

On the Role of Server Momentum in Federated Learning

Jianhui Sun^{1*}, Xidong Wu^{2*}, Heng Huang³, Aidong Zhang¹

¹Computer Science, University of Virginia, VA, USA

²Electrical and Computer Engineering, University of Pittsburgh, PA, USA

³Computer Science, University of Maryland College Park, MD, USA

js9gu@virginia.edu, xidong_wu@outlook.com, heng@umd.edu, aidong@virginia.edu

Abstract

Federated Averaging (FedAvg) is known to experience convergence issues when encountering significant clients system heterogeneity and data heterogeneity. Server momentum has been proposed as an effective mitigation. However, existing server momentum works are restrictive in the momentum formulation, do not properly schedule hyperparameters and focus only on system homogeneous settings, which leaves the role of server momentum still an under-explored problem. In this paper, we propose a general framework for server momentum, that (a) covers a large class of momentum schemes that are unexplored in federated learning (FL), (b) enables a popular stagewise hyperparameter scheduler, (c) allows heterogeneous and asynchronous local computing. We provide rigorous convergence analysis for the proposed framework. To our best knowledge, this is the first work that thoroughly analyzes the performances of server momentum with a hyperparameter scheduler and system heterogeneity. Extensive experiments validate the effectiveness of our proposed framework.

1 Introduction

Federated Averaging (FedAvg) (McMahan et al. 2017), which runs multiple epochs of Stochastic Gradient Descent (SGD) locally in each client and then averages the local model updates once in a while on the server, is probably the most popular algorithm to solve many federated learning (FL) problems, mainly due to its low communication cost and appealing convergence property.

Though it has seen great empirical success, vanilla FedAvg experiences an unstable and slow convergence when encountering *client drift*, i.e., the local client models move away from globally optimal models due to client heterogeneity (Karimireddy et al. 2020). On the server side, FedAvg is in spirit similar to an SGD with a constant learning rate one and updates the global model relying only on the averaged model update from the current round, thus extremely vulnerable to client drift. Note that in non-FL settings, SGD in its vanilla form has long been replaced by some momentum scheme (e.g. heavy ball momentum (SHB) and Nesterov’s accelerated gradient (NAG)) in many tasks, as mo-

mentum schemes achieve an impressive training time saving and generalization performance boosting compared to competing optimizers (Sutskever et al. 2013; Wilson et al. 2017), which promises a great potential to apply momentum in FL settings as well. Incorporating server momentum essentially integrates historical aggregates into the current update, which could conceptually stabilize the global update against dramatic local drifts.

Though various efforts have been made to understand the role of server momentum in FL, e.g. (Hsu, Qi, and Brown 2019; Rothchild et al. 2020), it is still largely an under-explored problem due to the following reasons:

(1) Lack of diversity in momentum schemes. Most existing server momentum works only focus on SHB (e.g. FedAvgM (Hsu, Qi, and Brown 2019)). It is unclear whether many momentum schemes that outperformed SHB in non-FL settings can also perform better in FL, and there is no unified analysis for momentum schemes other than SHB.

(2) No hyperparameter schedule. Properly scheduling hyperparameters is key to train deep models more efficiently and an appropriate selection of server learning rate η_t is also important in obtaining optimal convergence rate (Yang, Fang, and Liu 2021). Existing works either still employ a constant server learning rate one or consider a η_t schedule that is uncommonly used in practice, e.g., polynomially decay (i.e., $\eta_t \propto \frac{1}{t^\alpha}$) (Khanduri et al. 2021). Moreover, it is known that increasing momentum factor β is also a critical technique in deep model training (Sutskever et al. 2013; Smith, Kindermans, and Le 2018), while to our best knowledge, there is no prior work considering time-varying β in FL.

(3) Ignoring client system heterogeneity. Existing works make unrealistic assumptions on system homogeneity and client synchrony, e.g., clients are sampled uniformly at random, all participating clients synchronize at each round t , and all clients run identical number of local epochs, none of which holds in most cross device FL deployments (Kairouz et al. 2021). System heterogeneity (i.e., the violation of above assumptions), alongside with data heterogeneity, is also a main source client drift (Karimireddy et al. 2020). Thus, ignoring it would provide an incomplete understanding of the role of server momentum.

To address the above limitations, we propose a novel formulation which we refer to as Federated General Mo-

*These authors contributed equally.

mentum (FedGM). FedGM includes the following hyperparameters, learning rate η , momentum factor β , and instant discount factor ν . With different specifications of (η, β, ν) , FedGM subsumes the FL version of many popular momentum schemes, most of which have never been explored in FL yet.

We further incorporate a widely used hyperparameter scheduler “constant and drop” (a.k.a. “step decay”) in FedGM. We refer to this framework as multistage FedGM. Specifically, with a prespecified set of hyperparameters $\{\eta_s, \beta_s, \nu_s\}_{s=1}^S$ and training lengths $\{T_s\}_{s=1}^S$, the training process is divided into S stages, and at stage s , FedGM with $\{\eta_s, \beta_s, \nu_s\}$ is applied for T_s rounds. Compared to many unrealistic schedule in existing works, “constant and drop” is the de-facto scheduler in most model training (Sutskever et al. 2013; He et al. 2016; Huang et al. 2017). Multistage FedGM is extremely flexible, as it allows the momentum factor to vary stagewise as well, and subsumes single-stage training as a special case. We provide the convergence analysis of multistage FedGM. Our theoretical results reveal why stagewise training can provide empirically faster convergence.

Furthermore, in order to understand how server momentum behaves in the presence of system heterogeneity, we propose a framework that we refer to as Autonomous Multistage FedGM, in which clients can do heterogeneous and asynchronous computing. Specifically, we allow each client to (a) update local models based on an asynchronous view of the global model, (b) run a time-varying, client-dependent number of local epochs, and (c) participate at will. We provide convergence analysis of Autonomous Multistage FedGM. Autonomous Multistage FedGM is a more realistic characterization of real-world cross-device FL applications.

Finally, we conduct extensive experiments that validate, (a) FedGM is a much more capable momentum scheme compared with existing FedAvgM in both with and without system heterogeneity settings; and (b) multistage hyperparameter scheduler further improves FedGM effectively.

Our main contributions can be summarized as follow,

- We propose FedGM, which is a general framework for server momentum and covers a large class of momentum schemes that are unexplored in FL. We further propose Multistage FedGM, which incorporates a popular hyperparameter scheduler to FedGM.
- We show the convergence of multistage FedGM in both full and partial participation settings. We also empirically validate the superiority of multistage FedGM. To our best knowledge, this is the first work that provides convergence analysis of server-side hyperparameter scheduler.
- We propose Autonomous Multistage FedGM, which requires much less coordination between server and workers than most existing algorithms, and theoretically analyze its convergence. Our work is the first to study the interplay between server momentum and system heterogeneity.

The rest of the paper is organized as follows. In Section 2, we formally introduce federated optimization. In Section 3, we introduce Federated General Momentum (FedGM), followed by multistage FedGM and its convergence analysis

Algorithm 1: FedOPT (Reddi et al. 2020): A Generic Formulation of Federated Optimization

Input: Number of clients n , objective function $f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x)$, initialization x_0 , Number of communication rounds T , **local** learning rate η_l , **local** number of updates K , **global** hyperparameters \mathbb{H} ;

- 1 **for** $t \in \{1, \dots, T\}$ **do**
- 2 Randomly sample a subset \mathcal{S}_t of clients
- 3 Server sends x_t to subset \mathcal{S}_t of clients
- 4 **for each client** $i \in \mathcal{S}_t$ **do**
- 5 $\Delta_t^i = \text{LocalOPT}(i, \eta_l, K, x_t)$
- 6 **end**
- 7 Server aggregates $\Delta_t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \Delta_t^i$
- 8 $x_{t+1} = \text{ServerOPT}(x_t, \Delta_t, \mathbb{H})$
- 9 **end**
- 10 **return** x_T

Algorithm 2: LocalOPT (i, η_l, K, x_t)

Input: client index i , data distribution \mathcal{D}_i , **local** learning rate η_l , **local** updating number K , round t , **global** model x_t ;

- 1 Initialize $x_{t,0}^i \leftarrow x_t$
- 2 **for** $k \in \{0, 1, \dots, K-1\}$ **do**
- 3 Randomly sample a batch $\xi_{t,k}^i$ from \mathcal{D}_i
- 4 Compute $g_{t,k}^i = \nabla f_i(x_{t,k}^i, \xi_{t,k}^i)$
- 5 Update $x_{t,k+1}^i = x_{t,k}^i - \eta_l g_{t,k}^i$
- 6 **end**
- 7 $\Delta_t^i = x_t - x_{t,K}^i$ **return** Δ_t^i

in Section 4. In Section 5, we introduce Autonomous multistage FedGM and provide its convergence analysis. Section 6 presents the experimental results. Due to the page limit, we leave related work, all proofs, and additional experimental results to the Appendix.

2 Background: FedOPT and FedAvg

Many FL tasks can be formulated as solving the following optimization problems,

$$\min_{x \in \mathbb{R}^d} f(x) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(x), \quad \text{where } f_i(x) = \mathbb{E}_{\xi \sim \mathcal{D}_i} f_i(x, \xi). \quad (1)$$

where n is the total number of clients, x is the model parameter with d dimension. Each client i has a local data distribution \mathcal{D}_i and a local objective function $f_i(x) = \mathbb{E}_{\xi \sim \mathcal{D}_i} f_i(x, \xi)$. The global objective function is the averaged objective among all clients. \mathcal{D}_i can be very different from \mathcal{D}_j when $i \neq j$.

FedAvg (McMahan et al. 2017) and its variants are a special case of a more flexible formulation, **FedOPT** (Reddi et al. 2020), which is formalized in Algorithm 1. Suppose the total number of rounds is T , and the global model param-

eter is $\{x_t\}_{t=1}^T$. At each round t , the server randomly samples a subset of clients \mathcal{S}_t and sends the global model x_t to them. Upon receiving x_t , each participating client would do **LocalOPT** (Algorithm 2). Specifically, each client i would initialize their local model at x_t , run K steps of local SGD with local η_l and the local model is updated to $x_{t,K}^i$. The client then sends the local model update $\Delta_t^i = x_t - x_{t,K}^i$ back to the server. The server aggregates by averaging, i.e. $\Delta_t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \Delta_t^i$, and then triggers server-side optimization **ServerOPT**, which takes x_t , aggregated model update Δ_t , and a hyperparameter set \mathbb{H} as input, and outputs the next round’s global model parameter x_{t+1} .

In FedAvg, ServerOPT is simply $x_{t+1} = x_t - \Delta_t$, which is in spirit identical to SGD with a constant learning rate one if viewing Δ_t as a pseudo gradient.

3 FedGM: Federated Learning with General Momentum Acceleration

Partially due to its equivalence of constant learning rate SGD, FedAvg has two main limitations, (a) it is extremely vulnerable to client drift, as FedAvg relies entirely on its current aggregate Δ_t and ignores historical directions; (b) FedAvg may not be the best option in many applications, e.g. training large-scale vision or language models (Devlin et al. 2018; Dosovitskiy et al. 2021) where its counterpart SGD is known to be inferior to momentum or adaptive optimizers in non-FL settings (Wilson et al. 2017; Zhang et al. 2020).

Note that in FedOPT, ServerOPT could in principle be any type of gradient-based optimizers. In non-FL settings, the momentum scheme is known to not only exhibit convincing accelerating effect in training, it has also achieved better generalizability in many tasks than adaptive optimizers like Adam (Wilson et al. 2017; Cutkosky and Mehta 2020), which provides a strong motivation to incorporate server momentum.

Moreover, server-side momentum basically integrates historical aggregates into the current update and therefore could potentially make the global model more robust to drastic local drifts.

Existing server momentum works mostly focus on one specific type of momentum, i.e. stochastic heavy ball momentum (SHB) (Hsu, Qi, and Brown 2019; Rothchild et al. 2020; Khanduri et al. 2021), while ignoring many other momentum schemes that outperform SHB in many non-FL settings.

In order to systematically understand the role of server momentum schemes in FL, we propose a new algorithm which we refer to as Federated General Momentum (FedGM). FedGM replaces the ServerOPT $x_{t+1} = x_t - \Delta_t$ in FedAvg with the following,

$$\begin{aligned} d_{t+1} &= (1 - \beta)\Delta_t + \beta d_t, & h_{t+1} &= (1 - \nu)\Delta_t + \nu d_{t+1}, \\ x_{t+1} &= x_t - \eta h_{t+1}. \end{aligned} \quad (2)$$

where the hyperparameter set $\mathbb{H} = \{\eta, \beta, \nu\}$. η is server learning rate, β and ν are two hyperparameters which we call momentum factor and instant discount factor.

By setting ν as 0, FedGM becomes FedAvg with two-sided learning rates (Yang, Fang, and Liu 2021), i.e., choices of η other than 1 is allowed, which we refer to as FedSGD.

By setting $\nu = 1$, FedGM becomes FedAvgM (Hsu, Qi, and Brown 2019) (or FedSHB), which essentially applies server SHB, i.e. we update the model by a ‘‘momentum buffer’’ d_{t+1} . β controls how slowly the momentum buffer is updated. FedGM could be interpreted as a ν -weighted average of the FedAvgM update step and the plain FedAvg update step. ν is thus referred to as instant discount factor.

FedGM leverages the general formulation of QHM (Ma and Yarats 2019) and is much more general than just FedAvg and FedAvgM. It subsumes many other momentum variants that are never explored in FL. For example, if $\nu = \beta$, FedGM becomes a new algorithm which can be naturally referred to as FedNAG, i.e. application of the popular optimizer Nesterov’s accelerated gradient (NAG) to FL. Specifically, we update model by $x_{t+1} = x_t - \eta[(1 - \beta)\Delta_t + \beta d_{t+1}]$, where d_{t+1} is the momentum buffer.

FedGM could further recover the FL version of many other momentum schemes, e.g., SNV (Lessard, Recht, and Packard 2014), PID (An et al. 2018), ASGD (Kidambi et al. 2018), and Triple Momentum (Van Scoy, Freeman, and Lynch 2018), with different η, β, ν . Therefore, FedGM describes a family of momentum schemes, most of which have not been studied yet in FL.

4 Multistage FedGM and Convergence

4.1 Proposed Algorithm: Multistage FedGM

One major limitation in FedGM (2) is that all server-side hyperparameters are held constant, which are inconsistent with common practice. Adaptively adjusting hyperparameters throughout the training is key to the success of many optimizers. Learning rate scheduling has been thoroughly studied in non-FL settings, e.g., (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Goyal et al. 2017; Smith 2017). Scheduling other hyperparameters (e.g. momentum factor and batch size) is also shown to be very effective in many settings. For example, (Sutskever et al. 2013; Smith and Le 2018; Smith, Kindermans, and Le 2018) show a slowly increasing schedule for the momentum factor β is crucial in training deep models faster.

We focus on a simple yet effective hyperparameter scheduler, ‘‘constant and drop’’ (a.k.a. ‘‘step decay’’). In its non-FL SGD version (a.k.a. multistage SGD), with a prespecified set of learning rates $\{\eta_s\}_{s=1}^S$ and training lengths $\{T_s\}_{s=1}^S$ (measured by number of iterations/epochs), the training process is divided into S stages, and SGD with η_s is applied for T_s iterations/epochs at s -th stage, where $\{\eta_s\}_{s=1}^S$ is usually a non-increasing sequence¹. We concentrate on ‘‘constant and drop’’ as it is the de-facto learning rate scheduler in most large-scale neural networks (Krizhevsky, Sutskever, and Hinton 2012; Sutskever et al. 2013; He et al. 2016; Huang et al. 2017), and has been theoretically shown to achieve near-

¹The name ‘‘constant and drop’’ refers to learning rate is dropped by some constant factor after each stage.

