Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. Fedpaq: A communication-ef cient federated learning method with periodic averaging and quantization. In International Conference on Arti cial Intelligence and Statistics, pp. 2021–2031. PMLR, 2020.

Robbins, H. and Monro, S. A stochastic approximation method. The annals of mathematical statistics, pp. 400–407, 1951.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In Fifteenth annual conference of the international speech communication association. Citeseer, 2014.

Stich, S. U. Local sgd converges fast and communicates little. arXiv preprint arXiv:1805.09767, 2018.

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

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsi ed sgd with memory. arXiv preprint arXiv:1809.07599, 2018.

Tang, H., Gan, S., Awan, A. A., Rajbhandari, S., Li, C., Lian, X., Liu, J., Zhang, C., and He, Y. 1-bit adam: Communication ef cient large-scale training with adam's convergence speed. arXiv preprint arXiv:2102.02888, 2021.

Tieleman, T., Hinton, G., et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26–31, 2012.

Tong, Q., Liang, G., and Bi, J. Effective federated adaptive gradient methods with non-iid decentralized data. arXiv preprint arXiv:2009.06557, 2020.

Trockman, A. and Kolter, J. Z. Patches are all you need? arXiv preprint arXiv:2201.09792, 2022.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. Slowmo: Improving communication-ef cient distributed sgd with slow momentum. arXiv preprint arXiv:1910.00643, 2019.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. arXiv preprint arXiv:2007.07481, 2020.

Wang, Y., Lin, L., and Chen, J. Communication-compressed adaptive gradient method for distributed nonconvex optimization. In International Conference on Arti cial Intelligence and Statistics, pp. 6292–6320. PMLR, 2022.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-iid federated learning. arXiv preprint arXiv:2101.11203, 2021.

Yang, Z., Chen, M., Saad, W., Hong, C. S., and Shikh-Bahaei, M. Energy ef cient federated learning over wireless communication networks. IEEE Transactions on Wireless Communications, 20(3):1935–1949, 2020.

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

Zhou, D., Chen, J., Cao, Y., Tang, Y., Yang, Z., and Gu, Q. On the convergence of adaptive gradient methods for nonconvex optimization. arXiv preprint arXiv:1808.05671, 2018.

A. Proof in Section 4.1

A.1. Proof of Theorem 4.4

Similar to previous works studied adaptive methods (Chen et al., 2018; Zhou et al., 2018; Chen et al., 2020a), we introduce a Lyapunov sequence z_t: assume $x_0 = x_1$, for each $t \geq 1$, we have

$$z_t = x_t + \frac{1}{1-\beta_1}(x_t - x_{t-1}) = \frac{1}{1-\beta_1}x_t - \frac{\beta_1}{1-\beta_1}x_{t-1}. \tag{A.1}$$

For the difference of sequence z_t, we have

$$\begin{aligned} z_{t+1} - z_t &= \frac{1}{1-\beta_1}(x_{t+1} - x_t) - \frac{\beta_1}{1-\beta_1}(x_t - x_{t-1}) \\ &= \frac{1}{1-\beta_1}(-\hat{V}_t^{-1/2}m_t) - \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1} \\ &= \frac{1}{1-\beta_1}\hat{V}_t^{-1/2}\beta_1 m_{t-1} + (1-\beta_1)\hat{V}_t^{-1/2}g_t - \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1} \\ &= \hat{V}_t^{-1/2}g_t - \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}(I - \hat{V}_t^{1/2}\hat{V}_t^{-1/2})m_{t-1}. \end{aligned}$$

Since f is L-smooth, taking conditional expectation at time t, we have

$$\begin{aligned} &\mathbb{E}[f(z_{t+1})] - f(z_t) \\ &\leq \mathbb{E}[\langle \nabla f(z_t), z_{t+1} - z_t\rangle] + \frac{L}{2}\mathbb{E}[\|z_{t+1} - z_t\|^2] \\ &= \mathbb{E}\langle \nabla f(z_t), -\hat{V}_t^{-1/2}g_t\rangle - \mathbb{E}\langle \nabla f(z_t), \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}(I - \hat{V}_t^{1/2}\hat{V}_t^{-1/2})m_{t-1}\rangle \\ &\quad + \frac{L}{2}\mathbb{E}\left\|-\hat{V}_t^{-1/2}g_t + \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}(I - \hat{V}_t^{1/2}\hat{V}_t^{-1/2})m_{t-1}\right\|^2 \\ &= \underbrace{\mathbb{E}\langle \nabla f(x_t), -\hat{V}_t^{-1/2}g_t\rangle}_{I_1} \underbrace{- \mathbb{E}\langle \nabla f(z_t), \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}(I - \hat{V}_t^{1/2}\hat{V}_t^{-1/2})m_{t-1}\rangle}_{I_2} \\ &\quad + \underbrace{\frac{L}{2}\mathbb{E}\left\|-\hat{V}_t^{-1/2}g_t + \frac{\beta_1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}(I - \hat{V}_t^{1/2}\hat{V}_t^{-1/2})m_{t-1}\right\|^2}_{I_3} + \underbrace{\mathbb{E}\langle \nabla f(z_t) - \nabla f(x_t), -\hat{V}_t^{-1/2}g_t\rangle}_{I_4}; \end{aligned} \tag{A.2}$$

here we recall the notation $\hat{V}_t = \mathrm{diag}(\hat{v}_t) = \mathrm{diag}(\max(\hat{v}_{t-1}, v_t, \epsilon))$.

Bounding I_1: We have

$$\begin{aligned} I_1 &= \mathbb{E}\langle \nabla f(x_t), -\frac{g_t}{\sqrt{\hat{v}_t + \epsilon}}\rangle \\ &\quad \mathbb{E}\langle \nabla f(x_t), -\frac{g_t}{\sqrt{v_t + \epsilon}}\rangle \\ &= \frac{1}{2}\mathbb{E}\langle \nabla f(x_t), -\frac{g_t}{\sqrt{\hat{v}_{t-1} + \epsilon}}\rangle + \frac{1}{2}\mathbb{E}\langle \nabla f(x_t), -\frac{g_t}{\sqrt{v_t + \epsilon}}\rangle - \frac{1}{2}\mathbb{E}\langle \nabla f(x_t), -\frac{g_t}{\sqrt{\hat{v}_{t-1} + \epsilon}}\rangle; \end{aligned} \tag{A.3}$$

where the rst inequality follows by the fact that $\hat{v}_t \geq \frac{v_t + \epsilon}{2}$. For the second term in (A.3), we have

$$\begin{aligned}
&\sqrt{2}\, E\, \langle \nabla f(x_t), \frac{\eta}{\sqrt{v_t + \epsilon}} - \frac{\eta}{\sqrt{2v_{t-1} + \epsilon}} \rangle \\
&\leq \sqrt{2}\, \|\nabla f(x_t)\| E\, \Big\| \frac{\eta}{\sqrt{v_t + \epsilon}} - \frac{\eta}{\sqrt{2v_{t-1} + \epsilon}} \Big\| \cdot \|\delta_t\| \\
&\leq \frac{\sqrt{2(1-\beta_2)}G}{\epsilon} E[\|\delta_t\|^2];
\end{aligned} \tag{A.4}$$

where the second inequality follows from Lemma C.1 and C.4, and we will further apply the bound $E[\|\delta_t\|^2]$ following from Lemma C.5. For the rst term in (A.3), we have

$$\begin{aligned}
&\sqrt{2}\, E\, \langle \nabla f(x_t), \frac{\eta}{\sqrt{2v_{t-1} + \epsilon}} \rangle \\
&= \sqrt{2}\, E\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \delta_t + \beta_1 m_{t-1} - \beta_1 m_{t-1} \rangle \\
&= \sqrt{2}\beta_1 E\, \Big\| \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 + \sqrt{2}\, E\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \delta_t + \beta_1 m_{t-1} \rangle \\
&= \sqrt{2}\beta_1 E\, \Big\| \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 + \sqrt{2}\, \Big\langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, E\, \frac{1}{m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i \Big\rangle + \beta_1 \langle \nabla f(x_t), m_{t-1} \rangle \\
&= \sqrt{2}\beta_1 E\, \Big\| \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 + \sqrt{2}\, \Big\langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, E\, \frac{1}{m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i \Big\rangle + \frac{\beta_1}{m} \sum_{i=1}^{m} \langle \nabla F_i(x_t), \cdots \rangle;
\end{aligned} \tag{A.5}$$

where the third equality follows the local update rule. For the last term in (A.5), we have

$$\begin{aligned}
&\sqrt{2}\, \Big\langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, E\, \frac{\beta_1}{m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i + \frac{\beta_1}{m} \sum_{i=1}^{m} \nabla F_i(x_t) \Big\rangle \\
&= \sqrt{2}\, \Big\langle \frac{\sqrt{\beta_1}\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \frac{\sqrt{\beta_1}}{Km} \frac{1}{\sqrt{2v_{t-1} + \epsilon}} E \sum_{i=1}^{m} \sum_{k=0}^{K-1} (\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)) \Big\rangle \\
&= \frac{\sqrt{2}\beta_1}{2} E\, \Big\| \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 + \frac{\sqrt{2}\beta_1}{2Km^2} E\, \Big\| \frac{1}{\sqrt{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} (\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)) \Big\|^2 \\
&\quad \frac{\sqrt{2}\beta_1}{2Km^2} E\, \Big\| \frac{1}{\sqrt{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \Big\|^2 \\
&\leq \frac{\sqrt{2}\beta_1}{2} E\, \Big\| \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 + \frac{\sqrt{2}}{2m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} E\, \Big\| \frac{\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 \\
&\quad \frac{\sqrt{2}\beta_1}{2Km^2} E\, \Big\| \frac{1}{\sqrt{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \Big\|^2 \\
&\leq \frac{\sqrt{2}\beta_1}{2} E\, \Big\| \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 + \frac{\sqrt{2}\beta_1 L^2}{2m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} E\, \Big\| \frac{x_{t,k}^i - x_t}{\sqrt{2v_{t-1} + \epsilon}} \Big\|^2 \\
&\quad \frac{\sqrt{2}\beta_1}{2Km^2} E\, \Big\| \frac{1}{\sqrt{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \Big\|^2;
\end{aligned} \tag{A.6}$$

where the second equation follows from $\langle x, y \rangle = \frac{1}{2}[\|x\|^2 + \|y\|^2 - \|x - y\|^2]$, the rst inequality holds by applying Cauchy-Schwarz inequality, the second inequality follows from Assumption 4.1.

Hence by applying Lemma C.9 with the local learning rate condition: $\eta \le \frac{1}{8KL}$, we have

$$\frac{p}{2}\,\frac{r f(x_t)}{\sqrt{v_{t-1}+}};\, E\left[\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1} g_{t,k}^{i} + \frac{1}{m}\sum_{i=1}^{m} r F_i(x_t)\right]$$
$$\frac{3p}{4}\,\frac{2}{1}K E\left[\frac{r f(x_t)}{\sqrt{v_{t-1}+}}\right]^2 + \frac{5}{p}\,\frac{3K^2L^2}{2}\left(\frac{2}{1}+6K\frac{2}{g}\right)\frac{p}{2}\,\frac{2}{1}E\left[\frac{1}{\sqrt{v_{t-1}+}}\right]\sum_{i=1}^{m}\sum_{k=0}^{K-1} r F_i(x_{t,k}^{i})^2 : \tag{A.7}$$

Then merging pieces together, we have

$$\begin{aligned} I_1 \le\ & \frac{p}{2}\,\frac{2}{1}K E\left[\frac{r f(x_t)}{\sqrt{v_{t-1}+}}\right]^2 + \frac{5}{p}\,\frac{3K^2L^2}{2}\left(\frac{2}{1}+6K\frac{2}{g}\right) \\ & \frac{p}{2}\,\frac{2}{1}E\left[\frac{1}{\sqrt{v_{t-1}+}}\right]\sum_{i=1}^{m}\sum_{k=0}^{K-1} r F_i(x_{t,k}^{i})^2 + \frac{p}{2(1-2)}G E[k_t k^2] \\ & \frac{1}{4}K E\left[\frac{r f(x_t)}{\sqrt{v_{t-1}+}}\right]^2 + \frac{5}{p}\,\frac{3K^2L^2}{2}\left(\frac{2}{1}+6K\frac{2}{g}\right) \\ & \frac{1}{2}\,\frac{2}{1}E\left[\frac{1}{\sqrt{v_{t-1}+}}\right]\sum_{i=1}^{m}\sum_{k=0}^{K-1} r F_i(x_{t,k}^{i})^2 + \frac{p}{2(1-2)}G E[k_t k^2]: \end{aligned} \tag{A.8}$$

Bounding I_2: The bound for I_2 mainly follows by the update rule and de nition of virtual sequence z_t,

$$\begin{aligned} I_2 =\ & E\left\langle r f(z_t), \frac{1}{1-1}\left(\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right) m_{t-1}\right\rangle \\ =\ & E\left\langle r f(z_t) - f(x_t) + r f(x_t), \frac{1}{1-1}\left(\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right) m_{t-1}\right\rangle \\ & E\left\langle r f(x_t), \frac{1}{1-1}\left(\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right) m_{t-1}\right\rangle \\ & + {}^2 L E\left[\frac{1}{\sqrt{v_{t-1}}}\,\frac{1}{1-1} m_{t-1}\right]\left[\frac{1}{1-1}\left(\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right) m_{t-1}\right] \\ & \frac{1}{1-1}\,\frac{2}{1}KG^2 E\left[\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right]_1 + {}^2\frac{2}{(1-1)^2}L\,\frac{2}{1}K^2G^2\,{}^{1=2}E\left[\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right]_1; \end{aligned} \tag{A.9}$$

where the last inequality holds by applying Lemma C.4 and the fact of v_t .

Bounding I_3: It can be bounded as follows:

$$\begin{aligned} I_3 =\ & \frac{{}^2 L}{2} E\left[\sqrt{v_t}^{1=2}\ {}_t + \frac{1}{1-1}\left(\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right) m_{t-1}\right]^2 \\ & {}^2 L E\left[\sqrt{v_t}^{1=2}\ {}_t\right]^2 + {}^2 L E\left[\frac{1}{1-1}\left(\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right) m_{t-1}\right]^2 \\ & {}^2 L E\left[\sqrt{v_t}^{1=2}\ {}_t\right]^2 + {}^2 L\,\frac{2}{(1-1)^2}\,{}^2 K^2 G^2 E\left[\sqrt{v_t}^{1=2} - \sqrt{v_{t-1}}^{1=2}\right]^2; \end{aligned} \tag{A.10}$$

where the rst inequality follows by Cauchy-Schwarz inequality, and the second one follows by Lemma C.4.

Bounding I_4:

$$\begin{aligned} I_4 &= E\quad r f(z_t)\quad r\quad f(x_t);\ \hat{V}_t^{1=2}\ _t \\ &\quad E\ kr f(z_t)\quad r\quad f(x_t)k\quad \hat{V}_t^{1=2}\ _t \\ &\quad LE\ kz_t\quad x_tk\quad \hat{V}_t^{1=2}\ _t \\ &\quad \frac{^2L}{2}E\quad \hat{V}_t^{1=2}\ _t\ ^2 + \frac{^2L}{2}E\quad \frac{1}{1\quad _1}\hat{V}_t^{1=2}m_t\ ^2; \end{aligned}$$

where the rst inequality holds by the fact that $ha; bi \quad kakkbk$, the second one follows from Assumption 4.1 and the third one holds by the de nition of virtual sequence z_t and the fact of $kakkbk \quad \frac{1}{2}kak^2 + \frac{1}{2}kbk^2$. Then summing I_4 over $t = 1; \quad ; T$, we have

$$\begin{aligned} \sum_{t=1}^{T} I_4 \quad & \frac{^2L}{2}\sum_{t=1}^{T} E[k\ _tk^2] + \frac{^2L}{2}\sum_{t=1}^{T} E\quad \frac{1}{1\quad _1}m_t\ ^2 \\ & \frac{^2L}{2}\sum_{t=1}^{T} E[k\ _tk^2] + \frac{^2L}{2}\frac{^2_1}{(1\quad _1)^2}\sum_{t=1}^{T} E[km_tk^2]: \end{aligned} \tag{A.11}$$

By Lemma C.7, we have

$$\sum_{t=1}^{T} E[km_tk^2] \quad \frac{TK\ ^2_l}{m}\ ^2_l + \frac{^2_l}{m^2}\sum_{t=1}^{T} E\quad \sum_{i=1}^{m}\sum_{k=0}^{K\ 1} r F_i(x^i_{t;k})\ ^2:$$

Therefore, the summation of I_4 term is bounded by

$$\begin{aligned} \sum_{t=1}^{T} I_4 \quad & \frac{^2L}{2}\sum_{t=1}^{T} E[k\ _tk^2] + \frac{^2_1}{(1\quad _1)^2}\frac{^2L}{2}\frac{^2_l}{m^2}\sum_{t=1}^{T} E\quad \sum_{i=1}^{m}\sum_{k=0}^{K\ 1} r F_i(x^i_{t;k})\ ^2 \\ & + \frac{^2_1}{(1\quad _1)^2}\frac{^2L}{2}\frac{TK\ ^2_l}{m}\ ^2_l: \end{aligned} \tag{A.12}$$

Merging pieces together: Substituting (A.8), (A.9) and (A.10) into (A.2), summing over from $t = 1$ to T and then adding (A.11), we have

$$\begin{aligned} E[f(z_{T+1})] \quad f(z_1) = & \sum_{t=1}^{T}[I_1 + I_2 + I_3 + I_4] \\ & \frac{_lK}{4}\sum_{t=1}^{T} E\quad p\frac{r f(x_t)}{\sqrt[4]{2v_t\ _1 + }}\ ^2 + \frac{5\ ^3_lK\ ^2L\ ^2T}{p\overline{2}}(\ ^2_l + 6K\ ^2_g) + \frac{p\overline{2(1\quad _2)}\ G}{}\sum_{t=1}^{T} E[k\ _tk^2] \\ & \frac{_l}{2Km^2}\sum_{t=1}^{T} E\quad p\frac{1}{\sqrt[4]{2v_t\ _1 + }}\sum_{i=1}^{m}\sum_{k=0}^{K\ 1} r F_i(x^i_{t;k}))\ ^2 \\ & + \frac{1}{1\quad _1}\ _lKG\ ^2\sum_{t=1}^{T} E\quad \hat{V}_t^{1=2}\ _1\quad \hat{V}_t^{1=2}\ _1 + \frac{^2_1}{(1\quad _1)^2}\frac{^2\ ^2_lK\ ^2G^2}{p\ \ }\sum_{t=1}^{T} E\quad \hat{V}_t^{1=2}\ _1\quad \hat{V}_t^{1=2}\ _1 \\ & + \frac{^2_1}{(1\quad _1)^2}\ ^2\ ^2_lK\ ^2LG^2\sum_{t=1}^{T} E\quad \hat{V}_t^{1=2}\ _1\quad \hat{V}_t^{1=2}\ ^2 + \ ^2L\sum_{t=1}^{T} E\quad \hat{V}_t^{1=2}\ _t\ ^2 \\ & + \frac{^2L}{2}\sum_{t=1}^{T} E[k\ _tk^2] + \frac{^2L}{2}\frac{^2_1}{(1\quad _1)^2}\sum_{t=1}^{T} E[km_tk^2]: \end{aligned} \tag{A.13}$$

By applying Lemma C.5 into all terms containing the second moment estimate of model differences in (A.13), and using the fact that $(\sqrt{\hat{v}_{t-1}+\epsilon})^{-1}\|x\| \le (\sqrt{\epsilon})^{-1}\|x\|$ and $\|\frac{x}{\sqrt{\hat{v}_t+\epsilon}}\| \le \epsilon^{-1/2}\|x\|$, we have

$$
\begin{aligned}
&\mathbb{E}[f(z_{T+1})] - f(z_1) \\
&\le \frac{\eta K}{4}\sum_{t=1}^{T}\mathbb{E}\Big\|\frac{\nabla f(x_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}\Big\|^2 + \frac{5\eta^3 K^2 L^2 T}{2}(\sigma_l^2 + 6K\sigma_g^2) + \frac{1}{1-\beta_1}\frac{\eta K G^2 d}{\sqrt{\epsilon}} \\
&\quad + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{2\beta^2\eta^2 K^2 L G^2 d}{\epsilon\sqrt{\epsilon}}\frac{\eta}{2Km^2}\sum_{t=1}^{T}\mathbb{E}\Big\|\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t;k}^i)\Big\|^2 \\
&\quad + \Big(2L + \frac{\sigma^2 L}{2} + \sqrt{2(1-\beta_2)}G\Big)\frac{KT}{m}\sigma_l^2 + \frac{\sigma_l^2}{m^2}\sum_{t=1}^{T}\mathbb{E}\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t;k}^i)\Big\|^2 \\
&\quad + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\sigma^2 L}{2}\frac{\sigma_l^2}{m^2}\sum_{t=1}^{T}\mathbb{E}\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t;k}^i)\Big\|^2 + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\sigma^2 L}{2}\frac{TK}{m}\sigma_l^2 \\
&\le \frac{\eta K}{4\sqrt{\epsilon^2\sigma_l^2 K^2 G^2 + \epsilon}}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{5\eta^3 K^2 L^2 T}{2}(\sigma_l^2 + 6K\sigma_g^2) \\
&\quad + \frac{1}{1-\beta_1}\frac{\eta K G^2 d}{\sqrt{\epsilon}} + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{2\beta^2\eta^2 K^2 L G^2 d}{\epsilon\sqrt{\epsilon}} \\
&\quad + \Big(2L + \frac{\sigma^2 L}{2} + \sqrt{2(1-\beta_2)}G + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\sigma^2 L}{2}\frac{KT}{m}\sigma_l^2 \sum_{t=1}^{T}\mathbb{E}\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t;k}^i)\Big\|^2 \\
&\quad \frac{\eta}{2\sqrt{\epsilon^2\sigma_l^2 K^2 G^2 + \epsilon}Km^2}\Big(2L + \frac{\sigma^2 L}{2} + \sqrt{2(1-\beta_2)}G + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\sigma^2 L}{2}\frac{\sigma_l^2}{m^2}\Big) \\
&\le \frac{\eta K}{4\sqrt{\epsilon^2\sigma_l^2 K^2 G^2 + \epsilon}}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] + \frac{5\eta^3 K^2 L^2 T}{2}(\sigma_l^2 + 6K\sigma_g^2) \\
&\quad + \frac{1}{1-\beta_1}\frac{\eta K G^2 d}{\sqrt{\epsilon}} + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{2\beta^2\eta^2 K^2 L G^2 d}{\epsilon\sqrt{\epsilon}} \\
&\quad + \Big(2L + \frac{\sigma^2 L}{2} + \sqrt{2(1-\beta_2)}G + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\sigma^2 L}{2}\frac{KT}{m}\sigma_l^2. \qquad \text{(A.14)}
\end{aligned}
$$

The last inequality holds due to additional constraint of local learning rate with the inequality $\frac{\eta}{2\sqrt{\epsilon^2\sigma_l^2 K^2 G^2 + \epsilon}Km^2}\big(\frac{3\sigma^2 L}{2} + \sqrt{2(1-\beta_2)}G + \frac{\beta_1^2}{(1-\beta_1)^2}\frac{\sigma^2 L}{2}\frac{\sigma_l^2}{m^2}\big) \ge 0$, thus we obtain the constraint $\eta \le \frac{\sqrt{\epsilon}}{K\sqrt{\epsilon^2\sigma_l^2 K^2 G^2 + [(3+\beta_1^2)L + 2\sqrt{2(1-\beta_2)}G]}}$. Hence we have

$$
\begin{aligned}
&\frac{\eta K}{4\sqrt{\epsilon^2\sigma_l^2 K^2 G^2 + \epsilon}}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(x_t)\|^2] \\
&\le \frac{f(z_0) - \mathbb{E}[f(z_T)]}{T} + \frac{5\eta^3 K^2 L^2}{2}(\sigma_l^2 + 6K\sigma_g^2) + [(3+\beta_1^2)L + 2\sqrt{2(1-\beta_2)}G]\frac{K\sigma_l^2}{2m}\sigma_l^2 \\
&\quad + \frac{C_1}{T}\frac{\eta K G^2 d}{\sqrt{\epsilon}} + \frac{2C_1^2\beta^2\eta^2 K^2 L G^2 d}{T\epsilon\sqrt{\epsilon}}. \qquad \text{(A.15)}
\end{aligned}
$$

Therefore,

$$
\min_{t\in[T]}\mathbb{E}[\|\nabla f(x_t)\|^2] \le 4\sqrt{\frac{\epsilon^2\sigma_l^2 K^2 G^2 + \epsilon}{\eta K}}\Big(\frac{f_0 - f^*}{T} + \frac{\eta}{T} + \eta\Big), \qquad \text{(A.16)}
$$

where $\eta = \frac{C_1\eta G^2 d}{\sqrt{\epsilon}} + \frac{2C_1^2\beta^2\eta^2 KLG^2 d}{\epsilon\sqrt{\epsilon}}$; $\eta = \frac{5\eta^2 KL^2}{2}(\sigma_l^2 + 6K\sigma_g^2) + [(3+\beta_1^2)L + 2\sqrt{2(1-\beta_2)}G]\frac{\sigma_l^2}{2m}\sigma_l^2$, where $C_1 = \frac{1}{1-\beta_1}$.

A.2. Proof of Corollary 4.6

If we pick $_l = (\ \frac{1}{\sqrt{TK}})$ and $\ = (\ \sqrt{Km})$, we have $\min_{t2[T]} E[kr\ f(x_t)k^2] = O(\frac{1}{\sqrt{TKm}})$.

A.3. Proof of Theorem 4.9

Notations and equations For partial participation, i.e., $jS_t j = n; 8t\ 2\ [T]$. The global model difference is the average of local model difference from the subset S_t, i.e., $_t = \frac{1}{n}\sum_{i2S_t} {}^t_i$. Denote $_t = \frac{1}{m}\sum_{i=1}^{m} {}^t_i$, and for convenience, we follow the previous notation of $\hat{V}_t = \mathrm{diag}(\hat{v}_t + \ )$. Next we show that the global model difference $_t$ is an unbiased estimator of $_t$:

$$E_{S_t}[\ _t] = \frac{1}{n}E_{S_t}[\sum_{i=1}^{n} {}^t_{w_i}] = E_{S_t}[\ {}^t_{w_1}] = \frac{1}{m}\sum_{i=1}^{m} {}^t_i = \ _t: \tag{A.17}$$

Define the virtual sequence z_t same as previous: assume $x_0 = x_1$, for each $t\ \ 1$, we have

$$z_t = x_t + \frac{1}{1\ \ _1}(x_t\ \ x_{t\ 1}) = \frac{1}{1\ \ _1}x_t\ \ \frac{1}{1\ \ _1}x_{t\ 1}; \tag{A.18}$$

$$z_{t+1}\ \ z_t = \hat{V}_t\ _t\ \ \frac{1}{1\ \ _1}\hat{V}_t^{1=2}\ _1\hat{V}_t^{1=2}\ m_{t\ 1}: \tag{A.19}$$

By Assumption 4.1, we have

$$\begin{aligned}
&E[f(z_{t+1})]\ \ f(z_t) \\
&\quad E\ \ r\ f(z_t);\ \hat{V}_t^{1=2}\ _t\ \ \ E\ \ r\ f(z_t);\ \frac{1}{1\ \ _1}\hat{V}_t^{1=2}\ _1\hat{V}_t^{1=2}\ m_{t\ 1} \\
&\quad + \frac{^2L}{2}E\ \ \hat{V}_t^{1=2}\ _t\ \ \frac{1}{1\ \ _1}\hat{V}_t^{1=2}\ _1\hat{V}_t^{1=2}\ m_{t\ 1}\ ^2 \\
&= \underbrace{E\ \ r\ f(x_t);\ \hat{V}_t^{1=2}\ _t}_{I_1^0}\ \ \underbrace{E\ \ r\ f(z_t);\ \frac{1}{1\ \ _1}\hat{V}_t^{1=2}\ _1\hat{V}_t^{1=2}\ m_{t\ 1}}_{I_2^0} \\
&\quad + \frac{^2L}{2}\underbrace{E\ \ \hat{V}_t^{1=2}\ _t\ \ \frac{1}{1\ \ _1}\hat{V}_t^{1=2}\ _1\hat{V}_t^{1=2}\ m_{t\ 1}\ ^2}_{I_3^0} + \underbrace{E\ \ r\ f(z_t)\ \ r\ f(x_t);\ \hat{V}_t^{1=2}\ _t}_{I_4^0}\ :
\end{aligned} \tag{A.20}$$

Since $_t$ is an unbiased estimator of $_t$, the main difference of convergence analysis for partial participation cases is bounding $E[k\ _t k^2]$.