Algorithm 3: Multistage FedGM

Input:

Initialization x_0 , number of rounds T , **local** learning rate η_l , **local** updating number K ;

Number of stages S , stage lengths $\{T_s\}_{s=1}^S$;

Stagewise hyperparameters $\{\eta_s, \beta_s, \nu_s\}_{s=1}^S$;

```
1 for  $s \in \{1, \dots, S\}$  do
2   for  $t$  in stage  $s$  do
3     Randomly sample a subset  $\mathcal{S}_t$  of clients
4     Server sends  $x_t$  to subset  $\mathcal{S}_t$  of clients
5     for each client  $i \in \mathcal{S}_t$  do
6        $\Delta_t^i = \text{LocalOPT}(i, \eta_l, K, x_t)$ 
7     end
8     Server aggregates  $\Delta_t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \Delta_t^i$ 
9      $d_{t+1} = (1 - \beta_s)\Delta_t + \beta_s d_t$ 
10     $h_{t+1} = (1 - \nu_s)\Delta_t + \nu_s d_{t+1}$ 
11    Update  $x_{t+1} = x_t - \eta_s h_{t+1}$ 
12  end
13 end
14 return  $x_T$ 
```

optimal rate in non-FL settings (Ge et al. 2019b; Wang, Magnússon, and Johansson 2021).

The intuition behind “constant and drop” is straightforward: a large learning rate is held constant for a reasonably long period to take advantage of faster convergence until it saturates, and then the learning rate is dropped by a constant factor for more refined training.

We extend “constant and drop” to FedGM in Algorithm 3, which we refer to as Multistage FedGM. In Multistage FedGM (Algorithm 3), each stage has length T_s ($T = \sum_{s=1}^S T_s$), and has its triplet of stagewise hyperparameters $\{\eta_s, \beta_s, \nu_s\}_{s=1}^S$. The convergence analysis in Sec 4.2 also applies to single-stage FedGM by $S = 1$.

To our best knowledge, there is no prior work giving definitive theoretical guarantee or empirical performances of any hyperparameter schedule in FL, especially considering multistage FedGM is an extremely flexible framework that allows both learning rate and momentum factor to evolve.

4.2 Convergence Analysis of Multistage FedGM

We now analyze the convergence of Algorithm 3 under both full and partial participation settings.

We aim to optimize objective (1). Each local loss f_i (and therefore f) may be general nonconvex function. We study the general *non-i.i.d.* setting, i.e. $\mathcal{D}_i \neq \mathcal{D}_j$ when $i \neq j$. We state the assumptions that are needed in the analysis.

Assumption 1 (Smoothness). Each local loss $f_i(x)$ is differentiable and has L -Lipschitz continuous gradients, i.e., $\forall x, x' \in \mathbb{R}^d$, we have $\|\nabla f_i(x) - \nabla f_i(x')\| \leq L \|x - x'\|$. And $f^* \triangleq \min_x f(x)$ exists, i.e., $f^* > -\infty$.

Assumption 2 (Bounded Local Variance). $\forall t, i$, LocalOPT can access an unbiased estimator $g_{t,k}^i = \nabla f_i(x_{t,k}^i, \xi_{t,k}^i)$ of true gradient $\nabla f_i(x_{t,k}^i)$, where $g_{t,k}^i$ is the stochastic gradient estimated with minibatch $\xi_{t,k}^i$. And each stochastic gradient on the i -th client has a bounded local variance, i.e., we have $\mathbb{E} \left[\left\| g_{t,k}^i - \nabla f_i(x_{t,k}^i) \right\|^2 \right] \leq \sigma_l^2$.

Assumption 3 (Bounded Global Variance). The local loss $\{f_i(x)\}$ across all clients have bounded global variance, i.e., $\forall x$, we have $\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \sigma_g^2$.

Assumption 1-3 are standard assumptions in nonconvex optimization and FL research, and have been universally adopted in most existing works (Reddi, Kale, and Kumar 2018; Li et al. 2020b; Reddi et al. 2020; Bao, Gu, and Huang 2020; Yang, Fang, and Liu 2021; Wang, Lin, and Chen 2022; Wu et al. 2023b,c). $\sigma_g^2 = 0$ in Assumption 3 corresponds to the *i.i.d.* setting. And we do not require the restrictive bounded gradient assumption (Reddi, Kale, and Kumar 2018; Avdiukhin and Kasiviswanathan 2021; Wu et al. 2023a).

Recall $T = \sum_{s=1}^S T_s$ is the number of rounds. Denote the expected gradient square as $\{\mathcal{G}_t \triangleq \mathbb{E} [\|\nabla f(x_t)\|^2]\}_{t \leq T}$. Define the average expected gradient square at s -th stage as $\bar{\mathcal{G}}_s \triangleq \frac{1}{T_s} \sum_{t=T_1+\dots+T_{s-1}+1}^{T_1+\dots+T_s} \mathcal{G}_t$ and the average expected gradient square across S stages as $\bar{\mathcal{G}} \triangleq \frac{1}{S} \sum_{s=1}^S \bar{\mathcal{G}}_s$. Bounding $\bar{\mathcal{G}}$ generalizes from bounding $\frac{1}{T} \sum_{t=1}^T \mathbb{E} [\|\nabla f(x_t)\|^2]$ in single-stage to multistage setting.

To reflect the common hyperparameter scheduling practice that is adopted by existing works e.g. (Sutskever et al. 2013; Smith, Kindermans, and Le 2018; Liu, Gao, and Yin 2020), We request the stagewise hyperparameters fulfill the following constraints,

$$\begin{aligned} \eta_S \leq \eta_{S-1} \leq \dots \leq \eta_1 \quad \beta_1 \leq \beta_2 \leq \dots \leq \beta_S < 1 \\ W_1 \equiv \frac{\eta_s \beta_s \nu_s}{1 - \beta_s} \quad \text{and} \quad W_2 \equiv T_s \eta_s \end{aligned} \quad (3)$$

where W_1 and W_2 are two constants. Constraint (3) essentially requires learning rate to be non-increasing and momentum factor to be non-decreasing at a similar rate, which is consistent with common practice, e.g. for SHB and NAG, (Sutskever et al. 2013; Smith, Kindermans, and Le 2018; Liu, Gao, and Yin 2020) propose a scheduler for β to increase and close to 1 for faster convergence. And it is also natural for (3) to require $T_s \eta_s$ as a constant. As the learning rate is decaying, more rounds in later stages are necessary for sufficient refined training.

We now state the convergence guarantee of the multistage training regime in FL framework.

Full Participation If all clients are required to participate in each round, i.e. $\mathcal{S}_t = \{1, 2, \dots, n\}$, we have,

Theorem 4.1. *We optimize $f(x)$ with Algorithm 3 (Full Participation) under Assumptions 1-3. Denote $\bar{\eta} \triangleq \frac{1}{S} \sum_{s=1}^S \eta_s$ as the average server learning rate and $C_\eta \triangleq \frac{\eta_1}{\eta_S}$. Under*

the condition ² $\eta_l \leq \min \left\{ \frac{1}{8KL}, \frac{1}{KSC_\eta(L\bar{\eta}+1+L^2W_1^2C_\eta)} \right\}$, we would have:

$$\begin{aligned} \bar{G} &\triangleq \frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\ &\leq \frac{64}{17} \frac{f(x_0) - f^*}{SW_2\eta_l K} + \Psi_l \sigma_l^2 + \Psi_g \sigma_g^2 \end{aligned}$$

where $\Psi_l \triangleq \frac{32}{17} \frac{L^2 W_1^2 T \bar{\eta} \eta_l}{n W_2} + \frac{32}{17} \frac{L \bar{\eta} \eta_l}{n} + \frac{32}{17} \frac{\eta_l}{n} + \frac{160}{17} \eta_l^2 L^2 K$, and $\Psi_g \triangleq \frac{960}{17} \eta_l^2 L^2 K^2$.

Corollary 4.2 (Convergence Rate of Multistage FedAvg). Suppose $\nu_s = 0$, i.e., the FedAvg algorithm that allows learning rate vary across S stages. By setting $\bar{\eta} = \Theta(\sqrt{nK})$ and $\eta_l = \Theta\left(\frac{1}{\sqrt{TK}}\right)$, $W_2 = \Theta\left(\frac{T\sqrt{nK}}{S}\right)$, i.e. $T\bar{\eta}$ equally divided into S stages. $W_1 = 0$ as $\nu_s = 0$. Suppose T is sufficiently large, i.e. $T \geq nK$, we have a $\mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right)$ convergence rate.

Remark 4.3 (Why Multistage Helps?). Corollary 4.2 indicates multistage FedAvg recovers the best-known rate for general FL nonconvex optimization approaches, e.g. SCAF-FOLD (Karimireddy et al. 2020) and FedAdam (Reddi et al. 2020). Note single-stage FedAvg with two-sided learning rates also achieves the same rate (Yang, Fang, and Liu 2021). However, we do observe multistage FedAvg empirically converges much better than single-stage. We can obtain insights from Theorem 4.1 why multistage helps. We note that Ψ_l is only related to average learning rate $\bar{\eta}$ (instead of initial learning rate η_l). At initial rounds, the first term with $f(x_0) - f^*$ dominates, and thus we could select a relatively large η_l to ensure a more dramatic decay of this term. At later rounds, when $f(x_t) - f^*$ plateaus, we could enable smaller learning rate to control $\bar{\eta}$. Thus, Theorem 4.1 indicates a less stringent reliance on η_l , which enables us to flexibly select suitable η depending on which training stage we are in, that can still guarantee convergence.

Corollary 4.4 (Convergence Rate of Multistage FedGM). Suppose $S > 1$, i.e. the multistage regime, by setting $\bar{\eta} = \Theta(\sqrt{nK})$, $\eta_l = \Theta\left(\frac{1}{\sqrt{TK}}\right)$, $W_2 = \Theta\left(\frac{T\sqrt{nK}}{S}\right)$. Let $W_1^2 = \mathcal{O}\left(\frac{\sqrt{nK}}{S}\right)^3$. When $T \geq Kn$, we have a $\mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right)$ convergence rate.

Remark 4.5 (Why Momentum Helps?). We attribute the empirically superior performances of momentum to two reasons. (a) When clients are dynamically heterogeneous, historical gradient information has regularization effect to

²The condition could be fulfilled by typical value assignment, and would recover the typical $\eta_l \leq \min \left\{ \frac{1}{8KL}, \frac{1}{KL\eta} \right\}$ constraint in FedAvg analysis (Yang, Fang, and Liu 2021), by setting $S = 1$.

³It holds by setting an infinitesimal β or ν at early stages when η is large, but β or ν can go to 1 when η is reduced to $o\left(\frac{\sqrt{nK}}{\sqrt{S}}\right)$.

avoid the search direction from going wild. (b) Server learning rate η acts like a multiplier to client learning rate η_l in FedAvg, i.e. $\eta > 1$ effectively enhances the reliance on current round gradient. Due to the same reason as in (a), such reliance can harm convergence. In contrast, in FedGM, β and ν act as a buffer that could to some extent absorb the impact from a large η . We empirically observe in Appendix H.2, with same η_l , FedGM could sustain a much larger η , while FedAvg diverges very easily with a moderately large η .

Partial Participation Full participation rarely holds in reality, thus we further analyze multistage FedGM in partial participation setting ⁴.

Theorem 4.6. We optimize $f(x)$ with Algorithm 3 (Partial Participation) under Assumptions 1-3. Denote $\bar{\eta}$ and C_η as in Theorem 4.1. Under the condition $\eta_l \leq \frac{1}{8KL}$, and $\eta_l (C_\eta + L\bar{\eta}C_\eta + L^2W_1^2C_\eta) SK \leq \min \left\{ \frac{m(n-1)}{n(m-1)}, \frac{17m}{282} \right\}$, we would have:

$$\bar{G} \leq \frac{64}{17} \frac{f(z_0) - f^*}{SW_2\eta_l K} + \Psi_l \sigma_l^2 + \Psi_g \sigma_g^2$$

where $\Psi_l \triangleq \frac{\eta_l}{m} \Phi + \frac{15(n-m)K^2L^3\eta_l^3}{m(n-1)} \Phi + \frac{160}{17} \eta_l^2 L^2 K$, $\Psi_g \triangleq \frac{90(n-m)K^3L^3\eta_l^3}{m(n-1)} \Phi + \frac{3\eta_l(n-m)K}{m(n-1)} \Phi + \frac{960}{17} \eta_l^2 L^2 K^2$, and $\Phi \triangleq \frac{32T\bar{\eta}+32LT\bar{\eta}^2+32L^2W_1^2T\bar{\eta}}{17W_2}$.

Corollary 4.7 (Convergence Rate of Multistage FedGM). Suppose $S > 1$, i.e. the multistage regime, by setting $\bar{\eta} = \Theta(\sqrt{mK})$, $\hat{\eta}^2 = \Theta(mK)$, $\eta_l = \Theta\left(\frac{1}{\sqrt{TK}}\right)$, $W_2 = \Theta\left(\frac{T\sqrt{mK}}{S}\right)$ and $W_1^2 = \mathcal{O}\left(\sqrt{mK}\right)$, we have convergence rate as $\mathcal{O}\left(\sqrt{\frac{K}{Tm}}\right)$.

Remark 4.8. $\mathcal{O}\left(\sqrt{\frac{K}{Tm}}\right)$ recovers the best known convergence rate for FL nonconvex optimization (Yang, Fang, and Liu 2021). Similar to Remark 4.3, Theorem 4.6 shows an reliance on average learning rate, which explains why multistage scheme helps empirically. $\mathcal{O}\left(\sqrt{\frac{K}{Tm}}\right)$ indicates a slowdown effect from more local computation, which is supported by some existing works (Li et al. 2020b), while others observe a different effect of K (Lin et al. 2020). The exact impact of K on convergence warrants further investigation.

5 Momentum with System Heterogeneity

5.1 Autonomous Multistage FedGM

For a simplified abstraction of real world settings, most FL algorithms make the assumption that, all clients synchronize with the same global model and they conduct identical number of local updates at any given round. Though the assumption has been adopted in most existing works (McMahan

⁴In each round t , the server samples a subset of clients \mathcal{S}_t (suppose $|\mathcal{S}_t| = m < n$) uniformly at random without replacement, i.e. $\mathbb{P}\{i \in \mathcal{S}_t\} = \frac{m}{n}$ and $\mathbb{P}\{i, j \in \mathcal{S}_t\} = \frac{m(m-1)}{n(n-1)}$.

et al. 2017; Hsu, Qi, and Brown 2019; Li et al. 2020a; Karimireddy et al. 2020; Reddi et al. 2020; Bao et al. 2022; Wang, Lin, and Chen 2022), it rarely holds in reality.

In light of the limitations of existing works, we propose a general framework called Autonomous Multistage FedGM that enables the following three features, i.e. **heterogeneous local computing**, **asynchronous aggregation**, and **flexible client participation**, which is formalized in Algorithm 4.

Autonomous Multistage FedGM could effectively mitigate straggler effect and poor convergence issue in highly heterogeneous cross-device deployments. We leave a more detailed discussion of Algorithm 4 to Appendix B due to space limit.

Specifically, in Autonomous Multistage FedGM, the client decides when to participate in the training, and idling between rounds or even completely unavailable are both allowed. Once it decides to participate at round t , it retrieves current global model x_μ from the server and conduct $K_{t,i}$ local steps to update to $x_{\mu, K_{t,i}}^i$. Note in vanilla FedAvg, $K_{t,i} = K$ for any i and t . In contrast, we allow $K_{t,i}$ to be time-varying and device-dependent. The client then normalizes the model update by $K_{t,i}$ to avoid model biased towards clients with more local updates. Concurrently, the server collects the model updates from the clients. As every client may participate in training at a different round, the collected model update $\Delta_{t-\tau_{t,i}}^i$ may be from a historic timestamp, i.e. $\tau_{t,i}$ away from current time t . The server triggers global update whenever it collects m model updates and we denote the set of m responsive clients as \mathcal{S}_t . The global update is same as multistage FedGM (i.e. Lines 11-13). Note that server optimization is concurrent with clients, i.e., the global update happens whenever m model updates are collected, regardless of whether there are still some clients conducting local computation, thus ensuring there is no straggler.