Note that the bound for I_2^0 is exactly the same as the bound for I_2. For the corresponding three terms I_1^0, I_3^0 and I_4^0 which include the second-order momentum estimate $\hat{v}_t$. For I_1^0, we have

$$\begin{aligned}
I_1^0 &= E\ \ r\ f(x_t);\ \frac{\sqrt{_t}}{\sqrt{\hat{v}_t}\ \sqrt{}} \\
&\quad E\ \ r\ f(x_t);\ \frac{\sqrt{2}\ \sqrt{_t}}{\sqrt{v_t + }} \\
&= E\ \ r\ f(x_t);\ \frac{\sqrt{2}\ \sqrt{_t}}{\sqrt{_2 v_{t\ 1} + }}\ \ + E\ \ r\ f(x_t);\ \frac{\sqrt{2}\ \sqrt{_t}}{\sqrt{v_t + }}\ \ \frac{\sqrt{2}\ \sqrt{_t}}{\sqrt{_2 v_{t\ 1} + }}\ :
\end{aligned} \tag{A.21}$$

The first term in (A.21) does not change in partial participation scheme. The second term is changed due to the variance of $_t$ changes. For the second term I_1^0, we have

$$\sqrt{2}\ E\ \ r\ f(x_t);\ \frac{\sqrt{_t}}{\sqrt{v_t + }}\ \ \frac{\sqrt{_t}}{\sqrt{_2 v_{t\ 1} + }}\ \ \ \frac{\sqrt{2(1\ \ _2)}\ G}{}E[k\ _t k^2]: \tag{A.22}$$

For I_3^0, we have

$$\sum_{t=1}^{T} I_3^0 \quad \frac{{}^2L}{} \sum_{t=1}^{T} E[k \ {}_t k^2] + {}^2L \frac{{}_1^2}{(1 \quad {}_1)^2} \ {}_l^2 K^2 G^2 \sum_{t=1}^{T} E \quad \Big\| {}_t \ {}_1^{1=2} \quad {}_t \ {}^{1=2} \Big\|^2 ; \tag{A.23}$$

and for I_4^0, similar to (A.11), we have

$$\sum_{t=1}^{T} I_4^0 \quad \frac{{}^2L}{2} \sum_{t=1}^{T} E[k \ {}_t k^2] + \frac{{}^2L}{2} \frac{{}_1^2}{(1 \quad {}_1)^2} \sum_{t=1}^{T} E[k m_t k^2]: \tag{A.24}$$

From Lemma C.8, we have

$$\sum_{t=1}^{T} E[k m_t k^2] \quad \frac{KT \ {}_l^2}{n} \ {}_l^2 + \frac{{}_l^2}{n^2} \sum_{t=1}^{T} E \quad \Big\| \sum_{i2S_t} \sum_{k=0}^{K \ 1} r \ F_i(x_{t;k}^i) \Big\|^2 : \tag{A.25}$$

Then substituting (A.25) into (A.24), we have

$$\begin{aligned} &\sum_{t=1}^{T} I_4^0 \quad \frac{{}^2L}{2} \sum_{t=1}^{T} E[k \ {}_t k^2] + \frac{{}_1^2}{(1 \quad {}_1)^2} \frac{{}^2 \ {}_l^2 L}{2n^2} \sum_{t=1}^{T} E \quad \Big\| \sum_{i2S_t} \sum_{k=0}^{K \ 1} r \ F_i(x_{t;k}^i) \Big\|^2 + \frac{{}_1^2}{(1 \quad {}_1)^2} \frac{{}^2 \ {}_l^2 KTL}{2n} \ {}_l^2 \\ &\quad \frac{{}^2L}{2} \sum_{t=1}^{T} E[k \ {}_t k^2] + \frac{{}_1^2}{(1 \quad {}_1)^2} \frac{{}^2 \ {}_l^2 L}{2n^2} \sum_{t=1}^{T} E \quad \Big\| \sum_{i=1}^{m} \sum_{k=0}^{K \ 1} Pfi \ 2 \ S_t g r \ F_i(x_{t;k}^i) \Big\|^2 \\ &\quad + \frac{{}_1^2}{(1 \quad {}_1)^2} \frac{{}^2 \ {}_l^2 KTL}{2n} \ {}_l^2; \end{aligned} \tag{A.26}$$

where we will further apply the bound for $E[k \ {}_t k^2]$ following by Lemma C.6. The second term in (A.26) can be bounded from (C.10). Therefore, summing up (A.22), (A.23) and (A.9), summing over from $t = 1$ to T, then adding (A.24), we have

$$\begin{aligned} E[f(z_{T+1})] \quad f(z_1) = &\sum_{t=1}^{T} [I_1^0 + I_2 + I_3^0 + I_4^0] \\ &\frac{{}_l K}{4} \sum_{t=1}^{T} E \quad \Big\| \frac{r \ f(x_t)}{\sqrt{{}_2 v_t \ {}_1 + }} \Big\|^2 + \frac{5 \ {}_l^3 K^2 L^2 T}{p} (\ {}_l^2 + 6K \ {}_g^2) + \frac{p}{2(1 \quad {}_2)} G \sum_{t=1}^{T} E[k \ {}_t k^2] \\ &\frac{{}_l}{2Km^2} \sum_{t=1}^{T} E \quad \Big\| \frac{1}{\sqrt{{}_2 v_t \ {}_1 + }} \sum_{i=1}^{m} \sum_{k=0}^{K \ 1} r \ F_i(x_{t;k}^i) \Big\|^2 \\ &+ \frac{1}{1 \quad {}_1} \ {}_l K G^2 \sum_{t=1}^{T} E \quad \Big\| {}_t \ {}_1^{1=2} \quad {}_t \ {}^{1=2} \Big\|_1 + \frac{{}_1^2}{(1 \quad {}_1)^2} \frac{{}^2 \ {}_l^2 K^2 G^2}{p} \sum_{t=1}^{T} E \quad \Big\| {}_t \ {}_1^{1=2} \quad {}_t \ {}^{1=2} \Big\|_1 \\ &+ \frac{{}_1^2}{(1 \quad {}_1)^2} \ {}^2 \ {}_l^2 K^2 LG^2 \sum_{t=1}^{T} E \quad \Big\| {}_t \ {}_1^{1=2} \quad {}_t \ {}^{1=2} \Big\|^2 + {}^2L \sum_{t=1}^{T} E \quad \Big\| {}_t \ {}^{1=2} \quad {}_t \Big\|^2 \\ &+ \frac{{}^2L}{2} \sum_{t=1}^{T} E[k \ {}_t k^2] + \frac{{}^2L}{2} \frac{{}_1^2}{(1 \quad {}_1)^2} \sum_{t=1}^{T} E[k m_t k^2]: \end{aligned} \tag{A.27}$$

By applying Lemma C.5 into all terms containing the second moment estimate of model difference in (A.27), using the fact that $\Big(p \overline{{}_2 K^2 G^2 + } \Big) \ {}^1 k x_t \ x_{t \ 1} k \quad \Big(p \overline{{}_2 \ {}_l^2 K^2 G^2 + } \Big) \ {}^1 k x_t \ x_{t \ 1} k \quad \Big\| \frac{x_t \ x_{t \ 1}}{\sqrt{{}_2 v_t + }} \Big\|^{1=2} k x_t \ x_{t \ 1} k$, and applying Lemma C.8, we

have

$$
\begin{aligned}
&E[f(z_{T+1})] \quad f(z_1) \\
&\quad \frac{p}{4}\frac{{}_{l}K}{{}_{2}{}_{l}^{2}K^2G^2+} \sum_{t=1}^{T} E[kr\ f(x_t)k^2] + \frac{5\ {}_{l}^{3}K^2L^2T}{p\overline{2}}({}_{l}^{2}+6K\ {}_{g}^{2}) \\
&\quad + \frac{1}{1\quad{}_{1}}\frac{{}_{l}KG^2d}{p-} + \frac{{}_{1}^{2}}{(1\quad{}_{1})^2}\frac{2\ {}_{l}^{2}K^2LG^2d}{} \\
&\quad + \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{KT\ {}_{l}^{2}}{n}\ {}_{l}^{2}\sum_{t=1}^{T} E\left[\sum_{i=1}^{m}\frac{K}{K}\sum_{k=0}^{1} r\ F_i(x_{t;k}^{i})\right]^2 \\
&\quad \frac{p}{2}\frac{{}_{l}}{{}_{2}K^2G^2+Km^2} \quad \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{{}_{l}^{2}(n\quad 1)}{mn(m\quad 1)} \\
&\quad + \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{{}_{l}^{2}(m\quad n)}{mn(m\quad 1)}\ 15mK^3L^3\ {}_{l}^{2}({}_{l}^{2}+6K\ {}_{g}^{2})T \\
&\quad + (90mK^4L^2\ {}_{l}^{2}+3mK^2)\sum_{t=1}^{T} E[kr\ f(x_t)k^2] + 3mK^2T\ {}_{g}^{2}\ ; \qquad \text{(A.28)}
\end{aligned}
$$

then we have

$$
\begin{aligned}
&E[f(z_{T+1})] \quad f(z_1) \\
&\quad \frac{p}{4}\frac{{}_{l}K}{{}_{2}{}_{l}^{2}K^2G^2+} \sum_{t=1}^{T} E[kr\ f(x_t)k^2] + \frac{5\ {}_{l}^{3}K^2L^2T}{p\overline{2}}({}_{l}^{2}+6K\ {}_{g}^{2}) + \frac{1}{1\quad{}_{1}}\frac{{}_{l}KG^2d}{p-} \\
&\quad + \frac{{}_{1}^{2}}{(1\quad{}_{1})^2}\frac{2\ {}_{l}^{2}K^2LG^2d}{} + \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{KT\ {}_{l}^{2}}{n}\ {}_{l}^{2} \\
&\quad + \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{{}_{l}^{2}(m\quad n)}{mn(m\quad 1)}\ 15mK^3L^3\ {}_{l}^{2}({}_{l}^{2}+6K\ {}_{g}^{2})T \\
&\quad + (90mK^4L^2\ {}_{l}^{2}+3mK^2)\sum_{t=1}^{T} E[kr\ f(x_t)k^2] + 3mK^2T\ {}_{g}^{2}\ : \qquad \text{(A.29)}
\end{aligned}
$$

By adopting additional constraint of local learning rate, with the inequality $\frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{{}_{l}^{2}(n\quad 1)}{mn(m\quad 1)}\quad \frac{{}_{l}}{2}\frac{}{{}_{2}K^2G^2+Km^2}\quad 0$, thus we obtain the constraint ${}_{l}\ \frac{n(m\quad 1)}{m(n\quad 1)}\frac{}{{}_{2}K^2G^2+K\ (3\ L\ +C_1^2\ L\ +2}\frac{}{2(1\quad{}_{2})G)}$, and we further need satis es $\frac{p}{4}\frac{{}_{l}K}{{}_{2}{}_{l}^{2}K^2G^2+}\quad \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{{}_{l}^{2}(m\quad n)}{mn(m\quad 1)}(90mK^4L^2\ {}_{l}^{2}+3mK^2)\quad \frac{p}{8}\frac{{}_{l}K}{{}_{2}{}_{l}^{2}K^2G^2+}$. Hence we have the following condition on local learning rate ${}_{l}$,

$$
{}_{l}\quad \frac{n(m\quad 1)}{48m(n\quad 1)}\ K\ \frac{p}{{}_{2}K^2G^2+}\quad \frac{3\ L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ L + \frac{p}{2(1\quad{}_{2})}G\quad {}^{1}\ ; \qquad \text{(A.30)}
$$

then we have

$$
\begin{aligned}
&\frac{p}{8}\frac{{}_{l}K}{{}_{2}{}_{l}^{2}K^2G^2+\quad T}\sum_{i=1}^{T} E[kr\ f(x_t)k^2] \\
&\quad \frac{f(z_0)\quad E[f(z_T)]}{T} + \frac{C_1\ {}_{l}KG^2d}{Tp-} + \frac{2C_1^2\ {}_{1}^{2}\ {}_{l}^{2}K^2LG^2d}{T} \\
&\quad + \frac{5\ {}_{l}^{3}K^2L^2}{p\overline{2}}({}_{l}^{2}+6K\ {}_{g}^{2}) + \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{K\ {}_{l}^{2}}{n}\ {}_{l}^{2} \\
&\quad + \frac{3\ ^2L}{2} + \frac{{}_{1}^{2}}{2(1\quad{}_{1})^2}\ ^2L + \frac{p}{2(1\quad{}_{2})}G\ \frac{{}_{l}^{2}(m\quad n)}{mn(m\quad 1)}[15mK^3L^2\ {}_{l}^{2}({}_{l}^{2}+6K\ {}_{g}^{2})+3mK^2\ {}_{g}^{2}]: \qquad \text{(A.31)}
\end{aligned}
$$

Therefore,

$$\min E[\|\nabla f(x_t)\|^2] \quad 8\sqrt{\frac{2\sigma^2 K^2 G^2 + \quad}{\quad}}\left(\frac{f_0 \quad f}{\quad KT} + \frac{}{T} + \quad\right); \tag{A.32}$$

where $= \frac{C_1 G^2 d}{\sqrt{\quad}} + \frac{2C_1^2 \quad KLG^2 d}{\quad}$; $= \frac{5\sqrt{2}KL^2}{2}\left(\frac{2}{} + 6K\frac{2}{g}\right) + [(3 + C_1^2)L + 2\sqrt{2(1 \quad 2)}G]\frac{}{2n}\frac{2}{} + [(3 + C_1^2)L + 2\sqrt{2(1 \quad 2)}G]\frac{(m \quad n)}{2n(m \quad 1)}[15K^2L^2\frac{2}{}(\frac{2}{} + 6K\frac{2}{g}) + 3K\frac{2}{g}]$ and $C_1 = \frac{1}{1 \quad 1}$.

A.4. Proof of Corollary 4.11

If we choose $= (\frac{1}{\sqrt{TK}})$ and $= (\sqrt{Kn})$, we have $\min_{t\in[T]} E[\|\nabla f(x_t)\|^2] = O(\frac{\sqrt{K}}{\sqrt{Tn}})$.

B. Proof of Theorems in Section 4.2 and Partial Participation Setting for FedCAMS

B.1. Proof of Theorem 4.17

Notations and equations.From the update rule of Algorithm 2, we have $e_1 = 0$, $e_t = \frac{1}{m}\sum_{i=1}^{m} e_t^i$ and $m_t = (1 \quad 1)\sum_{i=1}^{t} {}_1^{t\ i} b_i$. Denote a global uncompressed difference $= \frac{1}{m}\sum_{i=1}^{m} {}_t^i$. Denote a virtual momentum sequence: $m_t^0 = {}_1 m_{t\ 1}^0 + (1 \quad 1)\ {}_t$, hence we have $m_t^0 = (1 \quad 1)\sum_{i=1}^{t} {}_1^{t\ i}\ {}_i$. By the aforementioned de nition and notation, we have

$$b_t \quad {}_t = \frac{1}{m}\sum_{i=1}^{m}(b_t^i \quad {}_t^i) = \frac{1}{m}\sum_{i=1}^{m}(e_t^i \quad e_{t+1}^i) = e_t \quad e_{t+1}: \tag{B.1}$$

Denote the weighted averaging error sequence $= (1 \quad 1)\sum_{=1}^{t} {}_1^{t}\ e$, with the impute $_1 = 0$, we obtain the relation between ${}_t$ and m_t as follows

$$m_t \quad m_t^0 = (1 \quad 1)\sum_{=1}^{t} {}_1^{t}\ (b \quad) = (1 \quad 1)\sum_{=1}^{t} {}_1^{t}\ (e \quad e_{+1}) = {}_t \quad {}_{t+1}; \tag{B.2}$$

where the last step holds due to $_{t+1} = (1 \quad 1)\sum_{=1}^{t+1} {}_1^{t}\ e_{+1} = (1 \quad 1)\sum_{=1}^{t} {}_1^{t}\ e_{+1} + {}_1^{t} e_1$.

Similar to previous works studied adaptive methods (Chen et al., 2020a; Zhou et al., 2018; Chen et al., 2018), we introduce a Lyapunov sequence z_t: assume $x_0 = x_1$, for each $t \quad 1$, we have

$$y_t = x_t + \frac{1}{1 \quad 1}(x_t \quad x_{t\ 1}) = \frac{1}{1 \quad 1}x_t \quad \frac{1}{1 \quad 1}x_{t+1}:$$

Therefore, by the update rule of x_t, we have

$$\begin{aligned} y_{t+1} &= x_{t+1} + \frac{1}{1 \quad 1}\sqrt{\quad}_t^{1=2} m_t \\ &= x_{t+1} + \frac{1}{1 \quad 1}\sqrt{\quad}_t^{1=2}[m_t^0 + \quad {}_t \quad {}_{t+1}] \\ &= x_{t+1} + \frac{1}{1 \quad 1}\sqrt{\quad}_t^{1=2} m_t^0 + \frac{1}{1 \quad 1}\sqrt{\quad}_t^{1=2}\ \frac{{}_{t+1} \quad (1 \quad 1)e_{t+1}}{1} \quad {}_{t+1} \\ &= x_{t+1} + \frac{1}{1 \quad 1}\sqrt{\quad}_t^{1=2} m_t^0 + \sqrt{\quad}_t^{1=2}\ {}_{t+1} \quad \sqrt{\quad}_t^{1=2} e_{t+1}: \end{aligned} \tag{B.3}$$

The third equation holds due to the fact that $_{t+1} = {}_1\ {}_t + (1 \quad 1)e_{t+1}$. We then introduce a new sequence based on the previous Lyapunov sequence y_t as follows

$$z_{t+1} = y_{t+1} + \sqrt{\quad}_t^{1=2} e_{t+1} = x_{t+1} + \frac{1}{1 \quad 1}\sqrt{\quad}_t^{1=2} m_t^0 + \sqrt{\quad}_t^{1=2}\ {}_{t+1}: \tag{B.4}$$

The sequence difference $z_{t+1} - z_t$ can be represented by

$$\begin{aligned} z_{t+1} - z_t &= x_{t+1} - x_t + \frac{1}{1-\beta_1}\hat{V}_t^{-1/2}m_t' - \frac{1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1}' + \hat{V}_t^{-1/2}\Delta_{t+1} - \hat{V}_{t-1}^{-1/2}\Delta_t \\ &= \hat{V}_t^{-1/2}m_t + \hat{V}_t^{-1/2}\Delta_{t+1} + \frac{1}{1-\beta_1}\hat{V}_t^{-1/2}m_t' - \frac{1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1}' - \hat{V}_{t-1}^{-1/2}\Delta_t; \end{aligned} \tag{B.5}$$

where the second equation follows the update rule of x_{t+1}. Following (B.2), then combining likely terms and applying the de nition of m_t', we have

$$\begin{aligned} z_{t+1} - z_t &= \hat{V}_t^{-1/2}m_t' + \hat{V}_t^{-1/2}\Delta_t + \frac{1}{1-\beta_1}\hat{V}_t^{-1/2}m_t' - \frac{1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1}' - \hat{V}_{t-1}^{-1/2}\Delta_t \\ &= \frac{1}{1-\beta_1}\hat{V}_t^{-1/2}m_t' - \frac{1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t \\ &= \frac{1}{1-\beta_1}\hat{V}_t^{-1/2}[\beta_1 m_{t-1}' + (1-\beta_1)\Delta_t] - \frac{1}{1-\beta_1}\hat{V}_{t-1}^{-1/2}m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t \\ &= \hat{V}_t^{-1/2}\Delta_t - \frac{1}{1-\beta_1}(\hat{V}_{t-1}^{-1/2} - \hat{V}_t^{-1/2})m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t. \end{aligned} \tag{B.6}$$

Therefore, we obtain a helpful Lyapunov sequence for our proof of FedCAMS. The proof of FedCAMS in full participation settings has a similar outline with the proof of FedAMS. By Assumption 4.1, we have

$$\begin{aligned} &\mathbb{E}[f(z_{t+1})] - f(z_t) \\ &\le \mathbb{E}[\langle \nabla f(z_t), z_{t+1} - z_t\rangle] + \frac{L}{2}\mathbb{E}[\|z_{t+1} - z_t\|^2] \\ &= \mathbb{E}\left\langle \nabla f(z_t), \hat{V}_t^{-1/2}\Delta_t\right\rangle \\ &\quad + \mathbb{E}\left\langle \nabla f(z_t), -\frac{1}{1-\beta_1}(\hat{V}_{t-1}^{-1/2} - \hat{V}_t^{-1/2})m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t\right\rangle \\ &\quad + \frac{L}{2}\mathbb{E}\left\|\hat{V}_t^{-1/2}\Delta_t - \frac{1}{1-\beta_1}(\hat{V}_{t-1}^{-1/2} - \hat{V}_t^{-1/2})m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t\right\|^2 \\ &= \underbrace{\mathbb{E}\left\langle \nabla f(x_t), \hat{V}_t^{-1/2}\Delta_t\right\rangle}_{T_1} \underbrace{- \mathbb{E}\left\langle \nabla f(z_t), \frac{1}{1-\beta_1}(\hat{V}_{t-1}^{-1/2} - \hat{V}_t^{-1/2})m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t\right\rangle}_{T_2} \\ &\quad + \underbrace{\frac{L}{2}\mathbb{E}\left\|\hat{V}_t^{-1/2}\Delta_t - \frac{1}{1-\beta_1}(\hat{V}_{t-1}^{-1/2} - \hat{V}_t^{-1/2})m_{t-1}' + \hat{V}_t^{-1/2}\Delta_t - \hat{V}_{t-1}^{-1/2}\Delta_t\right\|^2}_{T_3} \\ &\quad + \underbrace{\mathbb{E}\left\langle \nabla f(z_t) - \nabla f(x_t), \hat{V}_t^{-1/2}\Delta_t\right\rangle}_{T_4}; \end{aligned} \tag{B.7}$$

here we recall the notation $\hat{V}_t = \mathrm{diag}(\hat{v}_t) = \mathrm{diag}(\max(\hat{v}_{t-1}, v_t, \epsilon))$.

Bounding T_1: We have

$$\begin{aligned} T_1 &= \mathbb{E}\left\langle \nabla f(x_t), \frac{\sqrt{1-\beta_2}\Delta_t}{\sqrt{\hat{v}_t}}\right\rangle \\ &\le \mathbb{E}\left\langle \nabla f(x_t), \frac{\sqrt{1-\beta_2}\Delta_t}{\sqrt{v_t + \epsilon}}\right\rangle \\ &= \frac{\sqrt{1-\beta_2}}{2}\mathbb{E}\left\langle \nabla f(x_t), \frac{\Delta_t}{\sqrt{\epsilon v_{t-1} + \epsilon}}\right\rangle + \frac{\sqrt{1-\beta_2}}{2}\mathbb{E}\left\langle \nabla f(x_t), \frac{\Delta_t}{\sqrt{v_t + \epsilon}} - \frac{\Delta_t}{\sqrt{\epsilon v_{t-1} + \epsilon}}\right\rangle; \end{aligned} \tag{B.8}$$

where the rst inequality follows by the fact that $\hat{v}_t \geq \frac{v_t + \epsilon}{2}$. For the second term in (B.8), we have

$$\begin{aligned}
&\sqrt{2}\, E\, \langle \nabla f(x_t), \frac{\eta}{\sqrt{v_t + \epsilon}} - \frac{\eta}{\sqrt{2v_{t-1} + \epsilon}} \rangle \\
&\leq \sqrt{2}\, \|\nabla f(x_t)\| E\, \left\| \frac{\eta}{\sqrt{v_t + \epsilon}} - \frac{\eta}{\sqrt{2v_{t-1} + \epsilon}} \right\| \|\delta_t\| \\
&\leq \frac{\sqrt{2(1-\theta_2)}G}{\sqrt{\epsilon}} E[\|\delta_t\|^2];
\end{aligned} \tag{B.9}$$

where the second inequality follows from Lemma C.1 and C.4, and we will further apply the bound $E[\|\delta_t\|^2]$ by applying Lemma C.5. For the rst term in (B.8), we have

$$\begin{aligned}
&\sqrt{2}\, E\, \langle \nabla f(x_t), \frac{\eta}{\sqrt{2v_{t-1} + \epsilon}} \rangle \\
&= \sqrt{2}\, E\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \delta_t + \frac{1}{K} \sum_{k=0}^{K-1} \nabla f(x_t) - \frac{1}{K} \sum_{k=0}^{K-1} \nabla f(x_t) \rangle \\
&= \sqrt{2}\, \frac{1}{K} E\, \left\| \frac{\nabla f(x_t)}{\sqrt[4]{2v_{t-1} + \epsilon}} \right\|^2 + \sqrt{2}\, E\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \delta_t + \frac{1}{K} \sum_{k=0}^{K-1} \nabla f(x_t) \rangle \\
&= \sqrt{2}\, \frac{1}{K} E\, \left\| \frac{\nabla f(x_t)}{\sqrt[4]{2v_{t-1} + \epsilon}} \right\|^2 + \sqrt{2}\, E\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \frac{1}{m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i + \frac{1}{K} \sum_{k=0}^{K-1} \nabla f(x_t) \rangle \\
&= \sqrt{2}\, \frac{1}{K} E\, \left\| \frac{\nabla f(x_t)}{\sqrt[4]{2v_{t-1} + \epsilon}} \right\|^2 + \sqrt{2}\, E\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, \frac{1}{m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i + \frac{1}{K} \sum_{i=1}^{m} \nabla F_i(x_t) \rangle \quad : \quad \text{(B.10)}
\end{aligned}$$

For the last term in (B.10), we have

$$\begin{aligned}
&\sqrt{2}\, \langle \frac{\nabla f(x_t)}{\sqrt{2v_{t-1} + \epsilon}}, E\, \frac{1}{m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} g_{t,k}^i + \frac{1}{K} \sum_{i=1}^{m} \nabla F_i(x_t) \rangle \\
&= \sqrt{2}\, \langle \frac{\sqrt{K}}{\sqrt[4]{2v_{t-1} + \epsilon}} \nabla f(x_t), \frac{\sqrt{K}}{Km} \frac{1}{\sqrt[4]{2v_{t-1} + \epsilon}} E \sum_{i=1}^{m} \sum_{k=0}^{K-1} (\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)) \rangle \\
&= \frac{\sqrt{2}K}{2} \left\| \frac{\nabla f(x_t)}{\sqrt[4]{2v_{t-1} + \epsilon}} \right\|^2 + \frac{\sqrt{2}}{2Km^2} E \left\| \frac{1}{\sqrt[4]{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} (\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)) \right\|^2 \\
&\quad \frac{\sqrt{2}}{2Km^2} E \left\| \frac{1}{\sqrt[4]{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2 \\
&\quad \frac{\sqrt{2}K}{2} \left\| \frac{\nabla f(x_t)}{\sqrt[4]{2v_{t-1} + \epsilon}} \right\|^2 + \frac{\sqrt{2}}{2m} \sum_{i=1}^{m} \sum_{k=0}^{K-1} E \left\| \frac{\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)}{\sqrt[4]{2v_{t-1} + \epsilon}} \right\|^2 \\
&\quad \frac{\sqrt{2}}{2Km^2} E \left\| \frac{1}{\sqrt[4]{2v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2;
\end{aligned} \tag{B.11}$$

where the second equation follows from $\langle x, y \rangle = \frac{1}{2}[\|x\|^2 + \|y\|^2 - \|x - y\|^2]$, and the inequality holds by applying

Cauchy-Schwarz inequality. Then by Assumption 4.1, we have

$$\begin{aligned}
&\frac{\sqrt{2}}{\sqrt{\hat{v}_{t-1}+\epsilon}}\nabla f(x_t);\ \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1} g_{t,k}^{i} + \frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_t)\\
&\frac{\sqrt{2}}{2}\left\|\frac{\nabla f(x_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}\right\|^2 + \frac{\sqrt{2}L^2}{2m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\mathbb{E}\left\|\frac{x_{t,k}^{i}-x_t}{\sqrt{\hat{v}_{t-1}+\epsilon}}\right\|^2\\
&\frac{\sqrt{2}}{2Km^2}\mathbb{E}\left\|\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^{i})\right\|^2\\
&\frac{3\sqrt{2}}{4}\left\|\frac{\nabla f(x_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}\right\|^2 + \frac{5\sqrt{3}K^2L^2}{\sqrt{2}}(\sigma_l^2+6K\sigma_g^2)\\
&\frac{\sqrt{2}}{2Km^2}\mathbb{E}\left\|\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^{i})\right\|^2;
\end{aligned}\tag{B.12}$$

where the last inequality holds by applying Lemma C.9 and the constraint of local learning rate $\eta \le \frac{1}{8KL}$. Then we have

$$\begin{aligned}
T_1 \le\ &\frac{\sqrt{2}}{4}\left\|\frac{\nabla f(x_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}\right\|^2 + \frac{5\sqrt{3}K^2L^2}{\sqrt{2}}(\sigma_l^2+6K\sigma_g^2)\\
&\frac{\sqrt{2}}{2Km^2}\mathbb{E}\left\|\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^{i})\right\|^2 + \frac{\sqrt{2(1-\beta_2)}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2]\\
&\frac{\sigma_l^2}{4}\mathbb{E}\left\|\frac{\nabla f(x_t)}{\sqrt{\hat{v}_{t-1}+\epsilon}}\right\|^2 + \frac{5\sqrt{3}K^2L^2}{\sqrt{2}}(\sigma_l^2+6K\sigma_g^2)\\
&\frac{\sigma_l^2}{2Km^2}\mathbb{E}\left\|\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^{i})\right\|^2 + \frac{\sqrt{2(1-\beta_2)}G}{\epsilon}\mathbb{E}[\|\Delta_t\|^2];
\end{aligned}\tag{B.13}$$

Bounding T_2: The bound for T_2 mainly follows by the update rule and de nition of virtual sequence z_t.

$$\begin{aligned}
T_2 &= \mathbb{E}\left\langle \nabla f(z_t), \frac{1}{1-\beta_1}\hat{V}_t^{1/2}\hat{V}_t^{1/2}m_{t-1}' + \hat{V}_t^{1/2}\hat{V}_t^{1/2}\Delta_t\right\rangle\\
&= \mathbb{E}\left\langle \nabla f(z_t) + \nabla f(z_t) - \nabla f(x_t), \hat{V}_t^{1/2}\hat{V}_t^{1/2}\left(\frac{1}{1-\beta_1}m_{t-1}'\right) + \Delta_t\right\rangle\\
&\le \mathbb{E}\left\|\nabla f(x_t)\right\|\left\|\hat{V}_t^{1/2}\hat{V}_t^{1/2}\left(\frac{1}{1-\beta_1}m_{t-1}'\right) + \Delta_t\right\|\\
&\quad + \sqrt{2}L\mathbb{E}\left\|\frac{1}{\sqrt{\hat{v}_{t-1}+\epsilon}}\left(\frac{1}{1-\beta_1}m_{t-1}'\right) + \Delta_t\right\|\left\|\hat{V}_t^{1/2}\hat{V}_t^{1/2}\left(\frac{1}{1-\beta_1}m_{t-1}'\right) + \Delta_t\right\|\\
&\le C_1\sigma_l KG^2\mathbb{E}\left\|\hat{V}_t^{1/2}\hat{V}_t^{1/2}\right\|_1 + \sqrt{2}C_1^2L^2\sigma_l^2K^2G^2\left\|\mathbb{E}\left\|\hat{V}_t^{1/2}\hat{V}_t^{1/2}\right\|_1\right\|;
\end{aligned}\tag{B.14}$$

where the last inequality holds by Lemma C.4, here $C_1 = \frac{1}{1-\beta_1} + \frac{2q}{1-q^2}$.

Bounding T_3: It can be bounded as follows:

$$\begin{aligned}
T_3 &= \frac{\sqrt{2}L}{2}\mathbb{E}\left\|\hat{V}_t^{1/2}\Delta_t + \frac{1}{1-\beta_1}\hat{V}_t^{1/2}\hat{V}_t^{1/2}m_{t-1}'\right\|^2\\
&\le \sqrt{2}L\mathbb{E}\left\|\hat{V}_t^{1/2}\Delta_t\right\|^2 + \sqrt{2}L\mathbb{E}\left\|\frac{1}{1-\beta_1}\hat{V}_t^{1/2}\hat{V}_t^{1/2}m_{t-1}' + \hat{V}_t^{1/2}\hat{V}_t^{1/2}\Delta_t\right\|^2\\
&\le \sqrt{2}L\mathbb{E}\left\|\hat{V}_t^{1/2}\Delta_t\right\|^2 + \sqrt{2}LC_1^2\sigma_l^2K^2G^2\mathbb{E}\left\|\hat{V}_t^{1/2}\hat{V}_t^{1/2}\right\|^2;
\end{aligned}\tag{B.15}$$

where the rst inequality follows by Cauchy-Schwarz inequality, and the second one follows by Lemma C.4, where $\frac{1}{1-\beta_1} + \frac{2q}{1-q^2}$.