Autonomous multistage FedGM, i.e. Algorithm 4, will recover multistage FedGM, i.e. Algorithm 3, if we set $K_{t,i} = K$ and $\tau_{t,i} = 0$ for $\forall t, i$. Please note that varying $K_{t,i}$ and nonzero $\tau_{t,i}$ bring nontrivial extra complexity to the theoretical analysis as can be seen in our proof.

5.2 Convergence Analysis

We state the convergence guarantee of autonomous multistage FedGM as follows,

Theorem 5.1. *We optimize $f(x)$ with Algorithm 4 under assumptions 1-3. Suppose the maximum delay is bounded, i.e. $\tau_{t,i} \leq \tau < \infty$ for any $i \in \mathcal{S}_t$ and $t \in \{0, 1, \dots, T-1\}$. Under the condition $\eta_l \leq \min \left\{ \frac{1}{8K_{t,\max}L}, \sqrt{\frac{1}{120L^2C_\eta\tau K_{t,\max}^2}} \right\}$, where $K_{t,\max} = \max_{i \in \mathcal{S}_t} K_{t,i}$. And further assume each client is included in \mathcal{S}_t with probability $\frac{m}{n}$ uniformly and independently. With necessary abbreviation for ease of notation⁵, we would have:*

⁵We denote $\bar{\eta} \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \eta_s$ (average server learning rate), $\hat{\eta}^2 \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^2$, $\hat{\eta}^3 \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^3$, $\frac{1}{K_t} = \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}}$, $\bar{K}_t \triangleq \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}$, $\hat{K}_t^2 \triangleq \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}^2$, $\phi_1 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \bar{K}_t$,

Algorithm 4: Autonomous Multistage FedGM

Input: Same as Algorithm 3

```

1 for  $s \in \{1, \dots, S\}$  do
2   for  $t$  in stage  $s$  do
3     At Each Client (Concurrently)
4     Once decided to participate in the training,
       retrieve  $x_\mu$  from the server and its
       timestamp, set  $x_{\mu,0}^i = x_\mu$ .
5     Select a number of local steps  $K_{t,i}$ , which is
       time-varying and device-dependent.
6      $\Delta_\mu^i = \text{LocalOPT}(i, \eta_l, K_{t,i}, x_\mu)$ 
7     Normalize and send  $\Delta_\mu^i = \frac{\Delta_\mu^i}{K_{t,i}}$ 
8     At Server (Concurrently)
9     Collect  $m$  local updates  $\{\Delta_{t-\tau_{t,i}}^i, i \in \mathcal{S}_t\}$ 
       returned from the clients to form set  $\mathcal{S}_t$ ,
       where  $\tau_{t,i}$  is the random delay of the client
        $i$ 's local update,  $i \in \mathcal{S}_t$ 
10    Aggregate  $\Delta_t = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \Delta_{t-\tau_{t,i}}^i$ 
11     $d_{t+1} = (1 - \beta_s)\Delta_t + \beta_s d_t$ 
12     $h_{t+1} = (1 - \nu_s)\Delta_t + \nu_s d_{t+1}$ 
13    Update  $x_{t+1} = x_t - \eta_s h_{t+1}$ 
14  end
15 end
16 return  $x_T$ 

```

$$\bar{G} \leq \frac{4(f(x_0) - f^*)}{SW_2\eta_l} + \Phi_l\sigma_l^2 + \Phi_g\sigma_g^2$$

$$\Phi_l \triangleq \frac{20\eta_l^2 L^2 T \bar{\eta}}{W_2} \phi_1 + \frac{4L^2 \tau^2 \hat{\eta}^3 \eta_l^2 T}{mW_2} \phi_3 + \frac{2L^2 W_2^2 \bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2\bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2L \hat{\eta}^2 \eta_l}{mW_2} \phi_3, \text{ and } \Phi_g \triangleq \frac{120\eta_l^2 L^2 T \bar{\eta} \phi_2}{W_2}.$$

Corollary 5.2 (Convergence Rate). *Suppose an identical K for all t and i . By appropriately setting $\bar{\eta}$, η_l , W_1 , W_2 , we have the convergence rate as, $\mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{K^2}{T}\right)$.*

Remark 5.3. Corollary 5.2 indicates τ brings a slowdown in convergence. Fortunately, with a sufficiently large T (e.g. $T \geq mK^5$) and a manageable τ (e.g. $\tau \leq \frac{T^{\frac{1}{4}}}{(mK)^{\frac{1}{4}}}$), autonomous multistage FedGM obtains a $\mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right)$ rate. Note that we make an additional assumption that each client is included in \mathcal{S}_t with probability $\frac{m}{n}$ uniformly and independently, which is necessary as the following Corollary 5.4 indicates if without such assumption, the rate has a non-convergent $\mathcal{O}(\sigma_g^2)$ term that we cannot avoid (the lower bound is $\Omega(\sigma_g^2)$).

Corollary 5.4 (Convergence Rate w/o Uniform Sampling Assumption). *Suppose an identical K for all t and i . By appropriately setting $\bar{\eta}$, η_l , W_1 , W_2 , we have the convergence*

$$\phi_2 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \hat{K}_t^2, \text{ and } \phi_3 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K_t}, \text{ for ease of notation.}$$

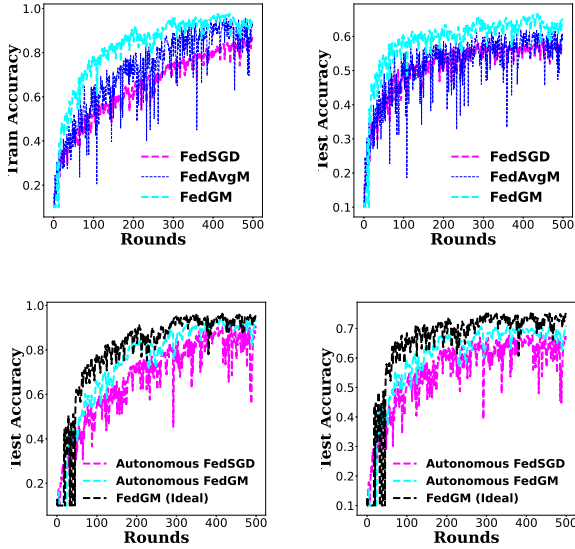


Figure 1: 1(a) Training and 1(b) Testing Curves for FedGM (ResNet on CIFAR-10). FedGM outperforms FedAvg/FedAvgM. 1(c) Training and 1(d) Testing for Autonomous FedGM (ResNet on CIFAR-10).

rate as, $\mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{K^2}{T}\right) + \mathcal{O}(\sigma_g^2)$, and the non-vanishing $\mathcal{O}(\sigma_g^2)$ is unavoidable.⁶

6 Experimental Results

In this section, we present empirical evidence to verify our theoretical findings. We train ResNet (He et al. 2016) and VGG (Simonyan and Zisserman 2015) on CIFAR10 (Krizhevsky 2009). To simulate data heterogeneity in CIFAR-10, we impose label imbalance across clients, i.e. each client is allocated a proportion of the samples of each label according to a Dirichlet distribution (Hsu, Qi, and Brown 2019; Yurochkin et al. 2019). The concentration parameter $\alpha > 0$ indicates the level of *non-i.i.d.*, with smaller α implies higher heterogeneity, and $\alpha \rightarrow \infty$ implies *i.i.d.* setting. Unless specified otherwise, we have 100 clients in all experiments, and the partial participation ratio is 0.05, i.e., 5 out of 100 clients are picked in each round, *non-i.i.d.* is $\alpha = 0.5$, and local epoch is 3. We defer many more results and details of hyperparameter settings to Appendix H.

6.1 Results on FedGM

Figure 1 shows the results for ResNet on CIFAR-10 with FedGM, FedAvgM, and FedAvg. We perform grid search over $\eta \in \{0.5, 1.0, 1.5, \dots, 5.0\}$, $\beta \in \{0.7, 0.9, 0.95\}$, and $\nu \in \{0.7, 0.9, 0.95\}$. We report their respective best results in Figure 1. We observe that though FedAvgM converges faster than FedAvg, it is only marginally better in terms of

⁶We informally state Corollary 5.4 due to page limit, please refer to Appendix G for a formal statement.

testing. FedGM, in contrast, outperforms FedAvgM and FedAvg in both measures. Therefore, a general momentum, instead of only SHB, is critical empirically. We analyze possible reasons and leave more results with VGG and different heterogeneity levels α to Appendix H.2.

6.2 Results on Multistage FedGM

Figure 2 shows the results for ResNet on CIFAR-10 with multistage vs. single-stage FedGM. The two black vertical lines at round 143 and 429 mark the end of 1st/2nd stage. For multistage FedGM, ($\eta_1 = 2.0, \eta_2 = 1.0, \eta_3 = 0.5$), the β also changes according to Eq. 3. From Figure 2, we observe multistage FedGM is better than single-stage FedGM, no matter what constant η it takes. Specifically, at first stage, $\eta_1 = 2.0$ makes the training curve fluctuate dramatically, but later into 2nd/3rd stage, the training stabilizes with smaller η_2 and η_3 . Multistage FedGM achieves a balance between early exploration and late exploitation. Multistage is also superior to its counterpart in testing. We leave more experiments to Appendix H.3.

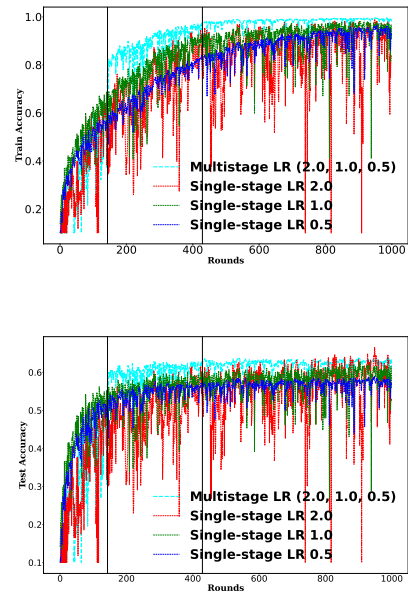


Figure 2: 2(a) Training and 2(b) Testing Curves for Multistage FedGM vs. Single-stage FedGM.

6.3 Results on Autonomous FedGM

Figure 1 shows the results for ResNet on CIFAR-10 with Autonomous FedGM (& FedAvg). Please refer to Appendix H.1 for detailed settings. We perform a grid search as in Section 6.1. We report their respective best curves. We plot an ideal FedGM (i.e. synchronous and identical local epochs) as reference line. We could observe Autonomous FedGM outperforms Autonomous FedAvg with system heterogeneity. Though Autonomous FedGM suffers a slowdown compared to the ideal FedGM, it is within a small margin, which

supports our theory in Corollary 5.2 and validates the effectiveness of Autonomous FedGM. We leave more experiments to Appendix H.4.

7 Conclusion

This paper systematically studied how the server momentum could help alleviate client drift that arises from both data heterogeneity and system heterogeneity. We demonstrated the critical role of momentum schemes and proper hyperparameter schedule by providing a rigorous convergence analysis and extensive empirical evidence, which pave a way for more widely and disciplined use of server momentum in the federated learning research community.

8 Acknowledgments

This work was partially supported by NSF 2217071, 2213700, 2106913, 2008208, 1955151 at UVA. This work was partially supported by NSF IIS 2347592, 2348169, 2348159, 2347604, CNS 2347617, CCF 2348306, DBI 2405416 at Pitt and UMD.

References

- An, W.; Wang, H.; Sun, Q.; Xu, J.; Dai, Q.; and Zhang, L. 2018. A PID Controller Approach for Stochastic Optimization of Deep Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8522–8531.
- Aydiukhin, D.; and Kasiviswanathan, S. 2021. Federated Learning under Arbitrary Communication Patterns. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 425–435. PMLR.
- Bao, R.; Gu, B.; and Huang, H. 2020. Fast OSCAR and OWL regression via safe screening rules. In *International Conference on Machine Learning*, 653–663. PMLR.
- Bao, R.; Wu, X.; Xian, W.; and Huang, H. 2022. Doubly sparse asynchronous learning for stochastic composite optimization. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, 1916–1922.
- Bao, W.; Wei, T.; Wang, H.; and He, J. 2023. Adaptive Test-Time Personalization for Federated Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Basu, D.; Data, D.; Karakus, C.; and Diggavi, S. 2019. *Qsparse-Local-SGD: Distributed SGD with Quantization, Sparsification, and Local Computations*. Red Hook, NY, USA: Curran Associates Inc.
- BUKATY, P. 2019. *The California Consumer Privacy Act (CCPA): An implementation guide*. IT Governance Publishing. ISBN 9781787781320.
- Chen, W.; Horváth, S.; and Richtárik, P. 2020. Optimal Client Sampling for Federated Learning. *ArXiv*, abs/2010.13723.
- Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; Anil, R.; Haque, Z.; Hong, L.; Jain, V.; Liu, X.; and Shah, H. 2016. Wide & Deep Learning for Recommender Systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, DLRS 2016*, 7–10. New York, NY, USA: Association for Computing Machinery. ISBN 9781450347952.
- Cutkosky, A.; and Mehta, H. 2020. Momentum Improves Normalized SGD. In *International Conference on Machine Learning*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, abs/1810.04805.
- 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. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

- European Commission. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance).
- Fallah, A.; Mokhtari, A.; and Ozdaglar, A. 2020. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*.
- Ge, R.; Kakade, S. M.; Kidambi, R.; and Netrapalli, P. 2019a. The Step Decay Schedule: A Near Optimal, Geometrically Decaying Learning Rate Procedure. In *NeurIPS*.
- Ge, R.; Kakade, S. M.; Kidambi, R.; and Netrapalli, P. 2019b. *The Step Decay Schedule: A near Optimal, Geometrically Decaying Learning Rate Procedure for Least Squares*. Red Hook, NY, USA: Curran Associates Inc.
- Goetz, J.; Malik, K.; Bui, D. V.; Moon, S.; Liu, H.; and Kumar, A. 2019. Active Federated Learning. *ArXiv*, abs/1909.12641.
- Goyal, P.; Dollár, P.; Girshick, R. B.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; Tulloch, A.; Jia, Y.; and He, K. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *CoRR*, abs/1706.02677.
- Gu, X.; Huang, K.; Zhang, J.; and Huang, L. 2021. Fast Federated Learning in the Presence of Arbitrary Device Unavailability. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.
- Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive Representation Learning on Large Graphs. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- He, F.; Liu, T.; and Tao, D. 2019. Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence. In *Advances in Neural Information Processing Systems 32*, 1143–1152. Curran Associates, Inc.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.
- Hsu, T.-M. H.; Qi; and Brown, M. 2019. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. *ArXiv*, abs/1909.06335.
- Hu, Z.; Wu, X.; and Huang, H. 2023. Beyond Lipschitz smoothness: a tighter analysis for nonconvex optimization. In *International Conference on Machine Learning*, 13652–13678. PMLR.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.
- Jee Cho, Y.; Wang, J.; and Joshi, G. 2022. Towards Understanding Biased Client Selection in Federated Learning. In Camps-Valls, G.; Ruiz, F. J. R.; and Valera, I., eds., *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 10351–10375. PMLR.
- Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K. A.; Charles, Z.; Cormode, G.; Cummings, R.; D’Oliveira, R. G. L.; Eichner, H.; Rouayheb, S. E.; Evans, D.; Gardner, J.; Garrett, Z.; Gascón, A.; Ghazi, B.; Gibbons, P. B.; Gruteser, M.; Harchaoui, Z.; He, C.; He, L.; Huo, Z.; Hutchinson, B.; Hsu, J.; Jaggi, M.; Javidi, T.; Joshi, G.; Khodak, M.; Konečný, J.; Korolova, A.; Koushanfar, F.; Koyejo, S.; Lepoint, T.; Liu, Y.; Mittal, P.; Mohri, M.; Nock, R.; Özgür, A.; Pagh, R.; Qi, H.; Ramage, D.; Raskar, R.; Raykova, M.; Song, D.; Song, W.; Stich, S. U.; Sun, Z.; Suresh, A. T.; Tramèr, F.; Vepakomma, P.; Wang, J.; Xiong, L.; Xu, Z.; Yang, Q.; Yu, F. X.; Yu, H.; and Zhao, S. 2021. Advances and Open Problems in Federated Learning. *Found. Trends Mach. Learn.*, 14(1-2): 1–210.
- Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 5132–5143. PMLR.
- Khanduri, P.; Sharma, P.; Yang, H.; Hong, M.; Liu, J.; Rajawat, K.; and Varshney, P. 2021. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34.
- Kidambi, R.; Netrapalli, P.; Jain, P.; and Kakade, S. M. 2018. On the insufficiency of existing momentum schemes for Stochastic Optimization. *CoRR*, abs/1803.05591.
- Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. 1097–1105.
- Lessard, L.; Recht, B.; and Packard, A. 2014. Analysis and Design of Optimization Algorithms via Integral Quadratic Constraints. *SIAM Journal on Optimization*, 26.
- Li, Q.; Diao, Y.; Chen, Q.; and He, B. 2022. Federated Learning on Non-IID Data Silos: An Experimental Study. In *IEEE International Conference on Data Engineering*.
- Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020a. Federated Optimization in Heterogeneous Networks. In Dhillon, I.; Papailiopoulos, D.; and Sze, V., eds., *Proceedings of Machine Learning and Systems*, volume 2, 429–450.
- Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020b. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.
- Lian, X.; Huang, Y.; Li, Y.; and Liu, J. 2015. Asynchronous Parallel Stochastic Gradient for Nonconvex Optimization. In

- Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, 2737–2745. Cambridge, MA, USA: MIT Press.
- Lin, T.; Stich, S. U.; Patel, K. K.; and Jaggi, M. 2020. Don't Use Large Mini-batches, Use Local SGD. In *ICLR - International Conference on Learning Representations*.
- Liu, Y.; Gao, Y.; and Yin, W. 2020. An Improved Analysis of Stochastic Gradient Descent with Momentum. *arXiv:2007.07989*.
- Ma, J.; and Yarats, D. 2019. Quasi-hyperbolic momentum and Adam for deep learning. In *International Conference on Learning Representations*.
- McMahan, H. B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics*.
- Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. A. 2013. Playing Atari with Deep Reinforcement Learning. *ArXiv*, abs/1312.5602.
- Nguyen, J.; Malik, K.; Zhan, H.; Yousefpour, A.; Rabbat, M. G.; Malek, M.; and Huba, D. 2021. Federated Learning with Buffered Asynchronous Aggregation. In *International Conference on Artificial Intelligence and Statistics*.
- Nishio, T.; and Yonetani, R. 2018. Client Selection for Federated Learning with Heterogeneous Resources in Mobile Edge. *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 1–7.
- Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečný, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.
- Reddi, S. J.; Kale, S.; and Kumar, S. 2018. On the Convergence of Adam and Beyond. In *International Conference on Learning Representations*.
- Ribero, M.; and Vikalo, H. 2020. Communication-Efficient Federated Learning via Optimal Client Sampling. *ArXiv*, abs/2007.15197.
- Rothchild, D.; Panda, A.; Ullah, E.; Ivkin, N.; Stoica, I.; Braverman, V.; Gonzalez, J.; and Arora, R. 2020. FetchSGD: Communication-Efficient Federated Learning with Sketching. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.
- Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Smith, L. N. 2017. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464–472.
- Smith, S.; and Le, Q. V. 2018. A Bayesian Perspective on Generalization and Stochastic Gradient Descent.
- Smith, S. L.; Kindermans, P.-J.; and Le, Q. V. 2018. Don't Decay the Learning Rate, Increase the Batch Size. In *International Conference on Learning Representations*.
- Sun, J.; Huai, M.; Jha, K.; and Zhang, A. 2022. Demystify Hyperparameters for Stochastic Optimization with Transferable Representations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 1706–1716. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850.
- Sun, J.; Sinha, S.; and Zhang, A. 2023. Enhance Diffusion to Improve Robust Generalization. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, 2083–2095. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701030.
- Sun, J.; Yang, Y.; Xun, G.; and Zhang, A. 2021. A Stagewise Hyperparameter Scheduler to Improve Generalization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, 1530–1540. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325.
- Sun, J.; Yang, Y.; Xun, G.; and Zhang, A. 2023. Scheduling Hyperparameters to Improve Generalization: From Centralized SGD to Asynchronous SGD. *ACM Trans. Knowl. Discov. Data*, 17(2).
- Suo, Q.; Yao, L.; Xun, G.; Sun, J.; and Zhang, A. 2019. Recurrent Imputation for Multivariate Time Series with Missing Values. In *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019, Xi'an, China, June 10-13, 2019*, 1–3. IEEE.
- Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the Importance of Initialization and Momentum in Deep Learning. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, III–1139–III–1147.
- Van Scoy, B.; Freeman, R. A.; and Lynch, K. M. 2018. The Fastest Known Globally Convergent First-Order Method for Minimizing Strongly Convex Functions. *IEEE Control Systems Letters*, 2(1): 49–54.
- Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713829546.
- Wang, S.; and Ji, M. 2022. A Unified Analysis of Federated Learning with Arbitrary Client Participation. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.
- Wang, X.; Magnússon, S.; and Johansson, M. 2021. On the Convergence of Step Decay Step-Size for Stochastic Optimization. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.
- Wang, Y.; Lin, L.; and Chen, J. 2022. Communication-Efficient Adaptive Federated Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 22802–22838. PMLR.

Wilson, A. C.; Roelofs, R.; Stern, M.; Srebro, N.; and Recht, B. 2017. The Marginal Value of Adaptive Gradient Methods in Machine Learning. In *Advances in Neural Information Processing Systems 30*, 4148–4158. Curran Associates, Inc.

Wu, X.; Huang, F.; Hu, Z.; and Huang, H. 2023a. Faster adaptive federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10379–10387.

Wu, X.; Sun, J.; Hu, Z.; Li, J.; Zhang, A.; and Huang, H. 2023b. Federated Conditional Stochastic Optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wu, X.; Sun, J.; Hu, Z.; Zhang, A.; and Huang, H. 2023c. Solving a Class of Non-Convex Minimax Optimization in Federated Learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wu, Y.; Hu, Z.; Zhang, H.; and Huang, H. 2023d. DiPmark: A Stealthy, Efficient and Resilient Watermark for Large Language Models. *ArXiv*, abs/2310.07710.

Xie, C.; Koyejo, O.; and Gupta, I. 2019. Asynchronous Federated Optimization. *ArXiv*, abs/1903.03934.

Xun, G.; Jha, K.; Sun, J.; and Zhang, A. 2020. Correlation Networks for Extreme Multi-label Text Classification. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Yan, Y.; Niu, C.; Ding, Y.; Zheng, Z.; Wu, F.; Chen, G.; Tang, S.; and Wu, Z. 2020. Distributed Non-Convex Optimization with Sublinear Speedup under Intermittent Client Availability. *ArXiv*, abs/2002.07399.

Yang, H.; Fang, M.; and Liu, J. 2021. Achieving Linear Speedup with Partial Worker Participation in Non-IID Federated Learning. In *International Conference on Learning Representations*.

Yang, H.; Zhang, X.; Khanduri, P.; and Liu, J. 2021. Anarchic Federated Learning. In *International Conference on Machine Learning*.

Yurochkin, M.; Agarwal, M.; Ghosh, S. S.; Greenewald, K. H.; Hoang, T. N.; and Khazaeni, Y. 2019. Bayesian Non-parametric Federated Learning of Neural Networks. In *International Conference on Machine Learning*.

Zhang, J.; Karimireddy, S. P.; Veit, A.; Kim, S.; Reddi, S.; Kumar, S.; and Sra, S. 2020. Why are Adaptive Methods Good for Attention Models? In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 15383–15393. Curran Associates, Inc.

Zhang, S.; Choromańska, A.; and LeCun, Y. 2014. Deep learning with Elastic Averaging SGD. In *NIPS*.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.

Zheng, S.; Meng, Q.; Wang, T.; Chen, W.; Yu, N.; Ma, Z.-M.; and Liu, T.-Y. 2017. Asynchronous Stochastic Gradient Descent with Delay Compensation. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17*, 4120–4129. JMLR.org.

Appendix

A Organization of Appendix

Appendix is organized as follows. In Section B, we discuss the omitted details of server momentum with system heterogeneity from Section 5. In Section C, we provide the discussion of related works. In Section D, we show the proof of Theorem 4.1, Corollary 4.2, and Corollary 4.4. In Section E, We provide the proof of Theorem 4.6 and Corollary 4.7. In Section F, we provide the proof of Theorem 5.1 and Corollary 5.2. In Section G, we provide the proof of Corollary 5.4. Note that we use a 0-indexing for T and S in most proofs, i.e. the rounds (stages) are denoted as $\{0, \dots, T - 1\}$ ($\{0, \dots, S - 1\}$), which is equivalent to the 1-indexing in main text, i.e. $\{1, \dots, T\}$ ($\{1, \dots, S\}$). Finally, in Section H, we provide experimental settings and extra experimental results that are omitted from main text.

B Autonomous Multistage FedGM

In this section, we discuss the omitted details of server momentum with system heterogeneity from Section 5.

B.1 Ubiquitous System Heterogeneity

For a simplified abstraction of real world settings, most FL algorithms make the assumption that, all clients initialize with the same global model and they conduct identical number of local updates at any given round.

More formally, we could observe from LocalOPT (Algorithm 2) that the following assumptions have been made, (a) *Homogeneous Local Updates* all participating clients would do local gradient descent for K steps; (b) *Uniform Client Participation* each client would participate in a given communication round uniformly according to a given distribution that is independent across rounds; (c) *Synchronous Local Clients* all participating clients always initialize at x_t , i.e., the global model at current timestamp.

Though these three assumptions have been adopted in most existing works (McMahan et al. 2017; Hsu, Qi, and Brown 2019; Li et al. 2020a; Karimireddy et al. 2020; Reddi et al. 2020; Wang, Lin, and Chen 2022; Hu, Wu, and Huang 2023), each of these assumptions rarely holds in reality. Due to unavoidable **heterogeneous client capability**, and **unpredictable availability**, enforcing identical local epochs and synchrony would incur straggler effect and unnecessary energy waste (Kairouz et al. 2021). Therefore, realistic FL system is more economical to allow different local epochs and **asynchronous aggregation**.

When studying client heterogeneity and the resulting client drift, most works focus explicitly on data heterogeneity (Li et al. 2020b; Yang, Fang, and Liu 2021), while ignoring the equally ubiquitous system heterogeneity, which casts doubt on the applicability of the corresponding algorithms in practice.

B.2 Autonomous Multistage FedGM

In light of the limitations of existing works, we aim to propose a general framework that enables all three features, i.e. **heterogeneous local computing**, **asynchronous aggregation**, and **flexible client participation**, which is formalized in Algorithm 4.

Specifically, in Autonomous Multistage FedGM, the client decides when to participate in the training, and idling between rounds or even completely unavailable are both allowed. Once it decides to participate at round t , it retrieves current global model x_μ from the server and initializes $x_{\mu,0}^i = x_\mu$ locally, and conduct $K_{t,i}$ local steps to update from $x_{\mu,0}^i$ to $x_{\mu,K_{t,i}}^i$. Note in vanilla FedAvg, $K_{t,i} = K$ for any i and t . In contrast, we allow $K_{t,i}$ to be time-varying and device-dependent. The client then normalizes the model update by $K_{t,i}$, i.e. $\Delta_\mu^i = \frac{x_{\mu,0}^i - x_{\mu,K_{t,i}}^i}{K_{t,i}}$, to avoid model biased towards clients with more local updates. Concurrently, the server collects the model updates from the clients. As every client may participate in training at a different round, the collected model update $\Delta_{t-\tau_{t,i}}^i$ may be from a historic timestamp, i.e. $\tau_{t,i}$ away from current time t . If we set the random delay $\tau_{t,i} = 0$, it would be ordinary synchronous aggregation. The server triggers global update whenever it collects m model updates and we denote the set of m responsive clients as \mathcal{S}_t . The global update is same as multistage FedGM (i.e. Lines 11-13). Note that server optimization is concurrent with clients, i.e., the global update happens whenever m model updates are collected, regardless of whether there are still some clients conducting local computation, thus ensuring there is no straggler.

Autonomous multistage FedGM will recover multistage FedGM, i.e. Algorithm 3, if we set $K_{t,i} = K$ and $\tau_{t,i} = 0$ for $\forall t, i$. Please note that varying $K_{t,i}$ and nonzero $\tau_{t,i}$ bring nontrivial extra complexity to the theoretical analysis as can be seen in our proof.

C Related Work

C.1 Tackling Client Heterogeneity and Client Drift in FedAvg

Deep learning models have been widely applied in many different domains, e.g. (Mnih et al. 2013; He et al. 2016; Devlin et al. 2019; Hamilton, Ying, and Leskovec 2017; Cheng et al. 2016; Suo et al. 2019; Xun et al. 2020; Wu et al. 2023d), mostly in centralized environment. Due to privacy concerns and regulatory requirements (European Commission 2016; BUKATY 2019), Federated Averaging (FedAvg) (McMahan et al. 2017) has been applied to avoid data transmission in collaborative training of

deep learning models in a wide range of settings (Li et al. 2020a; Rothchild et al. 2020; Wang et al. 2020; Fallah, Mokhtari, and Ozdaglar 2020; Li, He, and Song 2021; Bao et al. 2023; Wu et al. 2023b,c).

Client heterogeneity and its resulting client drift is known to destabilize FedAvg convergence (Zhao et al. 2018; Karimireddy et al. 2020). FedProx (Li et al. 2020a) proposes to regularize the difference against global model in local objective. SCAFFOLD (Karimireddy et al. 2020) leverages variance reduction technique to reduce client drift and achieves the best known $\mathcal{O}\left(\frac{1}{\sqrt{nKT}}\right)$ rate in full participation setting. However, SCAFFOLD is not stateless which restricts its application in cross-device FL. (Reddi et al. 2020; Wang, Lin, and Chen 2022) propose a family of federated adaptive optimizers e.g. FedADAM, FedADAGRAD, and FedAMS that are natural extensions from non-FL adaptive optimizers to FL settings. Recent works propose algorithms to alleviate client heterogeneity in FL bilevel optimization problems, e.g. minimax (Wu et al. 2023c) or conditional stochastic optimization (Wu et al. 2023b). The most relevant line of research to this paper is on server-side momentum. Server-side momentum is first empirically studied in (Hsu, Qi, and Brown 2019), where FedAvgM is observed to outperform FedAvg in *non-i.i.d.* settings by a significant margin. Recent explorations include (Rothchild et al. 2020) that studies the interplay between server momentum and compression, and (Khanduri et al. 2021) which studies a two-sided momentum scheme that allows both server momentum and client momentum. However, all existing works have the following limitations that our paper aims to address, (a) they do not provide a unified analysis for a family of momentum schemes; (b) they do not incorporate any realistic hyperparameter schedulers; (c) they ignore an important source of client heterogeneity, i.e. system heterogeneity.

C.2 Hyperparameter Scheduling

Adaptively adjusting hyperparameters throughout the training is key to the success of deep model training, but most of such explorations are in non-FL context. For example, previous works (He, Liu, and Tao 2019), (Sun et al. 2022), and (Sun, Sinha, and Zhang 2023) reveal the connection between hyperparameters and generalization capacity of optimizers. (Krizhevsky, Sutskever, and Hinton 2012) and (He et al. 2016) propose to decay learning rate η whenever the loss saturates; (Goyal et al. 2017) popularizes the heuristic of warmup to increase η from a small value to a very large value in the first few iterations; (Smith 2017) proposes to adopt a cyclic learning rate schedule between warmup and decay phases. Apart from learning rate, adaptively scheduling other hyperparameters (e.g. momentum factor and batch size) is also shown to be very effective in many settings. For example, (Sutskever et al. 2013) shows a slowly increasing schedule for the momentum factor is crucial; (Smith and Le 2018; Smith, Kindermans, and Le 2018) propose a procedure to enable large batch training where η and momentum factor β are increasing, and batch size B is scaled $B \propto \frac{\eta}{1-\beta}$. From a theoretical point of view, in non-FL context, (Ge et al. 2019a) and (Wang, Magnússon, and Johansson 2021) show the multistage learning rate scheduler achieves a near-optimal convergence rate of $\mathcal{O}\left(\frac{\log T}{T}\right)$ rate (faster than polynomial decay), in both convex and non-convex functions. (Sun et al. 2021) and (Sun et al. 2023) show the convergence of multistage scheduler for momentum schemes. However, to our best knowledge, there is no prior work studying multistage hyperparameter scheduler (both learning rate and momentum factor) in FL settings.