Bounding T_4:

$$\begin{aligned} T_4 &= \mathbb{E}\left\langle \nabla f(z_t) - \nabla f(x_t),\ \hat{\eta}_t^{1/2} \delta_t \right\rangle \\ &\le \mathbb{E}\left[\|\nabla f(z_t) - \nabla f(x_t)\| \cdot \|\hat{\eta}_t^{1/2} \delta_t\| \right] \\ &\le L\, \mathbb{E}\left[\|z_t - x_t\| \cdot \|\hat{\eta}_t^{1/2} \delta_t\| \right] \\ &\le \frac{\beta_1^2 L}{2} \mathbb{E}\left\| \hat{\eta}_t^{1/2} \delta_{t-1} \right\|^2 + \frac{\beta_1^2 L}{2} \mathbb{E}\left\| \frac{1}{1-\beta_1} \hat{\eta}_t^{1/2} m_t + \hat{\eta}_t^{1/2} \delta_t \right\|^2, \end{aligned}$$

where the rst inequality holds by the fact that $\langle a, b\rangle \le \|a\| \cdot \|b\| \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$, the second one follows from Assumption 4.1 and the third one holds by the de nition of virtual sequence z_t and the fact of $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$. Then summing T_4 over $t = 1, \ldots, T$, we have

$$\begin{aligned} \sum_{t=1}^{T} T_4 &\le \frac{\beta_1^2 L}{2} \sum_{t=1}^{T} \mathbb{E}\left\| \hat{\eta}_t^{1/2} \delta_{t-1} \right\|^2 + \frac{\beta_1^2 L}{2} \sum_{t=1}^{T} \mathbb{E}\left\| \frac{1}{1-\beta_1} m_t + \delta_t \right\|^2 \\ &\le \frac{\beta_1^2 L}{2} \sum_{t=1}^{T} \mathbb{E}[\|\delta_{t-1}\|^2] + \frac{\beta_1^2 L}{(1-\beta_1)^2} \sum_{t=1}^{T} \mathbb{E}\|m_t\|^2 + \sum_{t=1}^{T} \mathbb{E}\|\delta_t\|^2 . \end{aligned} \tag{B.16}$$

By Lemma C.7, we have

$$\sum_{t=1}^{T} \mathbb{E}[\|m_t\|^2] \le \frac{TK\sigma_l^2}{m} + \frac{\sigma_l^2}{m^2} \sum_{t=1}^{T} \mathbb{E}\left\| \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2 ;$$

and

$$\sum_{t=1}^{T} \mathbb{E}[\|\delta_t\|^2] \le \frac{4Tq^2}{(1-q^2)^2} \frac{K\sigma_l^2}{m} + \frac{\sigma_l^2}{m^2} \frac{4q^2}{(1-q^2)^2} \sum_{t=1}^{T} \mathbb{E}\left\| \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2 .$$

Therefore, the T_4 term is bounded by

$$\sum_{t=1}^{T} T_4 \le \frac{\beta_1^2 L}{2} \sum_{t=1}^{T} \mathbb{E}[\|\delta_{t-1}\|^2] + \frac{C_2 \beta_1^2 L \sigma_l^2}{m^2} \sum_{t=1}^{T} \mathbb{E}\left\| \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2 + \frac{C_2 \beta_1^2 L}{m} TK\sigma_l^2, \tag{B.17}$$

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

Merging pieces together: Substituting (B.13), (B.14) and (B.15) into (B.7), summing over from $t = 1$ to T and then adding

(B.17), we have

$$
\begin{aligned}
E[f(z_{T+1})] - f(z_1) = & \sum_{t=1}^{T} [T_1 + T_2 + T_3 + T_4] \\
& \frac{\alpha K}{4} \sum_{t=1}^{T} E\left[\left\| \frac{\nabla f(x_t)}{\sqrt{2 v_{t-1} + \epsilon}} \right\|^2 \right] + \frac{5 \alpha^3 K^2 L^2 T}{\sqrt{2}} \left(\sigma_l^2 + 6 K \sigma_g^2 \right) + \frac{\sqrt{2(1-\beta_2)} G}{1-\beta_1} \sum_{t=1}^{T} E[\|\delta_t\|^2] \\
& \frac{\alpha}{2Km^2} \sum_{t=1}^{T} E\left[\left\| \frac{1}{\sqrt{2 v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t;k}^i) \right\|^2 \right] + C_1 \alpha K G^2 \sum_{t=1}^{T} E\left[\left\| \beta_1^{t-1/2} - \beta_1^{t-1/2} \right\|_1 \right] \\
& + \frac{C_1^2 \beta_1^2 \alpha^2 K^2 G^2}{1-\beta_1} \sum_{t=1}^{T} E\left[\left\| \beta_1^{t-1/2} - \beta_1^{t-1/2} \right\|_1^2 \right] + C_1^2 \beta_1^2 \alpha^2 K^2 L G^2 \sum_{t=1}^{T} E\left[\left\| \beta_1^{t-1/2} - \beta_1^{t-1/2} \right\|^2 \right] \\
& + \alpha^2 L \sum_{t=1}^{T} E\left[\left\| \beta_1^{t-1/2} \right\|^2 \right] + \frac{\alpha^2 L}{2} \sum_{t=1}^{T} E\left[\left\| \beta_1^{t-1/2} \right\|^2 \right] + \frac{\alpha^2 L}{2} \frac{\beta_1^2}{(1-\beta_1)^2} \sum_{t=1}^{T} E[\|km_t^0\|^2] + \frac{\alpha^2 L}{2} \sum_{t=1}^{T} E[\|\delta_t\|^2].
\end{aligned}
\tag{B.18}
$$

Hence by organizing and applying Lemmas, we have

$$
\begin{aligned}
& E[f(z_{T+1})] - f(z_1) \\
& \frac{\alpha K}{4} \sum_{t=1}^{T} E\left[\left\| \frac{\nabla f(x_t)}{\sqrt{2 v_{t-1} + \epsilon}} \right\|^2 \right] + \frac{5 \alpha^3 K^2 L^2 T}{\sqrt{2}} \left(\sigma_l^2 + 6 K \sigma_g^2 \right) \\
& \frac{\alpha}{2Km^2} \sum_{t=1}^{T} E\left[\left\| \frac{1}{\sqrt{2 v_{t-1} + \epsilon}} \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t;k}^i) \right\|^2 \right] + \frac{C_1 \alpha K G^2 d}{1-\beta_1} + \frac{2 C_1^2 \beta_1^2 \alpha^2 K^2 L G^2 d}{\epsilon} \\
& + \alpha^2 L + \frac{\alpha^2 L}{2} + \frac{\sqrt{2(1-\beta_2)} G}{1-\beta_1} \left(\frac{KT \sigma_l^2}{m} + \sigma_l^2 + \frac{\sigma_l^2}{m^2} \right) \sum_{t=1}^{T} E\left[\left\| \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t;k}^i) \right\|^2 \right] \\
& + \frac{\alpha^2 L}{\epsilon} \frac{\sigma_l^2 C_2}{m^2} \sum_{t=1}^{T} E\left[\left\| \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t;k}^i) \right\|^2 \right] + \frac{\alpha^2 L}{\epsilon} \frac{TK \sigma_l^2 C_2}{m} \sigma_l^2;
\end{aligned}
\tag{B.19}
$$

by applying Lemma C.5 into all terms containing the second moment estimate of model difference in (B.18), using the fact that $\left\| \frac{1}{\sqrt{\frac{(1+\beta_2^2)^3}{(1-\beta_2^2)^2} K^2 G^2 + \epsilon}} \right\|_1 \le \frac{1}{\epsilon}$ and $\left\| \frac{1}{\sqrt{\frac{(1+\beta_2^2)^3}{(1-\beta_2^2)^2} \sigma_l^2 K^2 G^2 + \epsilon}} \right\|_1 \le \frac{1}{\epsilon}$ and $\left\| \frac{x}{\sqrt{2 v_t + \epsilon}} \right\|^{1/2} \le \|x\|$, and applying Lemma C.2 and C.8, we have

$$
\begin{aligned}
& E[f(z_{T+1})] - f(z_1) \\
& \frac{\alpha K}{4 \sqrt{4 \frac{(1+\beta_2^2)^3}{(1-\beta_2^2)^2} \sigma_l^2 K^2 G^2 + \epsilon}} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|^2] + \frac{5 \alpha^3 K^2 L^2 T}{\sqrt{2}} \left(\sigma_l^2 + 6 K \sigma_g^2 \right) \\
& + \frac{C_1 \alpha K G^2 d}{1-\beta_1} + \frac{2 C_1^2 \beta_1^2 \alpha^2 K^2 L G^2 d}{\epsilon} + \left(\frac{3 \alpha^2 L}{2} + C_2 \alpha^2 L + \frac{\sqrt{2(1-\beta_2)} G}{1-\beta_1} \right) \frac{KT \sigma_l^2}{m} \sigma_l^2 \\
& \sum_{t=1}^{T} E\left[\left\| \sum_{i=1}^{m} \sum_{k=0}^{K-1} \nabla F_i(x_{t;k}^i) \right\|^2 \right] \frac{\alpha}{2 \sqrt{4 \frac{(1+\beta_2^2)^3}{(1-\beta_2^2)^2} \sigma_l^2 K^2 G^2 + \epsilon}} \left(\frac{3 \alpha^2 L}{2} + C_2 \alpha^2 L + \frac{\sqrt{2(1-\beta_2)} G}{1-\beta_1} \right) \frac{\sigma_l^2}{m^2} \\
& \frac{\alpha K}{4 C_0} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|^2] + \frac{5 \alpha^3 K^2 L^2 T}{\sqrt{2}} \left(\sigma_l^2 + 6 K \sigma_g^2 \right) \\
& + \frac{C_1 \alpha K G^2 d}{1-\beta_1} + \frac{2 C_1^2 \beta_1^2 \alpha^2 K^2 L G^2 d}{\epsilon} + \left(\frac{3 \alpha^2 L}{2} + C_2 \alpha^2 L + \frac{\sqrt{2(1-\beta_2)} G}{1-\beta_1} \right) \frac{KT \sigma_l^2}{m} \sigma_l^2;
\end{aligned}
\tag{B.20}
$$

where the last inequality holds by $\sqrt{4\ _2(1+q^2)^3(1\ q^2)^{\ 2}K^{\ 2}G^2+\ } \le K(3\ L\ +2\ C_2\ L\ +2\sqrt{2(1\ \ _2)}G)$. Hence we have

$$\frac{q}{4}\sqrt{\frac{{}_lK}{4\ _2\frac{(1+q^2)^3}{(1\ q^2)^2}\ {}_l^2K^{\ 2}G^2+\ }}\frac{X^T}{T}\ _{t=1}E[kr\ f(x_t)k^2]$$
$$\frac{f(z_0)\ \ E[f(z_T)]}{T}+\frac{5}{\sqrt{2}}\frac{{}_l^3K^{\ 2}L^2}{}(\ {}_l^{\ 2}+6K\ \ {}_g^2)+\frac{C_1\ {}_lKG^{\ 2}d}{T\sqrt{\ }}+\frac{2C_1^2\ \ ^2\ {}_l^2K^{\ 2}LG^{\ 2}d}{T}$$
$$+\ \ 3\ ^2L+2C_2\ \ ^2L+2\sqrt{2(1\ \ _2)}\ G\ \ \frac{K\ {}_l^2}{2m}\ {}_l^2; \tag{B.21}$$

where$C_1=\frac{1}{1\ \ _1}+\frac{2q}{1\ q^2}$ and$C_2=\frac{{}_1^2}{(1\ \ _1)^2}+\frac{4q^2}{(1\ q^2)^2}$. (B.21) also implies,

$$\min E[kr\ f(x_t)k^2]\quad 4\ \ 4\ _2\frac{(1+q^2)^3}{(1\ q^2)^2}\ {}_l^2K^{\ 2}G^2+\ \ \frac{f_0\ \ f}{{}_lKT}+\frac{}{T}+\ \ ; \tag{B.22}$$

where $=\frac{C_1G^2d}{\sqrt{\ }}+\frac{2C_1^2\ {}_lKLG^{\ 2}d}{}$; $=\frac{5}{\sqrt{2}}\frac{{}_l^2KL^{\ 2}}{}(\ {}_l^{\ 2}+6K\ \ {}_g^2)+[(3+2\ C_2)\ L\ +2\sqrt{2(1\ \ _2)}G]\frac{{}_l}{2m}\ {}_l^{\ 2}$, $C_1=\frac{1}{1\ \ _1}+\frac{2q}{1\ q^2}$ and$C_2=\frac{{}_1^2}{(1\ \ _1)^2}+\frac{4q^2}{(1\ q^2)^2}$.

B.2. Proof of Corollary 4.19

Let ${}_l=(\ \frac{1}{\sqrt{TK}})$ and $=(\ \sqrt{Km})$, the convergence rate under full participation scheme can be$O(\frac{1}{\sqrt{TKm}})$.

B.3. Analysis on the Partial Participation Setting for FedCAMS

Let us present the theoretical analysis of the partial participation scheme of FedCAMS (Algorithm 2). Similar to partial participation scheme in Section 4.1, we have the following convergence analysis.

Theorem B.1. Under Assumption 4.1-4.3 and 4.14, if the local learning rate${}_l$ satis es the following condition: ${}_l\ \min\ \frac{1}{8KL};\frac{n(m\ \ 1)}{48m(n\ \ 1)}[K\sqrt{4\ _2(1+q^2)^3(1\ \ q^2)^{\ 2}K^{\ 2}G^2+\ }(\ L\ +\sqrt{2(1\ \ _2)}G)]^{\ 1}$, then the iterates of Algorithm 2 under partial participation scheme satisfy

$$\min_{t2[T]}Ekr\ f(x_t)k^2\quad 8\ \ 4\ _2\frac{(1+q^2)^3}{(1\ q^2)^2}\ {}_l^2K^{\ 2}G^2+\ \ \frac{f_0\ \ f}{{}_lKT}+\frac{}{T}+\ \ ; \tag{B.23}$$

where $=\frac{C_1G^2d}{\sqrt{\ }}+\frac{2C_1^2\ {}_lKLG^{\ 2}d}{}$, $=\frac{C_1\ {}_lKLG^{\ 2}}{}+\frac{5}{\sqrt{2}}\frac{{}_l^2KL^{\ 2}}{}(\ {}_l^{\ 2}+6K\ \ {}_g^2)+[\ L\ +\sqrt{2(1\ \ _2)}G]\frac{{}_l}{n}\ {}_l^{\ 2}+[\ L\ +\sqrt{2(1\ \ _2)}G]\frac{{}_l(m\ \ n)}{n(m\ \ 1)}[15K^{\ 2}L^{\ 2}\ {}_l^2(\ {}_l^{\ 2}+6K\ \ {}_g^2)+3K\ \ {}_g^2]$ and$C_1=\frac{1}{1\ \ _1}+\frac{2q}{1\ q^2}$.

Remark B.2. The upper bound for$\min_{t2[T]}Ekr\ f(x_t)k^2$ of partial participation is similar to full participation case but with a larger variance term. This is due to the fact that random sampling of participating workers introduces an additional variance during sampling.

Remark B.3. From Theorem B.1, constantC_1 is related to the compressor constant$q$. The stronger compression we apply to the model difference ${}_t^i$ corresponding to larger$q(q\ !\ \ 1)$ leads to worse convergence due to larger information losses.

Next, we provide theoretical proofs for the partial participation analysis for FedCAMS.

Proof of Theorem B.1:

Notations and equations:From the update rule of Algorithm 2, we have$e_0=0$, $e_t=\frac{1}{m}\sum_{i=1}^m e_t^i$ and$m_t=(1\ \ _1)\sum_{i=1}^t\ {}_1^{\ t\ i}b_i$. Denote a global uncompressed difference$_t=\frac{1}{jS_tj}\sum_{i2S_t}{}_t^i$. Denote a virtual momentum sequence: $m_t^0=\ _1m_{t\ 1}^0+(1\ \ _1)\ {}_t$ hence we have$m_t^0=(1\ \ _1)\sum_{i=1}^t\ {}_1^{\ t\ i}\ {}_i$. De ne additional two virtual sequences ${}_t^0=\frac{1}{n}\sum_{i=1}^m{}_t^i$ and$b_t^0=\frac{1}{n}\sum_{i=1}^m b_t^i$. Note that when the client$i$ does not take part in the round of participation at step t, we have ${}_t^i=b_t^i=0$, therefore, ${}_t^0=\ _t$ and$b_t^0=\ b_t$.

By the aforementioned de nition and notation, de ne a subset $S_t = \{w_1^t, w_2^t, \dots, w_n^t\}$, we have

$$b_t - \beta_t = \frac{1}{|S_t|}\sum_{i\in S_t}(b_t^i - \beta_t^i) = \frac{1}{n}\sum_{i=1}^m (b_t^i - \beta_t^i) = \frac{1}{n}\sum_{i=1}^m (e_t^i - e_{t+1}^i) = e_t^0 - e_{t+1}^0; \tag{B.24}$$

where the compression errors have the same structure, $e_t^0 = \frac{1}{n}\sum_{i=1}^m e_t^i$. Similar to the previous analysis, we de ne the following sequence:

$$\Delta_{t+1} := (1-\beta_1)\sum_{\tau=1}^{t+1}\beta_1^{t+1-\tau} e^0;$$

and keep using the Lyapunov function Φ_t from (B.4). For the expectation of model difference Δ_t, we have

$$E_{S_t}[\Delta_t] = \frac{1}{n}E_{S_t}\sum_{i=1}^n \Delta_{w_i}^t = E_{S_t}[\Delta_{w_1}^t] = \frac{1}{m}\sum_{i=1}^m \Delta_i^t = \Delta_t. \tag{B.25}$$

The proof of FedCAMS in partial participation settings has a similar outline combing the proof of partial participation in FedAMS and full participation in FedCAMS. By Assumption 4.1, we have

$$\begin{aligned}
&E[f(z_{t+1})] - f(z_t)\\
&\le E\underbrace{\langle \nabla f(x_t), \hat{\eta}_t^{-1/2}\Delta_t\rangle}_{T_1^0}\\
&\quad + E\underbrace{\langle \nabla f(z_t), \tfrac{1}{1-\beta_1}\hat{\eta}_t^{-1/2}\beta_1 m_{t-1}^0 + \hat{\eta}_t^{-1/2}\Delta_t - \hat{\eta}_{t-1}^{-1/2}\Delta_t\rangle}_{T_2^0}\\
&\quad + \frac{L^2}{2}E\underbrace{\Big\|\hat{\eta}_t^{-1/2}\Delta_t - \tfrac{1}{1-\beta_1}\hat{\eta}_{t-1}^{-1/2}\beta_1 m_{t-1}^0 - \hat{\eta}_t^{-1/2}\beta_1 m_{t-1}^0 + \hat{\eta}_{t-1}^{-1/2}\beta_1 m_{t-1}^0 - \hat{\eta}_{t-1}^{-1/2}\Delta_t\Big\|^2}_{T_3^0}\\
&\quad + E\underbrace{\langle \nabla f(z_t) - \nabla f(x_t), \hat{\eta}_t^{-1/2}\Delta_t\rangle}_{T_4^0}.
\end{aligned} \tag{B.26}$$

Note that the bound for T_2^0 is exactly the same as the bound for T_2. For the three corresponding terms T_1^0, T_3^0 and T_4^0 which include the second-order momentum estimate $\hat{\eta}_t$, For T_1^0, similar to the full participation settings, we have

$$T_1^0 \le \sqrt{2}E\langle \nabla f(x_t), \sqrt{\hat{\eta}_t}\hat{\eta}_t^{-1}\Delta_t\rangle + \sqrt{2}E\langle \nabla f(x_t), \sqrt{\hat{\eta}_t}\hat{\eta}_t^{-1}\Delta_t - \sqrt{\hat{\eta}_{t-1}}\hat{\eta}_{t-1}^{-1}\Delta_t\rangle. \tag{B.27}$$

The first term in (B.27) does not change in partial participation scheme. The second term is changed due to the variance of Δ_t changes. For the second term of T_1^0, we have

$$\sqrt{2}E\langle \nabla f(x_t), \sqrt{\hat{\eta}_t}\hat{\eta}_t^{-1}\Delta_t - \sqrt{\hat{\eta}_{t-1}}\hat{\eta}_{t-1}^{-1}\Delta_t\rangle \le \frac{\sqrt{2(1-\beta_2)}G}{1-\beta}E[\|\Delta_t\|^2]. \tag{B.28}$$

For T_3^0, similar to the proof of T_3, we have

$$\sum_{t=1}^T T_3^0 \le \frac{\alpha^2 L}{1-\beta}\sum_{t=1}^T E[\|\Delta_t\|^2] + \alpha^2 L C_1^2 d K^2 G^2 \sum_{t=1}^T E\Big\|\hat{\eta}_t^{-1/2} - \hat{\eta}_{t-1}^{-1/2}\Big\|^2; \tag{B.29}$$

where $C_1 = \frac{1}{1-\beta_1} + \frac{2q}{1-q^2}$. For T_4^0 in partial participation, we have

$$\begin{aligned} T_4^0 = & \quad E \quad f(z_t) \quad f(x_t); \hat{v}_t^{1=2} \, _t \\ & E \; [f(z_t) \quad f(x_t)] \hat{v}_t^{1=2} \, _t \\ & {}^2LE \quad \frac{1}{1-\beta_1} \hat{v}_t^{1=2} m_{t-1}^0 + \hat{v}_{t-1}^{1=2} \, _t \quad \hat{v}_t^{1=2} \, _t \\ & \frac{C_1 \, ^2 \, ^2K^2LG^2}{} : \end{aligned} \tag{B.30}$$

Hence, the summation from T_1^0 to T_4^0 over total iteration T is:

$$\begin{aligned} E[f(z_{T+1})] \quad f(z_1) = & \sum_{t=1}^{T} [T_1^0 + T_2^0 + T_3^0 + T_4^0] \\ & \frac{K}{4} \sum_{t=1}^{T} E \quad \frac{r f(x_t)}{\sqrt{2v_t + }} \quad ^2 + \frac{5 \, ^3K^2L^2T}{p\overline{2}} (\, ^2 + 6K \, _g^2) + \frac{p\overline{2(1 \quad _2)} \, G}{} \sum_{t=1}^{T} E[k \, _t k^2] \\ & \frac{}{2Km^2} \sum_{t=1}^{T} E \quad \frac{1}{\sqrt{2v_t + }} \sum_{i=1}^{m} \sum_{k=0}^{K-1} r F_i(x_t)) \quad ^2 + C_1 \, KG^2 \sum_{t=1}^{T} E \quad \hat{v}_t^{1=2} \, _{t-1} \quad \hat{v}_t^{1=2} \, _1 \\ & + C_1^2 \, ^2 \, ^2K^2LG^2 \quad ^{1=2} \sum_{t=1}^{T} E \quad \hat{v}_t^{1=2} \, _{t-1} \quad \hat{v}_t^{1=2} \, _1 + C_1^2 \, ^2 \, ^2K^2LG^2 \sum_{t=1}^{T} E \quad \hat{v}_t^{1=2} \, _{t-1} \quad \hat{v}_t^{1=2} \, ^2 \\ & + \frac{^2L}{} \sum_{t=1}^{T} E[k \, _t k^2] + \frac{C_1 T \, ^2 \, ^2K^2LG^2}{} \\ & \frac{q}{4} \frac{K}{4 \, _2 \frac{(1+q^2)^3}{(1-q^2)^2} \, ^2K^2G^2 + } \sum_{t=1}^{T} E[kr f(x_t)k^2] + \frac{5 \, ^3K^2L^2T}{p\overline{2}} (\, ^2 + 6K \, _g^2) + \frac{C_1 \, KG^2d}{T} p\overline{} \\ & + \frac{2C_1^2 \, ^2 \, ^2K^2LG^2d}{T} \quad \frac{q}{2} \frac{}{4 \, _2 \frac{(1+q^2)^3}{(1-q^2)^2} \, ^2K^2G^2 + Km^2} \sum_{t=1}^{T} E \quad \sum_{i=1}^{m} \sum_{k=0}^{K-1} r F_i(x_t)) \quad ^2 \\ & + \quad \frac{^2 \, ^2LKT}{n} + \frac{p\overline{2(1 \quad _2)} \, ^2KTG}{n} \quad ^2 + \frac{C_1 T \, ^2 \, ^2K^2LG^2}{} \\ & + \quad \frac{^2 \, ^2L}{} + \frac{p\overline{2(1 \quad _2)} \, ^2G}{} \quad \frac{m \quad n}{mn(m \quad 1)} \, 15mK^3L^3 \, ^2(\, ^2 + 6K \, _g^2)T \\ & + (90mK^4L^2 \, ^2 + 3mK^2) \sum_{t=1}^{T} E[kr f(x_t)k^2] + 3mK^2T \, _g^2 \\ & + \quad \frac{^2 \, ^2L}{} + \frac{p\overline{2(1 \quad _2)} \, ^2G}{} \quad \frac{n \quad 1}{mn(m \quad 1)} \sum_{t=1}^{T} E \quad \sum_{i=1}^{m} \sum_{k=0}^{K-1} r F_i(x_t)) \quad ^2 : \end{aligned} \tag{B.31}$$

The proof outline is similar with previous proof. We take the use of Lemma C.2, C.6, C.8 for corresponding terms. By additional constraints of local learning rate with the inequality $^2L + p\overline{2(1 \quad _2)} \, G \left[\frac{^2(n \quad 1)}{mn(m \quad 1)} \quad \frac{}{2Km^2} \right] \frac{q}{4 \, _2 \frac{(1+q^2)^3}{(1-q^2)^2} \, ^2K^2G^2 + } \quad 1 \quad 0$, we obtain the constraint $\frac{n(m \quad 1)}{m(n \quad 1)} \frac{}{2K} p\overline{} \frac{}{4 \, _2(1+q^2)^3(1 \quad q^2)^2 \, ^2K^2G^2 + } [L + p\overline{2(1 \quad _2)} G]$, and we further need satis es $p\overline{} \frac{K}{4 \, _2(1+q^2)^3(1 \quad q^2)^2 \, ^2K^2G^2 + } (\, ^2L + p\overline{2(1 \quad _2)} \, G) \frac{^2(m \quad n)}{mn(m \quad 1)} (90mK^4L^2 \, ^2 + 3mK^2) \quad p\overline{} \frac{K}{8 \, _2(1+q^2)^3(1 \quad q^2)^2 \, ^2K^2G^2 + }$. Hence for the convergence

rate, we have

$$\frac{q}{8}\frac{\sqrt{K}}{4\sqrt{2\frac{(1+q^2)^3}{(1-q^2)^2}\,{}^2K^2G^2+\quad}\sqrt{T}}\sum_{i=1}^{T}E[kr\ f(x_t)k^2]$$
$$\frac{f(z_0)\quad E[f(z_T)]}{T}+\frac{5\,p\overline{3}K^2L^2}{2}(\ {}^2_{\ l}+6K\ {}^2_{g})+\quad L\ +\sqrt{2(1\quad{}_2)}G\ \frac{{}^2_{\ l}K\ {}^2_{\ l}}{n}$$
$$+\ \frac{C_1\ {}_lKG^2d}{Tp-}+\frac{2C_1^2\ {}^2_{\ l}K^2LG^2d}{T}+\frac{C_1\ {}^2\ {}^2_{\ l}K^2LG^2}{}$$
$$+\quad\frac{{}^2\ {}^2_{\ l}L}{}+\frac{\sqrt{2(1\quad{}_2)}\ {}^2_{\ l}G}{}\ \frac{m\quad n}{mn(m\quad 1)}[15mK^3L^2\ {}^2_{\ l}(\ {}^2_{\ l}+6K\ {}^2_{g})+3mK^2\ {}^2_{g}]: \tag{B.32}$$

Therefore,

$$\min E[kr\ f(x_t)k^2]\quad 8\quad 4\sqrt{2\frac{(1+q^2)^3}{(1\quad q^2)^2}\,{}^2K^2G^2+\quad}\ \frac{f_0\quad f}{{}_lKT}+\frac{}{T}+\quad; \tag{B.33}$$

where $=\frac{C_1G^2d}{p-}+\frac{2C_1^2\ {}_lKLG^2d}{}$; $=\frac{C_1\ {}_lKLG^2}{}+\frac{5\,p\overline{2}K^2L^2}{2}(\ {}^2_{\ l}+6K\ {}^2_{g})+[\ L\ +\sqrt{2(1\quad{}_2)}G]\frac{{}_l\ {}^2_{\ l}}{n}+[\ L\ +\sqrt{2(1\quad{}_2)}G]\frac{{}_l(m\quad n)}{n(m\quad 1)}[15K^2L^2\ {}^2_{\ l}(\ {}^2_{\ l}+6K\ {}^2_{g})+3K\ {}^2_{g}]$ and $C_1=\frac{1}{1\quad{}_1}+\frac{2q}{1\quad q^2}$.

C. Supporting Lemmas

Lemma C.1. For the element-wise difference $W_t=\frac{1}{p\overline{v_t+}}\quad\frac{1}{p\overline{{}_2v_{t\ 1}+}}$, we have $kW_tk\quad\frac{p\overline{1\quad{}_2}}{}k\ {}_tk$.