C.3 Flexible Participation and Asynchronous Aggregation

The existing works that are dedicated to studying system heterogeneity can be mainly categorized into the following groups,

- Heterogeneous local computing but synchronous aggregation. (Wang et al. 2020) is probably the first work to show heterogeneous number of local updates results in the global model converges to a mismatched optimum which can be arbitrarily away from the true optimum and proposes an effective remedy FedNova to correct the mismatch. (Basu et al. 2019) considers how model compression works with different number of local updates ⁷. (Avdiukhin and Kasiviswanathan 2021) focuses on a similar setting and studies the upper bound of distances between two consecutive communications to reduce communication times as much as possible. Their theoretical analysis relies on bounded gradient assumption.
- Asynchronous aggregation. This line of research is most closely related to our proposed Autonomous Multistage FedGM. However, though asynchrony has been a decade-long topic in traditional distributed computing (Zhang, Choromańska, and LeCun 2014; Lian et al. 2015; Zheng et al. 2017), it has received very limited attention in federated learning. (Xie, Koyejo, and Gupta 2019) proposes FedAsync in which the server immediately updates the global model whenever it receives a single local model. Their theoretical analysis only applies to convex objective function which is not applicable to deep learning. Moreover, this proposal has negative implication in privacy, as it no longer hides one single update in an aggregate, which is one of the most important points to use FL in the first place. In light of this, (Nguyen et al. 2021) proposes FedBuff, in which a global update is triggered when the server receives m local updates, where m is a pre-specified hyperparameter. By maintaining a size m buffer, FedBuff could secure the identity of each local update and is empirically faster than FedAsync. However, it does not consider the heterogeneous local computing and does not provide a convergence rate that shows the dependency on m . (Yang et al. 2021) proposes anarchic FL in which the clients are free to determine how much local computation to conduct

⁷Though the authors call the proposed model 'asynchronous', the asynchrony refers to that updates occur after different number of local iterations but the local iterations are synchronous with respect to the global clock. However, 'asynchronous' in our context refers to local iterations are asynchronous with respect to the global clock, which is more challenging to analyze.

and the asynchronous communication is in the same fashion as FedBuff. However, (Yang et al. 2021) only considers the case of vanilla FedAvg, while our works subsumes anarchic FL as a special case.

There are many other works that enable flexible participation scheme but still synchronous aggregation, e.g. (Yan et al. 2020; Gu et al. 2021; Wang and Ji 2022; Nishio and Yonetani 2018; Chen, Horváth, and Richtárik 2020; Jee Cho, Wang, and Joshi 2022; Goetz et al. 2019; Ribero and Vikalo 2020). This line of research is less related to our proposed research.

D Proof of Theorem 4.1, Corollary 4.2, and Corollary 4.4

Proof of Multistage FedGM with Full Participation. Recall the formulation of General Momentum:

$$\begin{aligned} d_{t+1} &= (1 - \beta_t) \Delta_t + \beta_t d_t \\ x_{t+1} &= x_t - \eta_t [(1 - \nu_t) \Delta_t + \nu_t d_{t+1}] \end{aligned}$$

Denote the update sequence $y_t \triangleq x_{t+1} - x_t$. The updating rule is different from FedAvg in that $y_t \neq -\eta_t \Delta_t$. The proof hinges on the construction of an auxiliary sequence $\{z_t\}_{t=0}^T$, such that $z_{t+1} - z_t = -\eta_t \Delta_t$. This $\{z_t\}_{t=0}^T$ is more like vanilla FedAvg iterates and thus easier to deal with. We then study the property of $\{z_t\}_{t=0}^T$ and its connection to $\{x_t\}_{t=0}^T$. $\{z_t\}_{t=0}^T$ is devised as follows:

$$z_t = x_t - \frac{\eta_t \beta_t \nu_t}{1 - \beta_t} d_t \quad (4)$$

where $d_0 = 0$.

We now verify $z_{t+1} - z_t = -\eta_t \Delta_t$,

$$\begin{aligned} z_{t+1} - z_t &= x_{t+1} - \frac{\eta_{t+1} \beta_{t+1} \nu_{t+1}}{1 - \beta_{t+1}} d_{t+1} - x_t + \frac{\eta_t \beta_t \nu_t}{1 - \beta_t} d_t \\ &\stackrel{(i)}{=} -\eta_t y_t - W_1 (d_{t+1} - d_t) \\ &\stackrel{(ii)}{=} -\eta_t ((1 - \nu_t) \Delta_t + \nu_t d_{t+1}) - W_1 ((1 - \beta_t) \Delta_t + \beta_t d_t - d_t) \\ &= -\eta_t (1 - \nu_t) \Delta_t - \eta_t \beta_t \nu_t \Delta_t - \eta_t \nu_t (d_{t+1} - \beta_t d_t) \\ &= -\eta_t (1 - \nu_t) \Delta_t - \eta_t \beta_t \nu_t \Delta_t - \eta_t \nu_t (1 - \beta_t) \Delta_t = -\eta_t \Delta_t \end{aligned}$$

where (i) holds by the assumption $\frac{\eta_t \beta_t \nu_t}{1 - \beta_t}$ is a constant W_1 , (ii) holds by plugging in the updating rule for d_t and x_t . Since f is L -smooth, taking conditional expectation with respect to all randomness prior to step t , we have

$$\begin{aligned} \mathbb{E}[f(z_{t+1})] &\leq f(z_t) + \mathbb{E}[\langle \nabla f(z_t), z_{t+1} - z_t \rangle] + \frac{L}{2} \mathbb{E}[\|z_{t+1} - z_t\|^2] \\ &\leq f(z_t) + \mathbb{E}[\langle \nabla f(z_t), -\eta_t \Delta_t \rangle] + \frac{L}{2} \eta_t^2 \mathbb{E}[\|\Delta_t\|^2] \\ &\leq f(z_t) + \underbrace{\mathbb{E}[\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle]}_{A_1} + \underbrace{\mathbb{E}[\langle \nabla f(x_t), -\eta_t \Delta_t \rangle]}_{A_2} + \underbrace{\frac{L}{2} \eta_t^2 \mathbb{E}[\|\Delta_t\|^2]}_{A_3} \end{aligned}$$

Bounding A_1 :

$$\begin{aligned} A_1 &= \mathbb{E}[\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle] \\ &\stackrel{(i)}{\leq} \mathbb{E}[\|\sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t))\| \cdot \|\sqrt{\eta_t} \Delta_t\|] \\ &\stackrel{(ii)}{\leq} \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E}[\|d_t\|^2] + \frac{1}{2} \eta_t \mathbb{E}[\|\Delta_t\|^2] \end{aligned}$$

where (i) holds by applying Cauchy-Schwarz inequality, and (ii) follows by invoking the definition of z_t , Young's inequality and f is L -smooth.

Bounding A_2 :

$$\begin{aligned} A_2 &= \mathbb{E}[\langle \nabla f(x_t), -\eta_t \Delta_t \rangle] \\ &= \eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \Delta_t - \eta_t K \nabla f(x_t) \rangle] \\ &\stackrel{(i)}{=} -\eta_t \eta_t K \mathbb{E}[\|\nabla f(x_t)\|^2] + \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t g_{t,k}^i \right\rangle \right] \end{aligned}$$

where (i) follows from the definition $\Delta_t = \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t g_{t,k}^i$.

where we further bound $\eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t g_{t,k}^i \right\rangle \right]$,

$$\begin{aligned}
& \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t g_{t,k}^i \right\rangle \right] \\
& \stackrel{(i)}{=} \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t \nabla f_i(x_{t,k}^i) \right\rangle \right] \\
& \leq \eta_t \mathbb{E} \left[\left\langle \sqrt{\eta_t K} \nabla f(x_t), \frac{\sqrt{\eta_t K}}{Kn} \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right\rangle \right] \\
& \stackrel{(ii)}{\leq} \frac{\eta_t \eta_t K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{\eta_t \eta_t K}{2K^2 n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right\|^2 \right] \\
& - \frac{\eta_t}{2} \mathbb{E} \left[\left\| \sqrt{\eta_t K} (\nabla f(x_t) - \frac{1}{Kn} \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i))) \right\|^2 \right] = \frac{\eta_t \eta_t K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\
& + \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned} \tag{5}$$

where (i) holds as we take conditional expectation with respect to all randomness prior to step t and $\nabla f(x_t) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_t)$ by definition, (ii) holds as $\langle a, b \rangle = \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|^2 - \frac{1}{2} \|a - b\|^2$.

We further bound Equation 5,

$$\begin{aligned}
& \stackrel{(i)}{\leq} \frac{\eta_t \eta_t K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{\eta_t \eta_t}{2n} \sum_{i=1}^n \sum_{k=0}^{K-1} \mathbb{E} \left[\|\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& \stackrel{(ii)}{\leq} \frac{\eta_t \eta_t K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{\eta_t \eta_t L^2}{2n} \sum_{i=1}^n \sum_{k=0}^{K-1} \mathbb{E} \left[\|x_t - x_{t,k}^i\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

where (i) holds as $\|\sum_{i=1}^n x_i\|^2 \leq n \sum_{i=1}^n \|x_i\|^2$, and (ii) holds due to L -smoothness of f_i .

When $\eta_t \leq \frac{1}{8KL}$, for any k , we have the following from (Reddi et al. 2020),

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|x_t - x_{t,k}^i\|^2 \right] \leq 5K\eta_t^2 (\sigma_l^2 + 6K\sigma_g^2) + 30K^2\eta_t^2 \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right]$$

Thus, we have the following,

$$\begin{aligned}
& \frac{\eta_t \eta_t K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{\eta_t \eta_t L^2}{2n} \sum_{i=1}^n \sum_{k=0}^{K-1} \mathbb{E} \left[\|x_t - x_{t,k}^i\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& \leq \frac{\eta_t \eta_t K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& + \frac{\eta_t \eta_t L^2 K}{2} \left(5K\eta_t^2 (\sigma_l^2 + 6K\sigma_g^2) + 30K^2\eta_t^2 \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \right) \\
& \leq \left(\frac{\eta_t \eta_t K}{2} + 15\eta_t \eta_t^3 K^3 L^2 \right) \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{5}{2} \eta_t \eta_t^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2) \\
& \stackrel{(i)}{\leq} \frac{47}{64} \eta_t \eta_t K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{5}{2} \eta_t \eta_t^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2)
\end{aligned}$$

where (i) holds as $\eta_l \leq \frac{1}{8KL}$.

Merging all pieces together, we have the bound for A_2 ,

$$\begin{aligned} A_2 &= -\eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_l K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_l g_{t,k}^i \right\rangle \right] \\ &\leq -\frac{17}{64} \eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2) - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \end{aligned}$$

Bounding $\mathbb{E} \left[\|\Delta_t\|^2 \right]$:

$$\begin{aligned} \mathbb{E} \left[\|\Delta_t\|^2 \right] &= \mathbb{E} \left[\left\| \frac{\eta_l}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} g_{t,k}^i \right\|^2 \right] \\ &\stackrel{(i)}{=} \mathbb{E} \left[\left\| \frac{\eta_l}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} (g_{t,k}^i - \nabla f_i(x_{t,k}^i)) \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{\eta_l}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\ &\stackrel{(ii)}{=} \frac{\eta_l^2}{n^2} \sum_{i=1}^n \sum_{k=0}^{K-1} \mathbb{E} \left[\|g_{t,k}^i - \nabla f_i(x_{t,k}^i)\|^2 \right] + \mathbb{E} \left[\left\| \frac{\eta_l}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\ &\stackrel{(iii)}{\leq} \frac{K\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \end{aligned}$$

where (i) and (ii) hold as $\mathbb{E} \left[\left\| \sum_{i=1}^n x_i \right\|^2 \right] = \sum_{i=1}^n \mathbb{E} \left[\|x_i\|^2 \right]$ when $\mathbb{E} [x_i] = 0$, and we know $\mathbb{E} \left[g_{t,k}^i - \nabla f_i(x_{t,k}^i) \right] = 0$. (iii) holds due to bounded local variance assumption.

Bounding $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$:

It is straightforward to verify:

$$d_t = \sum_{p=0}^t a_{t,p} \Delta_p, \quad \text{where} \quad a_{t,p} = (1 - \beta_p) \prod_{q=p+1}^t \beta_q$$

With $d_t = \sum_{p=0}^t a_{t,p} \Delta_p$, we could get,

$$\begin{aligned} \mathbb{E} \left[\|d_t\|^2 \right] &= \mathbb{E} \left[\left\| \sum_{p=0}^t a_{t,p} \Delta_p \right\|^2 \right] \\ &= \sum_{e=1}^d \mathbb{E} \left[\left(\sum_{p=0}^t a_{t,p} \Delta_{p,e} \right)^2 \right] \stackrel{(i)}{\leq} \sum_{e=1}^d \mathbb{E} \left[\left(\sum_{p=0}^t a_{t,p} \right) \cdot \left(\sum_{p=0}^t a_{t,p} \Delta_{p,e}^2 \right) \right] \\ &\stackrel{(ii)}{\leq} \left(1 - \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t a_{t,p} \mathbb{E} \left[\|\Delta_p\|^2 \right] \stackrel{(iii)}{\leq} \frac{K\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \sum_{p=0}^t a_{t,p} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{p,k}^i) \right\|^2 \right] \end{aligned}$$

where $\Delta_{p,e}$ denotes the e -th element of vector Δ_p . (i) holds due to Cauchy–Schwarz inequality, (ii) holds as $\sum_{p=0}^t a_{t,p} = 1 - \prod_{q=1}^t \beta_q$, (iii) holds by plugging in the bound for $\mathbb{E} \left[\|\Delta_t\|^2 \right]$ and $\beta_q < 1$.

We sum over $t \in \{0, \dots, T-1\}$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] &\leq \frac{TK\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \sum_{t=0}^{T-1} \sum_{p=0}^t a_{t,p} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{p,k}^i) \right\|^2 \right] \\
&= \frac{TK\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \sum_{p=0}^{T-1} \left(\sum_{t=p}^{T-1} a_{t,p} \right) \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{p,k}^i) \right\|^2 \right]
\end{aligned}$$

Since $\{\beta_t\}_{t=0}^{T-1}$ is a non-decreasing sequence, we could verify $\sum_{t=p}^{T-1} a_{t,p} \leq \frac{1-\beta_0}{1-\beta_S} = C_\beta$.

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] &\leq \frac{TK\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \sum_{p=0}^{T-1} \left(\sum_{t=p}^{T-1} a_{t,p} \right) \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{p,k}^i) \right\|^2 \right] \\
&\leq \frac{TK\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

Bounding A_3 :

$$\begin{aligned}
A_3 &= \frac{L}{2} \eta_t^2 \mathbb{E} \left[\|\Delta_t\|^2 \right] \\
&\stackrel{(i)}{\leq} \frac{L}{2} \eta_t^2 \left(\frac{K\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \right) \\
&\leq \frac{LK\eta_l^2 \eta_t^2}{2n} \sigma_l^2 + \frac{L\eta_t^2 \eta_l^2}{2n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

where (i) holds by plugging in the bound for $\mathbb{E} \left[\|\Delta_t\|^2 \right]$.

Merging A_1, A_2, A_3 together,

$$\begin{aligned}
\mathbb{E} [f(z_{t+1})] - f(z_t) &\leq \underbrace{\mathbb{E} [\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle]}_{A_1} + \underbrace{\mathbb{E} [\langle \nabla f(x_t), -\eta_t \Delta_t \rangle]}_{A_2} + \underbrace{\frac{L}{2} \eta_t^2 \mathbb{E} \left[\|\Delta_t\|^2 \right]}_{A_3} \\
&\leq \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1-\beta_t} \right)^2 \mathbb{E} \left[\|d_t\|^2 \right] + \frac{1}{2} \eta_t \left(\frac{K\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \right) \\
&\quad - \frac{17}{64} \eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2) - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
&\quad + \frac{LK\eta_l^2 \eta_t^2}{2n} \sigma_l^2 + \frac{L\eta_t^2 \eta_l^2}{2n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

Reorganizing terms, we could get,

$$\begin{aligned}
\frac{17}{64} \eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] &\leq -(\mathbb{E}[f(z_{t+1})] - f(z_t)) + \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1-\beta_t} \right)^2 \mathbb{E} \left[\|d_t\|^2 \right] \\
&\quad + \frac{1}{2} \eta_t \left(\frac{K\eta_l^2}{n} \sigma_l^2 + \frac{\eta_l^2}{n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \right) + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2) \\
&\quad - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{LK\eta_l^2 \eta_t^2}{2n} \sigma_l^2 + \frac{L\eta_t^2 \eta_l^2}{2n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

that is,

$$\begin{aligned}
& \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq -\frac{64 \mathbb{E}[f(z_{t+1})] - f(z_t)}{17 \eta_t \eta_l K} + \frac{32 L^2}{17 \eta_l K} W_1^2 \mathbb{E} \left[\|d_t\|^2 \right] \\
& + \frac{32 \eta_l}{17 n} \sigma_l^2 + \frac{32 \eta_l}{17 K n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{160}{17} \eta_l^2 L^2 K (\sigma_l^2 + 6K\sigma_g^2) \\
& - \frac{32}{17 K^2 n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{32 L \eta_t \eta_l}{17 n} \sigma_l^2 + \frac{32 L \eta_t \eta_l}{17 K n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

Sum over all S stages and take average, we get,

$$\begin{aligned}
\bar{\mathcal{G}} & \triangleq \frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\
& \stackrel{(i)}{\leq} \frac{64 f(z_0) - \mathbb{E}[f(z_T)]}{17 S W_2 \eta_l K} + \frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_l K} \sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] + \frac{32 \eta_l}{17 n} \sigma_l^2 \\
& + \frac{32}{17 K n^2 W_2} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{160}{17} \eta_l^2 L^2 K (\sigma_l^2 + 6K\sigma_g^2) \\
& - \frac{32}{17 S W_2 K^2 n^2} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{32 L \bar{\eta} \eta_l}{17 n} \sigma_l^2 + \frac{32 L \eta_0 \bar{\eta} \eta_l}{17 W_2 K n^2} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& \stackrel{(ii)}{\leq} \frac{64 f(z_0) - \mathbb{E}[f(z_T)]}{17 S W_2 \eta_l K} + \frac{TK \eta_l^2}{n} \sigma_l^2 \frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_l K} + \frac{32 \eta_l}{17 n} \sigma_l^2 + \frac{160}{17} \eta_l^2 L^2 K (\sigma_l^2 + 6K\sigma_g^2) + \frac{32 L \bar{\eta} \eta_l}{17 n} \sigma_l^2 \\
& + \left(\frac{\eta_l^2}{n^2} C_\beta \frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_l K} + \frac{32}{17 K n^2 W_2} - \frac{32}{17 S W_2 K^2 n^2} + \frac{32 L \eta_0 \bar{\eta} \eta_l}{17 W_2 K n^2} \right) \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]
\end{aligned}$$

where $\bar{\eta} = \frac{1}{S} \sum_{s=1}^S \eta_s$. Due to η_t is stagewise, i.e. $\eta_t = \eta_s$ when $t \in \{T_0 + \dots + T_{s-1}, \dots, T_0 + \dots + T_s - 1\}$, and η_s is decaying, i.e. $\eta_s \leq \eta_s \leq \eta_0$, for any stage s , thus we have the following,

$$\begin{aligned}
& \frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \frac{32 L^2 W_1^2}{17 \eta_l K} \mathbb{E} \left[\|d_t\|^2 \right] = \frac{32 L^2 W_1^2}{17 \eta_l K} \frac{1}{S} \sum_{s=1}^S \frac{\eta_s}{T_s \eta_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|d_t\|^2 \right] \\
& = \frac{32 L^2 W_1^2}{17 \eta_l K} \frac{1}{S W_2} \sum_{s=1}^S \eta_s \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|d_t\|^2 \right] \leq \frac{32 L^2 W_1^2 \bar{\eta}}{17 \eta_l K W_2} \sum_{s=1}^S \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|d_t\|^2 \right] \\
& = \frac{32 L^2 W_1^2 \bar{\eta}}{17 \eta_l K W_2} \sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]
\end{aligned}$$

Similarly, we have,

$$\frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \frac{32 L \eta_t \eta_l}{17 K n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \leq \frac{32 L \eta_0 \bar{\eta} \eta_l}{17 W_2 K n^2} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]$$

and, we also have,

$$\frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \frac{32 L \eta_t \eta_l}{17 n} \sigma_l^2 = \frac{32 L \eta_l}{17 n} \sigma_l^2 \frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \eta_t = \frac{32 L \bar{\eta}}{17 n} \sigma_l^2$$

Thus, inequality (i) holds. (ii) holds by plugging into the bounds for $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$.

when the following two conditions hold,

$$\eta \leq \frac{1}{KSC_\eta(L\bar{\eta} + 1 + L^2W_1^2C_\eta)}$$

where $C_\eta = \frac{\eta_0}{\eta_S}$.

we could verify the coefficient of $\sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]$ is non-positive, by plugging in the learning rate constraints and using $C_\beta \leq C_\eta$.

$$\frac{\eta_L^2}{n^2} C_\beta \frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_L K} + \frac{32}{17} \frac{\eta \bar{\eta}}{K n^2 W_2} - \frac{32}{17} \frac{\eta_S}{S W_2 K^2 n^2} + \frac{32}{17} \frac{L \hat{\eta}^2 \eta}{W_2 K n^2} \leq 0$$

which results in,

$$\begin{aligned} \bar{\mathcal{G}} &\triangleq \frac{1}{S} \sum_{s=1}^S \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\ &\leq \frac{64}{17} \frac{f(z_0) - \mathbb{E}[f(z_T)]}{S W_2 \eta_L K} + \frac{TK \eta_L^2}{n} \sigma_i^2 \frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_L K} + \frac{32}{17} \frac{\eta}{n} \sigma_i^2 + \frac{160}{17} \eta_L^2 L^2 K (\sigma_i^2 + 6K \sigma_g^2) + \frac{32 L \bar{\eta} \eta}{17} \frac{\sigma_i^2}{n} \\ &\stackrel{(i)}{\leq} \frac{64}{17} \frac{f(x_0) - f^*}{S W_2 \eta_L K} + \left(\frac{32 L^2 W_1^2 T \bar{\eta} \eta}{17 n W_2} + \frac{32}{17} \frac{\eta}{n} + \frac{160}{17} \eta_L^2 L^2 K + \frac{32 L \bar{\eta} \eta}{17} \frac{1}{n} \right) \sigma_i^2 + \frac{960}{17} \eta_L^2 L^2 K^2 \sigma_g^2 \end{aligned}$$

where (i) holds as f is assumed to have minimum f^* .

Suppose $S = 1$, i.e. the typical constant hyperparameter regime, the total number of rounds are T , $\bar{\eta} = \eta_0 = \Theta(\sqrt{nK})$ and $\eta_L = \Theta(\frac{1}{\sqrt{TK}})$, $W_2 = \Theta(T\sqrt{nK})$ in this case. Suppose $W_1^2 = \mathcal{O}(\sqrt{nK})$. Considering in FedAvg, $\beta = 0$ and consequently $W_1 = 0$, thus, the condition $W_1^2 = \mathcal{O}(\sqrt{nK})$ naturally holds. We have the bound as,

$$\begin{aligned} \bar{\mathcal{G}} &\triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\ &\leq \mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right) (f(x_0) - f^*) + \left(\mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right) + \mathcal{O}\left(\frac{1}{TK}\right) + \mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right) \right) \sigma_i^2 + \mathcal{O}\left(\frac{1}{T}\right) \sigma_g^2 \end{aligned}$$

when T is sufficiently large, i.e. $T \geq Kn$, the dominant term is $\mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right)$.

Suppose $S > 1$, i.e. the multistage regime, the total number of rounds are T , $\bar{\eta} = \Theta(\sqrt{nK})$, $\eta_L = \Theta(\frac{1}{\sqrt{TK}})$, $W_2 = \Theta\left(\frac{T\sqrt{nK}}{S}\right)$, i.e. $T\bar{\eta}$ is equally divided into S stages. Suppose $W_1^2 = \mathcal{O}\left(\frac{\sqrt{nK}}{S}\right)$, we have the bound as,

$$\begin{aligned} \bar{\mathcal{G}} &\triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\ &\leq \mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right) (f(x_0) - f^*) + \left(\mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right) + \mathcal{O}\left(\frac{1}{TK}\right) + \mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right) \right) \sigma_i^2 + \mathcal{O}\left(\frac{1}{T}\right) \sigma_g^2 \end{aligned}$$

The dominant term is $\mathcal{O}\left(\frac{1}{\sqrt{TKn}}\right)$. □

E Proof of Theorem 4.6 and Corollary 4.7

Proof of Multistage GM with Partial Participation. Recall the formulation of General Momentum:

$$\begin{aligned} d_{t+1} &= (1 - \beta_t) \Delta_t + \beta_t d_t \\ x_{t+1} &= x_t - \eta_t [(1 - \nu_t) \Delta_t + \nu_t d_{t+1}] \end{aligned}$$

We first show Δ_t is an unbiased estimator of a virtual average $\Delta'_t \triangleq \frac{1}{n} \sum_{i=1}^n \Delta_t^i$,

$$\begin{aligned}
\mathbb{E}[\Delta_t] &= \mathbb{E}\left[\frac{1}{m} \sum_{i \in \mathcal{S}_t} \Delta_t^i\right] = \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^n \mathbf{1}(i \in \mathcal{S}_t) \Delta_t^i\right] \\
&= \frac{1}{m} \mathbb{E}\left[\sum_{i=1}^n \mathbb{P}\{i \in \mathcal{S}_t\} \Delta_t^i\right] \stackrel{(i)}{=} \frac{1}{n} \sum_{i=1}^n \Delta_t^i = \Delta'_t
\end{aligned}$$

where (i) follows from $\mathbb{P}\{i \in \mathcal{S}_t\} = \frac{m}{n}$.

We study the following Lyapunov sequence $\{z_t\}_{t=0}^{T-1}$, which is devised as follows:

$$z_t = x_t - \frac{\eta_t \beta_t \nu_t}{1 - \beta_t} d_t \quad (6)$$

where $d_0 = 0$.

We now verify $z_{t+1} - z_t = -\eta_t \Delta_t$,

$$\begin{aligned}
z_{t+1} - z_t &= x_{t+1} - \frac{\eta_{t+1} \beta_{t+1} \nu_{t+1}}{1 - \beta_{t+1}} d_{t+1} - x_t + \frac{\eta_t \beta_t \nu_t}{1 - \beta_t} d_t \\
&= -\eta_t y_t - W_1(d_{t+1} - d_t) \\
&= -\eta_t((1 - \nu_t)\Delta_t + \nu_t d_{t+1}) - W_1((1 - \beta_t)\Delta_t + \beta_t d_t - d_t) \\
&= -\eta_t(1 - \nu_t)\Delta_t - \eta_t \beta_t \nu_t \Delta_t - \eta_t \nu_t(d_{t+1} - \beta_t d_t) \\
&= -\eta_t(1 - \nu_t)\Delta_t - \eta_t \beta_t \nu_t \Delta_t - \eta_t \nu_t(1 - \beta_t)\Delta_t = -\eta_t \Delta_t
\end{aligned}$$

Since f is L -smooth, taking conditional expectation with respect to all randomness prior to step t , we have

$$\begin{aligned}
\mathbb{E}[f(z_{t+1})] &\leq f(z_t) + \mathbb{E}[\langle \nabla f(z_t), z_{t+1} - z_t \rangle] + \frac{L}{2} \mathbb{E}[\|z_{t+1} - z_t\|^2] \\
&\leq f(z_t) + \mathbb{E}[\langle \nabla f(z_t), -\eta_t \Delta_t \rangle] + \frac{L}{2} \eta_t^2 \mathbb{E}[\|\Delta_t\|^2] \\
&\leq f(z_t) + \underbrace{\mathbb{E}[\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle]}_{A_1} + \underbrace{\mathbb{E}[\langle \nabla f(x_t), -\eta_t \Delta_t \rangle]}_{A_2} + \underbrace{\frac{L}{2} \eta_t^2 \mathbb{E}[\|\Delta_t\|^2]}_{A_3}
\end{aligned}$$

Bounding A_1 :

$$\begin{aligned}
A_1 &= \mathbb{E}[\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle] \\
&\stackrel{(i)}{\leq} \mathbb{E}[\|\sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t))\| \cdot \|\sqrt{\eta_t} \Delta_t\|] \\
&\stackrel{(ii)}{\leq} \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t}\right)^2 \mathbb{E}[\|d_t\|^2] + \frac{1}{2} \eta_t \mathbb{E}[\|\Delta_t\|^2]
\end{aligned} \quad (7)$$

where (i) holds by applying Cauchy-Schwarz inequality, and (ii) follows from Young's inequality and f is L -smooth.

Bounding A_2 :

$$\begin{aligned}
A_2 &= \mathbb{E}[\langle \nabla f(x_t), -\eta_t \Delta_t \rangle] \\
&= \eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \Delta_t - \eta_t K \nabla f(x_t) \rangle] \\
&= -\eta_t \eta_t K \mathbb{E}[\|\nabla f(x_t)\|^2] + \eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \Delta_t \rangle]
\end{aligned}$$

where we further bound $\eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \Delta_t \rangle]$,

$$\begin{aligned}
& \eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \Delta_t \rangle] \\
& \stackrel{(i)}{=} \eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \Delta'_t \rangle] = \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t g_{t,k}^i \right\rangle \right] \\
& \stackrel{(ii)}{=} \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \eta_t K \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} \eta_t \nabla f_i(x_{t,k}^i) \right\rangle \right] \\
& \stackrel{(iii)}{=} \eta_t \mathbb{E} \left[\left\langle \nabla f(x_t), \frac{\eta_t}{n} \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right\rangle \right] \\
& = \eta_t \left\langle \sqrt{\eta_t K} \nabla f(x_t), \frac{\sqrt{\eta_t K}}{Kn} \mathbb{E} \left[\sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right] \right\rangle \\
& \stackrel{(iv)}{=} \frac{\eta_t \eta_t K}{2} \mathbb{E} [\|\nabla f(x_t)\|^2] + \frac{\eta_t \eta_t K}{2K^2 n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right\|^2 \right] \\
& \quad - \frac{\eta_t}{2} \mathbb{E} \left[\left\| \sqrt{\eta_t K} \left(\nabla f(x_t) - \frac{1}{Kn} \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right) \right\|^2 \right]
\end{aligned}$$

where (i) holds as Δ_t is an unbiased estimator of Δ'_t , (ii) holds as we take conditional expectation with respect to all randomness prior to step t . (iii) holds due to the following equality and the definition of $\nabla f(x_t) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_t)$. (iv) holds as $\langle a, b \rangle = \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|^2 - \frac{1}{2} \|a - b\|^2$.