Proof. Note that we have:

$$\begin{aligned}
kW_tk&=\frac{1}{p\overline{v_t+}}\quad\frac{1}{p\overline{{}_2v_{t\ 1}+}}\\
&=\frac{(p\overline{{}_2v_{t\ 1}+}\quad p\overline{v_t+})(p\overline{{}_2v_{t\ 1}+}+p\overline{v_t+})}{p\overline{v_t+}p\overline{{}_2v_{t\ 1}+}(p\overline{{}_2v_{t\ 1}+}+p\overline{v_t+})}\\
&=p\overline{v_t+}p\overline{{}_2v_{t\ 1}+}(\frac{{}_2v_t}{p\overline{{}_2v_{t\ 1}+}}+p\overline{v_t+})\\
&=p\overline{v_t+}p\overline{{}_2v_{t\ 1}+}(\frac{(1\quad{}_2)\ {}^2_t}{p\overline{{}_2v_{t\ 1}+}}+p\overline{v_t+})\\
&\quad p\overline{v_t+}p\overline{{}_2v_{t\ 1}+}p\overline{1\quad{}_2\ {}_t}\\
&\quad\sqrt{1\quad{}^2_2}k\ {}_tk;
\end{aligned} \tag{C.1}$$

where the forth equation holds by the update rule of v_t, i.e., $v_t={}_2v_{t\ 1}+(1\quad{}_2)\ {}^2_t$, and the rst inequality holds due $p\overline{v_t+}\quad p\overline{v_t}\quad p\overline{1\quad{}_2\ {}_t}$ and $p\overline{{}_2v_{t\ 1}+}\quad 0$. This concludes the proof. □

Lemma C.2. For the variance difference sequence $\{\ {}^{1=2}_{\ 1}\quad\ {}^{1=2}_{\ t}\}$, we have

$$\sum_{t=1}^{T}\ {}^{1=2}_{\ 1}\quad\ {}^{1=2}_{\ t\quad 1}\quad p\overline{=};\ \sum_{t=1}^{T}\ {}^{1=2}_{\ 1}\quad\ {}^{1=2\ \ 2}_{\ t\quad 1}\quad\frac{d}{}: \tag{C.2}$$

Proof. By the de nition of variance matrix $_t$, and the non-decreasing update of FedCAMS, we $_{t\ 1}\quad\ {}_t=$

$\max(\hat{v}_{t-1}, v_t, \epsilon)$, we have

$$\begin{aligned}\sum_{t=1}^{T} \|\psi_{t,1}^{1/2} - \psi_{t-1,1}^{1/2}\|_1 &= \sum_{t=1}^{T} \sum_{i=1}^{d} \left|\frac{1}{\sqrt{\hat{v}_{t-1,i}}} - \frac{1}{\sqrt{\hat{v}_{t,i}}}\right| \\ &= \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{\sqrt{\hat{v}_{t,i}} - \sqrt{\hat{v}_{t-1,i}}}{\sqrt{\hat{v}_{t-1,i}}\sqrt{\hat{v}_{t,i}}} \\ &= \sum_{i=1}^{d} \frac{\sqrt{\hat{v}_{1,i}}}{\sqrt{\hat{v}_{0,i}}} - \frac{\sqrt{\hat{v}_{T,i}}}{\sqrt{\hat{v}_{T,i}}} \\ &\le d, \end{aligned} \tag{C.3}$$

where the inequality holds by the de nition of $\hat{v}_t \in \mathbb{R}^d$. For the sum of the variance difference under ℓ_2-norm, we have

$$\begin{aligned}\sum_{t=1}^{T} \|\psi_{t,1}^{1/2} - \psi_{t-1,1}^{1/2}\|^2 &= \sum_{t=1}^{T} \sum_{i=1}^{d} \left|\frac{1}{\sqrt{\hat{v}_{t-1,i}}} - \frac{1}{\sqrt{\hat{v}_{t,i}}}\right|^2 \\ &= \sum_{t=1}^{T} \sum_{i=1}^{d} \left|\frac{1}{\sqrt{\hat{v}_{t-1,i}}} - \frac{1}{\sqrt{\hat{v}_{t,i}}}\right|^2 \\ &\le \sum_{t=1}^{T} \sum_{i=1}^{d} \frac{1}{\hat{v}_{t-1,i}} - \frac{1}{\hat{v}_{t,i}} \\ &\le \sum_{i=1}^{d} \frac{1}{\hat{v}_{0,i}} - \frac{1}{\hat{v}_{T,i}} \\ &\le d, \end{aligned} \tag{C.4}$$

where the first inequality holds by the element-wise operation. For $x, y \in \mathbb{R}^d, 0 \le y \le x$, we have $(x - y)^2 \le (x - y)(x + y) = x^2 - y^2$. It concludes the proof. □

Lemma C.3. The compression error has the following absolute bound

$$\|e_t^i\|^2 \le \frac{4q^2}{(1-q^2)^2} \lambda^2 K^2 G^2, \quad \|e_t\|^2 \le \frac{4q^2}{(1-q^2)^2} \lambda^2 K^2 G^2. \tag{C.5}$$

Proof. For all $t \in [T]$, by Assumption 4.14 and Young's inequality, we have

$$\begin{aligned}\|e_{t+1}^i\|^2 &= \|\Delta_t^i + e_t^i - C(\Delta_t^i + e_t^i)\|^2 \\ &\le q^2 \|\Delta_t^i + e_t^i\|^2 \\ &\le q^2(1+\gamma)\|e_t^i\|^2 + q^2\left(1 + \frac{1}{\gamma}\right)\|\Delta_t^i\|^2 \\ &\le \frac{1+q^2}{2}\|e_t^i\|^2 + \frac{2q^2}{1-q^2}\|\Delta_t^i\|^2, \end{aligned}$$

where the last inequality holds by choosing $\gamma = \frac{1-q^2}{2q^2}$. Thus we obtain the absolute bound for the error terms,

$$\begin{aligned}&\|e_t^i\|^2 \le \frac{4q^2}{(1-q^2)^2} \lambda^2 K^2 G^2, \\ &\|e_t\|^2 = \left\|\frac{1}{m}\sum_{i=1}^{m} e_t^i\right\|^2 \le \frac{1}{m}\sum_{i=1}^{m} \|e_t^i\|^2 \le \frac{4q^2}{(1-q^2)^2} \lambda^2 K^2 G^2. \end{aligned} \tag{C.6}$$

It concludes the proof. □

Lemma C.4. Under Assumptions 4.2 and 4.14, for FedAMS, we have$k r f(x)k$ $G, k_t k$ $_l KG, km_t k$ $_l KG$ and $kv_t k$ $^2_l K^2G^2$. For FedCAMS, we have$k r f(x)k$ $G, k\hat{}_t k^2$ $\frac{4(1+q^2)^3}{(1-q^2)^2}$ $^2_l K^2G^2, km'_t k$ $_l KG$ and $kv_t k$ $\frac{4(1+q^2)^3}{(1-q^2)^2}$ $^2_l K^2G^2$, where$m'_t = {}_1 m'_{t-1} + (1-{}_1)\ _t$.

Proof. Since$f$ has$G$-bounded stochastic gradients, for any$x$ and$\ $, we have$kr f(x;\ )k$ G, we have

$$kr f(x)k = kE\ r f(x;\ )k \quad E\ kr f(x;\ )k \quad G:$$

For FedAMS, the model difference$_t$, by de nition, has the following formula,

$$_t = x^i_{t;K} - x_t = -{}_l \sum_{k=1}^{K} g^i_{t;k};$$

therefore,

$$k\ _t k \quad {}_l K\, kg^i_{t;k} k \quad {}_l KG:$$

Thus the bound for momentumm_t and variancev_t has the formula of

$$km_t k = (1-{}_1)\sum_{=1}^{t} {}_1^{t-} k\ _t k \quad {}_l KG;$$
$$kv_t k = (1-{}_2)\sum_{=1}^{t} {}_2^{t-} k\ _t k^2 \quad {}^2_l K^2G^2:$$

For the compressed version, FedCAMS, we have

$$\begin{aligned} k\hat{}_t k^2 &\quad kC(\ _t + e_t)k^2 \\ &\quad kC(\ _t + e_t) - (\ _t + e_t) + (\ _t + e_t)k^2 \\ &\quad 2(q^2+1)\, k\ _t + e_t k^2 \\ &\quad 4(q^2+1)[k\ _t k^2 + ke_t k^2] \\ &\quad \frac{4(1+q^2)^3}{(1-q^2)^2}\ ^2_l K^2G^2; \end{aligned}$$

where the third inequality holds due to Assumption 4.14,and the last inequality holds due to Lemma C.3. The virtual momentum sequence$m'_t k$ has the same bound as$_t$ of FedAMS. For the variance sequence of FedCAMS, we have

$$kv_t k = (1-{}_2)\sum_{=1}^{t} {}_2^{t-} k\hat{}_t k^2 \quad \frac{4(1+q^2)^3}{(1-q^2)^2}\ ^2_l K^2G^2:$$

This concludes the proof. □

Lemma C.5. The global model difference$_t = \sum_{i=1}^{m} {}^i_t$ in full participation cases satisfy

$$E[k\ _t k^2] \quad \frac{K}{m}\ ^2_l + \frac{^2_l}{m^2} E \sum_{i=1}^{m}\sum_{k=0}^{K-1} r F_i(x^i_{t;k}) \ ^2 :$$

Proof. For $\mathbb{E}[\|\Delta_t\|^2]$ in full participation case, we have

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \mathbb{E}\left\|\frac{1}{m}\sum_{i=1}^{m}\sum_{k=0}^{K-1}\eta_l g_{t,k}^i\right\|^2 \\
&= \frac{\eta_l^2}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1} g_{t,k}^i\right\|^2 \\
&= \frac{\eta_l^2}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}(g_{t,k}^i - \nabla F_i(x_{t,k}^i))\right\|^2 + \frac{\eta_l^2}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2 \\
&\le \frac{K\eta_l^2}{m}\sigma_l^2 + \frac{\eta_l^2}{m^2}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2,
\end{aligned} \tag{C.7}$$

where the inequality holds by Assumption 4.2. This concludes the proof. □

Lemma C.6. The global model difference $\Delta_t = \sum_{i\in S_t}\Delta_t^i$ in partial participation cases satisfy

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] \le{}& \frac{K\eta_l^2}{n}\sigma_l^2 + \frac{\eta_l^2(m-n)}{mn(m-1)}[15mK^3L^3\eta_l^2(\sigma_l^2+6K\sigma_g^2) + 90mK^4L^2\eta_l^2\sigma_l^2 + 3mK^2\|\nabla f(x_t)\|^2 \\
&+ 3mK^2\sigma_g^2] + \frac{\eta_l^2(n-1)}{mn(m-1)}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2.
\end{aligned}$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[\|\Delta_t\|^2] &= \mathbb{E}\left\|\frac{1}{n}\sum_{i\in S_t}\Delta_t^i\right\|^2 \\
&= \frac{1}{n^2}\mathbb{E}\left\|\sum_{i=1}^{m}\mathbb{I}\{i\in S_t\}\Delta_t^i\right\|^2 \\
&= \frac{\eta_l^2}{n^2}\mathbb{E}\left\|\sum_{i=1}^{m}\mathbb{I}\{i\in S_t\}\sum_{k=0}^{K-1}[g_{t,k}^i - \nabla F_i(x_{t,k}^i)]\right\|^2 + \left\|\sum_{i=1}^{m}\mathbb{I}\{i\in S_t\}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2 \\
&= \frac{\eta_l^2}{n^2}\mathbb{E}\left\|\sum_{i=1}^{m}\mathbb{P}\{i\in S_t\}\sum_{k=0}^{K-1}[g_{t,k}^i - \nabla F_i(x_{t,k}^i)]\right\|^2 + \left\|\sum_{i=1}^{m}\mathbb{P}\{i\in S_t\}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2 \\
&= \frac{\eta_l^2}{mn}\mathbb{E}\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}[g_{t,k}^i - \nabla F_i(x_{t,k}^i)]\right\|^2 + \frac{\eta_l^2}{n^2}\mathbb{E}\left\|\sum_{i=1}^{m}\mathbb{P}\{i\in S_t\}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2 \\
&\le \frac{K\eta_l^2}{n}\sigma_l^2 + \frac{\eta_l^2}{n^2}\mathbb{E}\left\|\sum_{i=1}^{m}\mathbb{P}\{i\in S_t\}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2,
\end{aligned} \tag{C.8}$$

where the fth equation holds due to $\mathbb{P}\{i\in S_t\} = \frac{n}{m}$. Note that we have

$$\begin{aligned}
\left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2 &= \left\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i) + \sum_{i\neq j}\sum_{k=0}^{K-1}\sum_{k=0}^{K-1}\langle\nabla F_i(x_{t,k}^i), \nabla F_j(x_{t,k}^j)\rangle\right. \\
&= \left.\sum_{i=1}^{m}m\left\|\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\right\|^2 - \frac{1}{2}\sum_{i\neq j}\sum_{k=0}^{K-1}\sum_{k=0}^{K-1}\|\nabla F_i(x_{t,k}^i) - \nabla F_j(x_{t,k}^j)\|^2\right.;
\end{aligned} \tag{C.9}$$

where the second equation holds due to $\|\sum_{i=1}^{m}x_i\|^2 = \sum_{i=1}^{m}m\|x_i\|^2 - \frac{1}{2}\sum_{i\neq j}\|x_i - x_j\|^2$. By the sampling strategy

(without replacement), we have $\mathbb{P}\{i \in S_t\} = \frac{n}{m}$ and $\mathbb{P}\{i, j \in S_t\} = \frac{n(n-1)}{m(m-1)}$, thus we have

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{k=0}^{K-1} \mathbb{P}\{i \in S_t\} \left\| \nabla F_i(x_{t,k}^i) \right\|^2 \\
&= \sum_{i=1}^{m} \mathbb{P}\{i \in S_t\} \sum_{k=0}^{K-1} \left\| \nabla F_i(x_{t,k}^i) \right\|^2 + \sum_{i \neq j} \mathbb{P}\{i, j \in S_t\} \sum_{k=0}^{K-1} \left\langle \nabla F_i(x_{t,k}^i), \sum_{k=0}^{K-1} \nabla F_j(x_{t,k}^j) \right\rangle \\
&= \frac{n}{m} \sum_{i=1}^{m}\sum_{k=0}^{K-1} \left\| \nabla F_i(x_{t,k}^i) \right\|^2 + \frac{n(n-1)}{m(m-1)} \sum_{i \neq j} \sum_{k=0}^{K-1} \left\langle \nabla F_i(x_{t,k}^i), \sum_{k=0}^{K-1} \nabla F_j(x_{t,k}^j) \right\rangle \\
&= \frac{n^2}{m} \sum_{i=1}^{m}\sum_{k=0}^{K-1} \left\| \nabla F_i(x_{t,k}^i) \right\|^2 - \frac{n(n-1)}{2m(m-1)} \sum_{i \neq j} \sum_{k=0}^{K-1} \left\| \nabla F_i(x_{t,k}^i) - \sum_{k=0}^{K-1} \nabla F_j(x_{t,k}^j) \right\|^2 \\
&= \frac{n(m-n)}{m(m-1)} \sum_{i=1}^{m}\sum_{k=0}^{K-1} \left\| \nabla F_i(x_{t,k}^i) \right\|^2 + \frac{n(n-1)}{m(m-1)} \left\| \sum_{i=1}^{m}\sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2,
\end{aligned} \tag{C.10}$$

where the third equation holds due to $\langle x, y \rangle = \frac{1}{2}[\|x\|^2 + \|y\|^2 - \|x-y\|^2]$ and the last equation holds due to $\frac{1}{2}\sum_{i \neq j} \|x_i - x_j\|^2 = \sum_{i=1}^{m} m\|x_i\|^2 - \|\sum_{i=1}^{m} x_i\|^2$. Therefore, for the last term in (C.8), we have

$$\mathbb{E}[\|\bar{e}_t\|^2] = \frac{K}{n}\frac{\sigma_l^2}{\sigma_l^2} + \frac{\sigma_l^2(m-n)}{mn(m-1)} \sum_{i=1}^{m} \mathbb{E}\left\| \sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2 + \frac{\sigma_l^2(n-1)}{mn(m-1)} \mathbb{E}\left\| \sum_{i=1}^{m}\sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2. \tag{C.11}$$