We further bound the above terms as,

$$\begin{aligned}
& = \frac{\eta_t \eta_t K}{2} \mathbb{E} [\|\nabla f(x_t)\|^2] + \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} (\nabla f_i(x_t) - \nabla f_i(x_{t,k}^i)) \right\|^2 \right] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& \stackrel{(i)}{\leq} \frac{\eta_t \eta_t K}{2} \mathbb{E} [\|\nabla f(x_t)\|^2] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{\eta_t \eta_t L^2}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} (x_t - x_{t,k}^i) \right\|^2 \right] \\
& \stackrel{(ii)}{\leq} \frac{\eta_t \eta_t K}{2} \mathbb{E} [\|\nabla f(x_t)\|^2] - \frac{\eta_t \eta_t}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{\eta_t \eta_t L^2}{2n} \sum_{i=1}^n \sum_{k=0}^{K-1} \mathbb{E} [\|x_t - x_{t,k}^i\|^2]
\end{aligned}$$

where (i) holds due to L -smoothness of f_i , (ii) holds as $\|\sum_{i=1}^n x_i\|^2 \leq n \sum_{i=1}^n \|x_i\|^2$.

when $\eta_t \leq \frac{1}{8KL}$, we have the following bound,

$$\mathbb{E} [\|x_t - x_{t,k}^i\|^2] \leq 5K\eta_t^2 (\sigma_t^2 + 6K\sigma_g^2) + 30K^2\eta_t^2 \|\nabla f(x_t)\|^2$$

Plug in the above bound, we would have,

$$\begin{aligned}
& \frac{\eta_t \eta_l K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{\eta_t \eta_l L^2}{2n} \sum_{i=1}^n \sum_{k=0}^{K-1} \mathbb{E} \left[\|x_t - x_{t,k}^i\|^2 \right] \\
& \leq \frac{\eta_t \eta_l K}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& \quad + \frac{\eta_t \eta_l L^2 K}{2} \left(5K\eta_l^2 (\sigma_l^2 + 6K\sigma_g^2) + 30K^2\eta_l^2 \|\nabla f(x_t)\|^2 \right) \\
& \leq \left(\frac{\eta_t \eta_l K}{2} + 30K^2\eta_l^2 \frac{\eta_t \eta_l L^2 K}{2} \right) \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
& \quad + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2) \\
& \stackrel{(i)}{\leq} \frac{47}{64} \eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2)
\end{aligned}$$

where (i) holds by using the constraint $\eta_l \leq \frac{1}{8KL}$.

Therefore, merging all pieces together, we have,

$$\begin{aligned}
A_2 & = -\eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] + \eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_l K \nabla f(x_t) - \Delta_t \rangle] \\
& \leq -\frac{17}{64} \eta_t \eta_l K \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2)
\end{aligned}$$

Bounding $\mathbb{E} \left[\|\Delta_t\|^2 \right]$:

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

Bounding $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$:

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] &\leq \frac{KT\eta_l^2}{m} \sigma_l^2 + \frac{\eta_l^2}{m^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
&\leq \frac{KT\eta_l^2}{m} \sigma_l^2 + \frac{\eta_l^2}{m^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbb{P} \{i \in \mathcal{S}_t\} \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
&\leq \frac{KT\eta_l^2}{m} \sigma_l^2 + \frac{\eta_l^2}{m^2} \frac{m(n-m)}{n(n-1)} C_\beta \sum_{t=0}^{T-1} \sum_{i=1}^n \left\| \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 + \frac{\eta_l^2}{m^2} \frac{m(m-1)}{n(n-1)} C_\beta \sum_{t=0}^{T-1} \left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \\
&\leq \frac{KT\eta_l^2}{m} \sigma_l^2 + \frac{\eta_l^2 (m-1)}{mn(n-1)} C_\beta \sum_{t=0}^{T-1} \left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 + \\
&\quad \frac{\eta_l^2 (n-m)}{mn(n-1)} C_\beta \sum_{t=0}^{T-1} \sum_{i=1}^n \left\| \sum_{k=0}^{K-1} [\nabla f_i(x_{t,k}^i) - \nabla f_i(x_t^i)] + [\nabla f_i(x_t^i) - \nabla f(x_t)] + \nabla f(x_t) \right\|^2 \\
&\leq \frac{KT\eta_l^2}{m} \sigma_l^2 + \frac{\eta_l^2 (m-1)}{mn(n-1)} C_\beta \sum_{t=0}^{T-1} \left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 + \\
&\quad \frac{\eta_l^2 (n-m)}{mn(n-1)} C_\beta \sum_{t=0}^{T-1} \left(15nK^3L^3\eta_l^2 (\sigma_l^2 + 6K\sigma_g^2) + (90nK^4L^2\eta_l^2 + 3nK^2) \|\nabla f(x_t)\|^2 + 3nK^2\sigma_g^2 \right)
\end{aligned}$$

Merging A_1, A_2, A_3 together,

$$\begin{aligned}
\mathbb{E} [f(z_{t+1})] - f(z_t) &\leq \underbrace{\mathbb{E} [\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle]}_{A_1} + \underbrace{\mathbb{E} [\langle \nabla f(x_t), -\eta_t \Delta_t \rangle]}_{A_2} + \underbrace{\frac{L}{2} \eta_t^2 \mathbb{E} [\|\Delta_t\|^2]}_{A_3} \\
&\leq \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E} [\|d_t\|^2] + \frac{1}{2} \eta_t \mathbb{E} [\|\Delta_t\|^2] + \frac{L}{2} \eta_t^2 \mathbb{E} [\|\Delta_t\|^2] \\
&\quad - \frac{17}{64} \eta_t \eta_l K \mathbb{E} [\|\nabla f(x_t)\|^2] - \frac{\eta_t \eta_l}{2Kn^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{5}{2} \eta_t \eta_l^3 L^2 K^2 (\sigma_l^2 + 6K\sigma_g^2)
\end{aligned}$$

Reorganizing terms, we have the following,

$$\begin{aligned}
\mathbb{E} [\|\nabla f(x_t)\|^2] &\leq \frac{64}{17} \frac{f(z_t) - \mathbb{E} [f(z_{t+1})]}{\eta_t \eta_l K} + \frac{32}{17\eta_l K} L^2 W_1^2 \mathbb{E} [\|d_t\|^2] + \left(\frac{32}{17\eta_l K} + \frac{32L}{17} \frac{\eta_t}{\eta_l K} \right) \mathbb{E} [\|\Delta_t\|^2] \\
&\quad - \frac{32}{17K^2 n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] + \frac{160}{17} \eta_l^2 L^2 K (\sigma_l^2 + 6K\sigma_g^2)
\end{aligned}$$

Sum over all S stages and take average, we get,

$$\begin{aligned}
\bar{\mathcal{G}} &\triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\
&\leq \frac{64}{17} \frac{f(z_0) - \mathbb{E}[f(z_T)]}{SW_2\eta_l K} + \frac{32}{17} \frac{L^2 W_1^2 \bar{\eta}}{W_2 \eta_l K} \sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] - \frac{32}{17} \frac{\eta_S}{SW_2 K^2 n^2} \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
&\quad + \frac{160}{17} \eta_l^2 L^2 K (\sigma_l^2 + 6K\sigma_g^2) + \left(\frac{32\bar{\eta}}{17\eta_l K W_2} + \frac{32L}{17} \frac{\hat{\eta}^2}{\eta_l W_2 K} \right) \sum_{t=0}^{T-1} \mathbb{E} \left[\|\Delta_t\|^2 \right] \\
&\leq \frac{64}{17} \frac{f(z_0) - \mathbb{E}[f(z_T)]}{SW_2\eta_l K} + \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right] \\
&\quad \left(-\frac{32}{17} \frac{\eta_S}{SW_2 K^2 n^2} + \frac{\eta_l^2 (m-1)}{nm(n-1)} \left(\frac{32\bar{\eta}}{17\eta_l K W_2} + \frac{32L}{17} \frac{\hat{\eta}^2}{\eta_l W_2 K} \right) + \frac{\eta_l^2 (m-1)}{mn(n-1)} \frac{32}{17} \frac{L^2 W_1^2 \bar{\eta}}{W_2 \eta_l K} \right) \\
&\quad + \left(\frac{32}{17} \frac{L^2 W_1^2 \bar{\eta}}{W_2 \eta_l K} \frac{\eta_l^2 (n-m)}{mn(n-1)} (90nK^4 L^2 \eta_l^2 + 3nK^2) \right) \cdot \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \\
&\quad + \left(\left(\frac{32\bar{\eta}}{17\eta_l K W_2} + \frac{32L}{17} \frac{\hat{\eta}^2}{\eta_l W_2 K} \right) \frac{\eta_l^2 (n-m)}{nm(n-1)} (90nK^4 L^2 \eta_l^2 + 3nK^2) \right) \cdot \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \\
&\quad + \left(\frac{\eta_l}{m} \Phi + \frac{15(n-m)K^2 L^3 \eta_l^3}{m(n-1)} \Phi + \frac{160}{17} \eta_l^2 L^2 K \right) \sigma_l^2 \\
&\quad + \left(\frac{90(n-m)K^3 L^3 \eta_l^3}{m(n-1)} \Phi + \frac{3\eta_l(n-m)K}{m(n-1)} \Phi + \frac{960}{17} \eta_l^2 L^2 K \right) \sigma_g^2
\end{aligned}$$

where we denote Φ for ease of notation,

$$\Phi \triangleq \frac{32T\bar{\eta} + 32LT\hat{\eta}^2 + 32L^2 W_1^2 T\bar{\eta}}{17W_2}$$

We can verify, when the following condition holds,

$$\eta_l \leq \frac{1}{(C_\eta + L\bar{\eta}C_\eta + L^2 W_1^2 C_\eta) SK} \frac{m(n-1)}{n(m-1)}$$

where $C_\eta = \frac{\eta_0}{\eta_S}$.

we have the coefficient for $\sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \sum_{k=0}^{K-1} \nabla f_i(x_{t,k}^i) \right\|^2 \right]$,

$$-\frac{32}{17} \frac{\eta_S}{SW_2 K^2 n^2} + \frac{\eta_l^2 (m-1)}{nm(n-1)} \left(\frac{32\bar{\eta}}{17\eta_l K W_2} + \frac{32L}{17} \frac{\hat{\eta}^2}{\eta_l W_2 K} \right) + \frac{\eta_l^2 (m-1)}{mn(n-1)} \frac{32}{17} \frac{L^2 W_1^2 \bar{\eta}}{W_2 \eta_l K} \leq 0$$

With the following inequality,

$$\frac{1}{SW_2} \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 = \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s \eta_S} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{1}{\eta_S} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2$$

We could verify, when the following condition holds,

$$\eta_l \leq \frac{17}{282} \frac{m}{(C_\eta + LC_\eta \bar{\eta} + L^2 W_1^2 C_\eta) SK}$$

We have the following,

$$\begin{aligned}
& \left(\frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_l K} \frac{\eta_l^2 (n-m)}{m n (n-1)} (90 n K^4 L^2 \eta_l^2 + 3 n K^2) \right) \cdot \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \\
& + \left(\left(\frac{32 \bar{\eta}}{17 \eta_l K W_2} + \frac{32 L}{17} \frac{\hat{\eta}^2}{\eta_l W_2 K} \right) \frac{\eta_l^2 (n-m)}{n m (n-1)} (90 n K^4 L^2 \eta_l^2 + 3 n K^2) \right) \cdot \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \\
& \stackrel{(i)}{\leq} \frac{\eta_l^2 (n-m)}{n m (n-1)} \frac{141 n K^2}{32} \left(\frac{32 \bar{\eta}}{17 \eta_l K W_2} + \frac{32 L}{17} \frac{\hat{\eta}^2}{\eta_l W_2 K} + \frac{32 L^2 W_1^2 \bar{\eta}}{17 W_2 \eta_l K} \right) \cdot \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \\
& \leq \frac{\eta_l^2 (n-m)}{n m (n-1)} \frac{141 n K^2}{32} \frac{32 \eta_S}{17 \eta_l K W_2} (C_\eta + L C_\eta \bar{\eta} + L^2 W_1^2 C_\eta) \cdot \sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \\
& \stackrel{(ii)}{\leq} \frac{1}{2} \cdot \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2
\end{aligned}$$

where (i) holds as $90 n K^4 L^2 \eta_l^2 + 3 n K^2 \leq \frac{141}{32} n K^2$ when $\eta_l \leq \frac{1}{8 K L}$, (ii) holds by plugging in the learning rate constraint $\eta_l \leq \frac{17}{282} \frac{m}{(C_\eta + L C_\eta \bar{\eta} + L^2 W_1^2 C_\eta) S K}$.

Merging everything together, we have the following,

$$\begin{aligned}
\bar{g} & \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \\
& \leq \frac{64}{17} \frac{f(z_0) - \mathbb{E}[f(z_T)]}{S W_2 \eta_l K} + \left(\frac{\eta_l}{m} \Phi + \frac{15 (n-m) K^2 L^3 \eta_l^3}{m (n-1)} \Phi + \frac{160}{17} \eta_l^2 L^2 K \right) \sigma_l^2 \\
& \quad + \left(\frac{90 (n-m) K^3 L^3 \eta_l^3}{m (n-1)} \Phi + \frac{3 \eta_l (n-m) K}{m (n-1)} \Phi + \frac{960}{17} \eta_l^2 L^2 K^2 \right) \sigma_g^2
\end{aligned}$$

Suppose $S = 1$, i.e. the typical constant hyperparameter regime, the total number of rounds are T , $\bar{\eta} = \eta_0 = \Theta(\sqrt{mK})$ and $\eta_l = \Theta\left(\frac{1}{\sqrt{TK}}\right)$, $W_2 = \Theta(T\sqrt{mK})$ in this case. Assume $W_1^2 = \mathcal{O}(\sqrt{mK})$, recall $\Phi \triangleq \frac{32 T \bar{\eta} + 32 L T \hat{\eta}^2 + 32 L^2 W_1^2 T \bar{\eta}}{17 W_2}$, we could verify $\Phi = \Theta(\sqrt{mK})$.

We have the bound as,

$$\begin{aligned}
\bar{g} & \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\
& \leq \mathcal{O} \left(\frac{1}{\sqrt{TKm}} \right) (f(x_0) - f^*) + \left(\mathcal{O} \left(\frac{1}{\sqrt{TKm}} \right) + \mathcal{O} \left(\frac{1}{\sqrt{T^3 K m}} \right) + \mathcal{O} \left(\frac{1}{TK} \right) \right) \sigma_l^2 \\
& \quad + \left(\mathcal{O} \left(\sqrt{\frac{K}{T^3 m}} \right) + \mathcal{O} \left(\sqrt{\frac{K}{T m}} \right) + \mathcal{O} \left(\frac{1}{T} \right) \right) \sigma_g^2
\end{aligned}$$

Only keeping the dominant terms,

$$\begin{aligned}
\bar{g} & \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\
& \leq \mathcal{O} \left(\frac{1}{\sqrt{TKm}} \right) (f(x_0) - f^*) + \mathcal{O} \left(\frac{1}{\sqrt{TKm}} \right) \sigma_l^2 + \mathcal{O} \left(\sqrt{\frac{K}{T m}} \right) \sigma_g^2
\end{aligned}$$

Suppose $S > 1$ but $S = \Theta(1)$, i.e. the multistage regime, the total number of rounds are T , $\bar{\eta} = \Theta(\sqrt{mK})$, and $\hat{\eta}^2 = \Theta(mK)$, $\eta_l = \Theta\left(\frac{1}{\sqrt{TK}}\right)$, $W_2 = \Theta\left(\frac{T\sqrt{mK}}{S}\right)$, i.e. $T\bar{\eta}$ is equally divided into S stages. Assume $W_1^2 = \mathcal{O}(\sqrt{mK})$, we could verify $\Phi = \Theta(\sqrt{mK})$. Thus, we have the bound,

$$\begin{aligned}\bar{g} &\triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\ &\leq \mathcal{O} \left(\frac{1}{\sqrt{TKm}} \right) (f(x_0) - f^*) + \mathcal{O} \left(\frac{1}{\sqrt{TKm}} \right) \sigma_l^2 + \mathcal{O} \left(\sqrt{\frac{K}{Tm}} \right) \sigma_g^2\end{aligned}$$

In both cases, the dominant term is $\mathcal{O} \left(\sqrt{\frac{K}{Tm}} \right)$.