The second term in (C.11) is bounded partially following Reddi et al. (2020),

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{k=0}^{K-1} \left\| \nabla F_i(x_{t,k}^i) \right\|^2 &= \sum_{i=1}^{m} \mathbb{E}\sum_{k=0}^{K-1} \left\| [\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t^i) + \nabla F_i(x_t^i) - \nabla f(x_t) + \nabla f(x_t)] \right\|^2 \\
&\leq 3\sum_{i=1}^{m} \mathbb{E}\sum_{k=0}^{K-1} \left\| [\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t^i)] \right\|^2 + 3mK^2\sigma_g^2 + 3mK^2 \left\| \nabla f(x_t) \right\|^2 \\
&\leq 3K L^2 \sum_{i=1}^{m}\sum_{k=0}^{K-1} \mathbb{E}[\|x_{t,k}^i - x_t^i\|^2] + 3mK^2\sigma_g^2 + 3mK^2 \left\| \nabla f(x_t) \right\|^2 \\
&\leq 15mK^3L^3\sigma_l^2(\sigma_l^2 + 6K\sigma_g^2) + (90mK^4L^2\sigma_l^2 + 3mK^2) \left\| \nabla f(x_t) \right\|^2 + 3mK^2\sigma_g^2,
\end{aligned} \tag{C.12}$$

where the last inequality holds by applying Lemma C.9 (also follows from Reddi et al. (2020)). Substituting (C.12) into (C.11), this concludes the proof. □

Lemma C.7. Under Assumptions 4.1-4.3 and Assumption 4.14, for the momentum sequence $q_t = (1-\eta_1)\sum_{\tau=1}^{t} \eta_1^{t-\tau} g_\tau$ and accumulated error sequence $\bar{e}_t = (1-\eta_1)\sum_{\tau=1}^{t} \eta_1^{t-\tau} e_\tau$ in full participation settings, we have

$$\sum_{t=1}^{T} \mathbb{E}[\|m_t\|^2] \leq \frac{TK\sigma_l^2}{m}\sigma_l^2 + \frac{\sigma_l^2}{m^2} \sum_{t=1}^{T} \mathbb{E}\left\| \sum_{i=1}^{m}\sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2,$$

and

$$\sum_{t=1}^{T} \mathbb{E}[\|\bar{e}_t\|^2] \leq \frac{4T\eta_1^2}{(1-\eta_1^2)^2}\frac{K\sigma_l^2}{m}\sigma_l^2 + \frac{\sigma_l^2}{m^2}\frac{4\eta_1^2}{(1-\eta_1^2)^2} \sum_{t=1}^{T} \mathbb{E}\left\| \sum_{i=1}^{m}\sum_{k=0}^{K-1} \nabla F_i(x_{t,k}^i) \right\|^2.$$

Proof. By the updating rule, we have

$$\begin{aligned}
E[\|m_t\|^2] &= E\Big\|(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}g_\tau\Big\|^2 \\
&\le (1-\beta_1)^2\sum_{i=1}^{d}E\Big[\sum_{\tau=1}^{t}\beta_1^{t-\tau}g_{\tau,i}\Big]^2 \\
&\le (1-\beta_1)^2\sum_{i=1}^{d}E\Big[\sum_{\tau=1}^{t}\beta_1^{t-\tau}\sum_{\tau=1}^{t}\beta_1^{t-\tau}g_{\tau,i}^2\Big] \\
&\le (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}E[\|g_\tau\|^2] \\
&\le \frac{K\sigma_l^2}{m}\sigma_l^2+\frac{\sigma_l^2}{m^2}(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}E\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{\tau,k}^i)\Big\|^2, \qquad \text{(C.13)}
\end{aligned}$$

where the second inequality holds by applying Cauchy-Schwarz inequality, and the third inequality holds by summation of series. The last inequality holds by Lemma C.5. Hence summing over $t = 1, \ldots, T$, we have

$$\sum_{t=1}^{T}E[\|m_t\|^2] \le \frac{TK\sigma_l^2}{m}\sigma_l^2+\frac{\sigma_l^2}{m^2}\sum_{t=1}^{T}E\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\Big\|^2. \qquad \text{(C.14)}$$

For the compression error e_t, following the proof from Lemma C.3, for t and each local worker i, we have the induction

$$\begin{aligned}
E[\|e_{t+1}^i\|^2] &\le \frac{2q^2}{1-q^2}\sum_{\tau=1}^{t}\Big(\frac{1+q^2}{2}\Big)^{t-\tau}E[\|g_\tau^i\|^2] \\
&\le \frac{4q^2}{(1-q^2)^2}\frac{K\sigma_l^2}{m}\sigma_l^2+\frac{\sigma_l^2}{m^2}\sum_{\tau=1}^{t}\Big(\frac{1+q^2}{2}\Big)^{t-\tau}E\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{\tau,k}^i)\Big\|^2. \qquad \text{(C.15)}
\end{aligned}$$

For the sequence δ_t, similar as the previous analysis, we have

$$\begin{aligned}
E[\|\delta_t\|^2] &\le E\Big\|(1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}e_\tau\Big\|^2 \\
&\le (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}E[\|e_\tau\|^2] \\
&\le (1-\beta_1)\sum_{\tau=1}^{t}\beta_1^{t-\tau}\frac{1}{m}\sum_{i=1}^{m}E[\|e_\tau^i\|^2] \\
&\le \frac{4q^2}{(1-q^2)^2}\frac{K\sigma_l^2}{m}\sigma_l^2+\frac{\sigma_l^2}{m^2}\frac{2q^2(1-\beta_1)}{1-q^2}\sum_{\tau=1}^{t}\beta_1^{t-\tau}\sum_{j=1}^{\tau}\Big(\frac{1+q^2}{2}\Big)^{\tau-j}E\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{j,k}^i)\Big\|^2. \qquad \text{(C.16)}
\end{aligned}$$

Summing over $t = 1, \ldots, T$, we have

$$\begin{aligned}
\sum_{t=1}^{T}E[\|\delta_t\|^2] &\le \frac{4Tq^2}{(1-q^2)^2}\frac{K\sigma_l^2}{m}\sigma_l^2+\frac{\sigma_l^2}{m^2}\frac{2q^2}{1-q^2}\sum_{t=1}^{T}\sum_{\tau=1}^{t}\Big(\frac{1+q^2}{2}\Big)^{t-\tau}E\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{\tau,k}^i)\Big\|^2 \\
&\le \frac{4Tq^2}{(1-q^2)^2}\frac{K\sigma_l^2}{m}\sigma_l^2+\frac{\sigma_l^2}{m^2}\frac{4q^2}{(1-q^2)^2}\sum_{t=1}^{T}E\Big\|\sum_{i=1}^{m}\sum_{k=0}^{K-1}\nabla F_i(x_{t,k}^i)\Big\|^2. \qquad \text{(C.17)}
\end{aligned}$$

□

Lemma C.8. *Under Assumptions 4.1-4.3 and Assumption 4.14, for the momentum sequence* $\mathsf{m}_t = (1-\ _1)\sum_{\tau=1}^{t}\ _1^{t-\tau}\Delta_\tau$ *in partial participation settings, we have*

$$\sum_{t=1}^{T}\mathsf{E}[\|\mathsf{m}_t\|^2] \le \frac{KT\ _l^2}{n}\ _l^2 + \frac{\ _l^2}{n^2}\sum_{t=1}^{T}\mathsf{E}\Big\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\nabla F_i(\mathsf{x}_{t,k}^i)\Big\|^2:$$

Proof. The proof outline is the same as the proof of Lemma C.7, the main difference is $\mathsf{E}[\|\Delta_t\|^2]$ has changed, so we need to apply Lemma C.6 instead of Lemma C.5 during the proof. □

Lemma C.9. *(This lemma directly follows from Lemma 3 in FedAdam (Reddi et al., 2020). For local learning rate which satisfying* $_l \le \frac{1}{8KL}$*, the local model difference after* k *(*$\forall k \in \{0;1;:::;K-1\}$*) steps local updates satisfies*

$$\frac{1}{m}\sum_{i=1}^{m}\mathsf{E}[\|\mathsf{x}_{t,k}^i - \mathsf{x}_t\|^2] \le 5K\ _l^2(\ _l^2 + 6K\ _g^2) + 30K^2\ _l^2\mathsf{E}[\|\nabla f(\mathsf{x}_t)\|^2]: \tag{C.18}$$

Proof. The proof of Lemma C.9 is exactly same as the proof of Lemma 3 in Reddi et al. (2020). □

D. Additional Discussions

The additional server-to-worker communication: Our current analysis only focus on one-way compression from worker to server while the server-to-worker broadcasting is still uncompressed since of cost of broadcasting is in general cheaper than worker to server uploading. Note that it is also straightforward to compress x for server-to-worker communication in the full participation scheme (with guarantees). So we can indeed achieve high communication efficiency even for two-way compression. Table 1 shows the communication bits comparison for scaled sign and top-k compressors, where T is the total iteration of training and d denotes the dimension of x. Specifically, Table 2 shows the communication bits corresponding to the experiments showing by Figure 4. However, for the partial participating setting, it will encounter a synchronization issue, which is highly non-trivial to solve. Thus we leave the two-way compression strategy for future work.

Method	Uncompressed	One-way Compression	Two-way Compression
Scaled sign	$32d \times 2T$	$(32+d)\times T + 32d \times T$	$(32+d)\times 2T$
Top-k	$32d \times 2T$	$\approx 32(2k+d)\times T$	$\approx 32 \times 2k \times 2T$

Table 1. Communication bits comparisons for scaled sign and top-k compressors.

Method	Uncompressed	One-way Compression	Two-way Compression
Scaled sign	$3{:}58 \times 10^{11}$	$1{:}84 \times 10^{11}$	$1{:}12 \times 10^{10}$
Top-k with $r = 1/64$	$3{:}58 \times 10^{11}$	$1{:}84 \times 10^{11}$	$1{:}12 \times 10^{10}$
Top-k with $r = 1/128$	$3{:}58 \times 10^{11}$	$1{:}82 \times 10^{11}$	$5{:}59 \times 10^{9}$
Top-k with $r = 1/256$	$3{:}58 \times 10^{11}$	$1{:}80 \times 10^{11}$	$2{:}79 \times 10^{9}$

Table 2. Approximate communication bits comparisons for scaled sign and top-k with $r = 1/64$, $r = 1/128$ and $r = 1/256$ compressors when training CIFAR-10 on ResNet-18 model for 500 rounds.

E. Additional Experimental Results

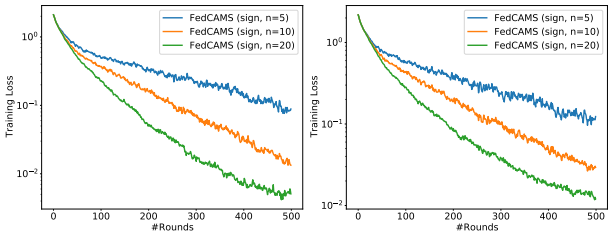
E.1. Hyperparameter Settings

We conduct detailed hyperparameter searches to find the best hyperparameters for each baseline methods including ours. In details, we grid search over the local learning rate $_l \in \{0{:}0001; 0{:}001; 0{:}01; 0{:}1; 1{:}0\}$, the global learning rate $\in \{0{:}001; 0{:}01; 0{:}1; 1{:}0\}$ for all methods. For adaptive federated optimization methods, we set $_1 = 0{:}9$, $_2 = 0{:}99$. For FedAdam, FedYogi, and FedAMSGrad, we search the best from $\{10^{-8}; 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}; 10^{0}\}$. For FedAMS and FedCAMS, we search the max stabilization from $\{10^{-8}; 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}; 10^{0}\}$.

Specifically, for our ResNet-18 experiments, we set the local learning rate $_l = 0.01$ and the global learning rate $= 1.0$ for FedAvg, set $_l = 0.01$, $= 0.1$ and $= 0.1$ for FedAdam and FedAMSGrad, set $_l = 0.01$, $= 1.0$ and $= 0.1$ for FedYogi, set $_l = 0.01$, $= 1.0$ and max stabilization $= 0.001$ for FedAMS and FedCAMS. For our ConvMixer-256-8 experiments, we set the local learning rate $_l = 0.01$ and the global learning rate $= 1.0$ for FedAvg, set $_l = 0.01$, $= 1.0$ and $= 0.1$ for FedAdam, FedYogi and FedAMSGrad, set $_l = 0.01$, $= 1.0$ and max stabilization $= 0.001$ for FedAMS and FedCAMS.

E.2. Additional Experiments

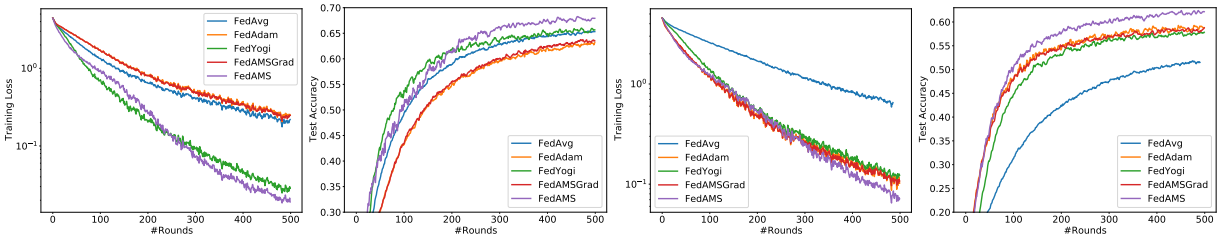
Figure 6 shows the effect of parameter n on the convergence rate of FedCAMS with choosing groups of parameters: $n \in \{5, 10, 20\}$. For both ResNet-18 and ConvMixer-256-8 models, it is shown that a larger number of participating clients n achieves a faster convergence rate, this backs up our theory.



(a) ResNet-18 (b) ConvMixer-256-8

Figure 6. The learning curves for FedCAMS with different participating number of clients n in training CIFAR-10 data on ResNet-18 and ConvMixer-256-8 models.

Figure 7 shows the convergence result of FedAMS and other federated learning baselines on training CIFAR-100 dataset with the ResNet-18 model and the ConvMixer-256-8 model. We compare the training loss and test accuracy against the global rounds for each model. For the ResNet-18 model, FedAMS and FedYogi achieve significantly better performance comparing with other three baselines. In particular, FedYogi has a fast convergence rate at the beginning status, while FedAMS performs the best in terms of the final training loss and test accuracy. FedAvg achieves a slightly better training loss to FedAdam and FedAMSGrad but much higher test accuracy which is close to FedYogi and FedAMS.



(a) ResNet-18 (b) ResNet-18 (c) ConvMixer-256-8 (d) ConvMixer-256-8

Figure 7. The learning curves for FedAMS and other federated learning baselines on training CIFAR-100 data (a)(b) show the results for ResNet-18 model and (c)(d) show the results for ConvMixer-256-8 model.

For the ConvMixer-256-8 model, which is typically trained via adaptive gradient method, we observe that all adaptive federated optimization methods (FedAdam, FedYogi, FedAMSGrad and FedAMS) achieve much better performance in terms of both training loss and test accuracy compared with FedAvg. In details, FedAMS again achieves a significantly better result than other baselines in terms of training loss and test accuracy. Other adaptive methods, including FedAdam, FedYogi, and FedAMSGrad, have similar convergence behaviour when training the ConvMixer-256-8 model. Such results empirically show the effectiveness of our proposed FedAMS method.

E.3. Additional Ablation Study

The ablation on : We conduct an ablation study with $\in \{10^{-1}; 10^{-2}; 10^{-3}; 10^{-4}; 10^{-6}; 10^{-8}\}$ on CIFAR-10 in Table 3 and our value in experiments is chosen by its relatively higher test accuracy.

	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-6}	10^{-8}
Test acc (%)	90.45	90.51	**90.94**	90.72	90.49	90.30

Table 3. Ablation study on ϵ.