□

F Proof of Theorem 5.1 and Corollary 5.2

Proof of Autonomous Multistage GM with Uniform Arrival. We introduce a Lyapunov sequence $\{z_t\}_{t=0}^{T-1}$ which is devised as follows:

$$z_t = x_t - \frac{\eta_t \beta_t \nu_t}{1 - \beta_t} d_t \quad (8)$$

where $d_0 = 0$.

We could easily verify $z_{t+1} - z_t = -\eta_t \Delta_t$. Let $-\eta_t y_t = x_{t+1} - x_t$, we first bound $\mathbb{E} \left[\|\Delta_t\|^2 \right]$, $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$, and $\sum_{t=0}^{T-1} \mathbb{E} \left[\|y_t\|^2 \right]$.

Bounding $\mathbb{E} \left[\|\Delta_t\|^2 \right]$:

$$\begin{aligned}\mathbb{E} \left[\|\Delta_t\|^2 \right] &= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \Delta_{t-\tau_t, i}^i \right\|^2 \right] \\ &\stackrel{(i)}{=} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} g_{t-\tau_t, i, k}^i \right\|^2 \right] \\ &= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \left(g_{t-\tau_t, i, k}^i - \nabla f_i(x_{t-\tau_t, i, k}^i) + \nabla f_i(x_{t-\tau_t, i, k}^i) \right) \right\|^2 \right] \\ &\stackrel{(ii)}{=} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \left\{ g_{t-\tau_t, i, k}^i - \nabla f_i(x_{t-\tau_t, i, k}^i) \right\} \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_t, i, k}^i) \right\|^2 \right] \\ &\stackrel{(iii)}{\leq} \frac{1}{m^2} \sum_{i \in \mathcal{S}_t} \frac{\eta_l^2}{K_{t,i}^2} \sum_{k=0}^{K_{t,i}-1} \sigma_l^2 + \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_t, i, k}^i) \right\|^2 \right] \\ &\leq \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_t, i, k}^i) \right\|^2 \right]\end{aligned}$$

where $\frac{1}{K_t} = \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}}$. (i) follows from the definition of $\Delta_{t-\tau_t, i}^i$, (ii) and (iii) hold as $\mathbb{E} \left[\left\| \sum_{i=1}^n x_i \right\|^2 \right] = \sum_{i=1}^n \mathbb{E} \left[\|x_i\|^2 \right]$ when $\mathbb{E} [x_i] = 0$, and we know $\mathbb{E} \left[g_{t-\tau_t, i, k}^i - \nabla f_i(x_{t-\tau_t, i, k}^i) \right] = 0$.

Bounding $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$:

We could verify:

$$d_t = \sum_{p=0}^t a_{t,p} \Delta_p, \quad \text{where} \quad a_{t,p} = (1 - \beta_p) \prod_{q=p+1}^t \beta_q$$

We further get,

$$\begin{aligned}
\mathbb{E} \left[\|d_t\|^2 \right] &= \mathbb{E} \left[\left\| \sum_{p=0}^t a_{t,p} \Delta_p \right\|^2 \right] \\
&\leq \sum_{e=1}^d \mathbb{E} \left[\sum_{p=0}^t a_{t,p} \Delta_{p,e} \right]^2 \leq \sum_{e=1}^d \mathbb{E} \left[\left(\sum_{p=0}^t a_{t,p} \right) \left(\sum_{p=0}^t a_{t,p} \Delta_{p,e}^2 \right) \right] \leq \left(1 - \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t a_{t,p} \mathbb{E} \left[\|\Delta_p\|^2 \right] \\
&\leq \left(1 - \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t a_{t,p} \left\{ \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \right\} \\
&\leq \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^t a_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Summing over $t \in \{0, 1, \dots, T-1\}$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] &\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{t=0}^{T-1} \sum_{p=0}^t a_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^{T-1} \left(\sum_{t=p}^{T-1} a_{t,p} \right) \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Bounding $\sum_{t=0}^{T-1} \mathbb{E} \left[\|y_t\|^2 \right]$:

We could verify:

$$y_t = \sum_{p=0}^t b_{t,p} \Delta_p,$$

where $b_{t,p}$ is defined as follows,

$$b_{t,p} = \begin{cases} 1 - \beta_t \nu_t & p = t \\ \nu_t (1 - \beta_p) \prod_{q=p+1}^t \beta_q & p < t \end{cases}$$

We further get,

$$\begin{aligned}
\mathbb{E} \left[\|y_t\|^2 \right] &= \mathbb{E} \left[\left\| \sum_{p=0}^t b_{t,p} \Delta_p \right\|^2 \right] \\
&\leq \sum_{e=1}^d \mathbb{E} \left[\sum_{p=0}^t b_{t,p} \Delta_{p,e} \right]^2 \leq \sum_{e=1}^d \mathbb{E} \left[\left(\sum_{p=0}^t b_{t,p} \right) \left(\sum_{p=0}^t b_{t,p} \Delta_{p,e}^2 \right) \right] \leq \left(1 - \nu_t \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t b_{t,p} \mathbb{E} \left[\|\Delta_p\|^2 \right] \\
&\leq \left(1 - \nu_t \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t b_{t,p} \left\{ \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \right\} \\
&\leq \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^t b_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Summing over $t \in \{0, 1, \dots, T-1\}$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} [\|y_t\|^2] &\leq \frac{\eta_t^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_t^2 + \frac{\eta_t^2}{m^2} \sum_{t=0}^{T-1} \sum_{p=0}^t b_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_t^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_t^2 + \frac{\eta_t^2}{m^2} \sum_{p=0}^{T-1} \left(\sum_{t=p}^{T-1} b_{t,p} \right) \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_p\} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_t^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_t^2 + \frac{\eta_t^2}{m^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Since f is L -smooth, taking conditional expectation with respect to all randomness prior to step t , we have

$$\begin{aligned}
\mathbb{E} [f(z_{t+1})] &\leq f(z_t) + \mathbb{E} [\langle \nabla f(z_t), z_{t+1} - z_t \rangle] + \frac{L}{2} \mathbb{E} [\|z_{t+1} - z_t\|^2] \\
&\leq f(z_t) + \mathbb{E} [\langle \nabla f(z_t), -\eta_t \Delta_t \rangle] + \frac{L}{2} \eta_t^2 \mathbb{E} [\|\Delta_t\|^2] \\
&\leq f(z_t) + \underbrace{\mathbb{E} [\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle]}_{A_1} + \underbrace{\mathbb{E} [\langle \nabla f(x_t), -\eta_t \Delta_t \rangle]}_{A_2} + \underbrace{\frac{L}{2} \eta_t^2 \mathbb{E} [\|\Delta_t\|^2]}_{A_3}
\end{aligned}$$

Bounding A_1 :

$$\begin{aligned}
A_1 &= \mathbb{E} [\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle] \\
&\leq \mathbb{E} [\|\sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t))\| \cdot \|\sqrt{\eta_t} \Delta_t\|] \\
&\stackrel{(i)}{\leq} \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E} [\|d_t\|^2] + \frac{1}{2} \eta_t \mathbb{E} [\|\Delta_t\|^2]
\end{aligned}$$

where (i) holds by applying Cauchy-Schwarz inequality, and (ii) follows from Young's inequality and f is L -smooth.

Bounding A_2 :

$$\begin{aligned}
A_2 &= \mathbb{E} [\langle \nabla f(x_t), -\eta_t \Delta_t \rangle] \\
&= \eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t - \eta_t \nabla f(x_t) \rangle] \\
&= -\eta_t \mathbb{E} [\|\nabla f(x_t)\|^2] + \eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t \rangle]
\end{aligned}$$

where we further bound $\eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t \rangle]$,

$$\begin{aligned}
\eta_t \mathbb{E} [\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t \rangle] &= \eta_t \mathbb{E} \left[\left\langle \sqrt{\eta_t} \nabla f(x_t), \frac{\sqrt{\eta_t}}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f(x_t) - g_{t-\tau_{t,i},k}^i) \right\rangle \right] \\
&\stackrel{(i)}{=} \eta_t \mathbb{E} \left[\left\langle \sqrt{\eta_t} \nabla f(x_t), \frac{\sqrt{\eta_t}}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f(x_t) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\rangle \right] \\
&\stackrel{(ii)}{=} \eta_t \mathbb{E} \left[\left\langle \sqrt{\eta_t} \nabla f(x_t), \frac{\sqrt{\eta_t}}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\rangle \right] \\
&\stackrel{(iii)}{=} \frac{\eta_t \eta_t}{2} \mathbb{E} [\|\nabla f(x_t)\|^2] - \frac{\eta_t \eta_t}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&\quad + \frac{\eta_t \eta_t}{2} \mathbb{E} \left[\left\| \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

where (i) holds as we take conditional expectation with respect to all randomness prior to step t . (ii) holds due to the following equality and the definition of $\nabla f(x_t) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_t)$. (iii) holds as $\langle a, b \rangle = \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|^2 - \frac{1}{2} \|a - b\|^2$.

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right] = \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^n \mathbf{1}(i \in \mathcal{S}_t) \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right] \\ &= \frac{1}{m} \mathbb{E} \left[\sum_{i=1}^n \mathbb{P}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right] \stackrel{(i)}{=} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right] \end{aligned}$$

where (i) holds due to uniform arrival assumption, in which $\mathbb{P}\{i \in \mathcal{S}_t\} = \frac{m}{n}$.

We further have,

$$\begin{aligned} & \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \nabla f(x_t) - \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\ &= \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\|^2 \right] \\ &= \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i}}) + \nabla f_i(x_{t-\tau_{t,i}}) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\|^2 \right] \\ &\stackrel{(i)}{\leq} \eta_t \eta_l \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n (\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i}})) \right\|^2 \right] + \eta_t \eta_l \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f_i(x_{t-\tau_{t,i}}) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\|^2 \right] \\ &\stackrel{(ii)}{\leq} \frac{\eta_t \eta_l}{n} \sum_{i=1}^n \mathbb{E} \left[\|\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i}})\|^2 \right] + \frac{\eta_t \eta_l}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \mathbb{E} \left[\|\nabla f_i(x_{t-\tau_{t,i}}) - \nabla f_i(x_{t-\tau_{t,i},k}^i)\|^2 \right] \\ &\stackrel{(iii)}{\leq} \frac{\eta_t \eta_l L^2}{n} \sum_{i=1}^n \mathbb{E} \left[\|x_t - x_{t-\tau_{t,i}}\|^2 \right] + \frac{\eta_t \eta_l L^2}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \mathbb{E} \left[\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2 \right] \end{aligned}$$

where (i) and (ii) hold as $\|\sum_{i=1}^n x_i\|^2 \leq n \sum_{i=1}^n \|x_i\|^2$, (iii) holds as f_i is L -smooth.

Thus, we have,

$$\begin{aligned} A_2 &\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\ &+ \frac{\eta_t \eta_l L^2}{n} \sum_{i=1}^n \mathbb{E} \left[\|x_t - x_{t-\tau_{t,i}}\|^2 \right] + \frac{\eta_t \eta_l L^2}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \mathbb{E} \left[\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2 \right] \end{aligned}$$

When $\eta_l \leq \frac{1}{8K_{t,i}L}$, we have,

$$\mathbb{E} \left[\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2 \right] \leq 5K_{t,i} \eta_l^2 (\sigma_l^2 + 6K_{t,i} \sigma_g^2) + 30K_{t,i}^2 \eta_l^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right]$$

We can further bound $\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|x_t - x_{t-\tau_{t,i}}\|^2 \right]$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[\|x_t - x_{t-\tau_{t,i}}\|^2 \right] &\stackrel{(i)}{\leq} \mathbb{E} \left[\|x_t - x_{t-\tau_{t,u}}\|^2 \right] = \mathbb{E} \left[\left\| \sum_{k=t-\tau_{t,u}}^{t-1} (x_{k+1} - x_k) \right\|^2 \right] \\ &\stackrel{(ii)}{\leq} \mathbb{E} \left[\left\| \sum_{k=t-\tau_{t,u}}^{t-1} \eta_k y_k \right\|^2 \right] \stackrel{(iii)}{\leq} \tau \eta_0^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] \end{aligned}$$

where (i) holds as we define $u = \arg \max_{i \in \{1, 2, \dots, n\}} \mathbb{E} \left[\|x_t - x_{t-\tau_{t,i}}\|^2 \right]$, (ii) follows from the definition of y_k , (iii) holds as bounded maximum delay assumption, i.e. $\tau_{t,i} \leq \tau$ for any t and i , and learning rate is decaying, i.e. $\eta_t \leq \eta_0$.

Merging all pieces together,

$$\begin{aligned} A_2 &\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\ &\quad + \eta_t \eta_l L^2 \tau \eta_0^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] + \frac{\eta_t \eta_l L^2}{n} \sum_{i=1}^n \left\{ 5K_{t,i} \eta_l^2 (\sigma_l^2 + 6K_{t,i} \sigma_g^2) + 30K_{t,i}^2 \eta_l^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] \right\} \\ &= -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + \eta_t \eta_l L^2 \tau \eta_0^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] \\ &\quad + \frac{30\eta_t \eta_l^3 L^2}{n} \sum_{i=1}^n K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] + 5\bar{K}_t \eta_t \eta_l^3 L^2 \sigma_l^2 + 30\hat{K}_t^2 \eta_t \eta_l^3 L^2 \sigma_g^2 \end{aligned}$$

where $\bar{K}_t \triangleq \frac{1}{n} \sum_{i=1}^n K_{t,i}$ and $\hat{K}_t^2 \triangleq \frac{1}{n} \sum_{i=1}^n K_{t,i}^2$.

Plug all pieces back in $\mathbb{E} [f(z_{t+1})] \leq f(z_t) + A_1 + A_2 + A_3$,

$$\begin{aligned} \mathbb{E} [f(z_{t+1})] - f(z_t) &\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\ &\quad + \eta_t \eta_l L^2 \tau \eta_0^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] + \frac{30\eta_t \eta_l^3 L^2}{n} \sum_{i=1}^n K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] + 5\bar{K}_t \eta_t \eta_l^3 L^2 \sigma_l^2 + 30\hat{K}_t^2 \eta_t \eta_l^3 L^2 \sigma_g^2 \\ &\quad + \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E} \left[\|d_t\|^2 \right] + \frac{1}{2} \eta_t \mathbb{E} \left[\|\Delta_t\|^2 \right] + \frac{L}{2} \eta_t^2 \mathbb{E} \left[\|\Delta_t\|^2 \right] \end{aligned}$$

Reorganizing terms and we have,

$$\begin{aligned} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] &\leq \frac{2(f(z_t) - \mathbb{E} [f(z_{t+1})])}{\eta_t \eta_l} - \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\ &\quad + 2L^2 \tau \eta_0^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] + \frac{60\eta_t^2 L^2}{n} \sum_{i=1}^n K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] + 10\bar{K}_t \eta_t^2 L^2 \sigma_l^2 + 60\hat{K}_t^2 \eta_t^2 L^2 \sigma_g^2 \\ &\quad + \frac{L^2 W_1^2}{\eta_l} \mathbb{E} \left[\|d_t\|^2 \right] + \frac{1}{\eta_l} \mathbb{E} \left[\|\Delta_t\|^2 \right] + \frac{L\eta_t}{\eta_l} \mathbb{E} \left[\|\Delta_t\|^2 \right] \end{aligned}$$

Sum over all S stages and take average, we get,

$$\begin{aligned}
\bar{G} &\triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(z_0) - \mathbb{E}[f(z_T)])}{SW_2\eta_l} \\
&- \frac{\eta_S}{SW_2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i,k}}^i) \right\|^2 \right] + \frac{2L^2\tau\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} [\|y_k\|^2] \\
&+ \frac{60\eta_l^2 L^2}{n} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \sum_{i=1}^n K_{t,i}^2 \mathbb{E} [\|\nabla f(x_{t-\tau_{t,i}})\|^2] \\
&+ \frac{L^2 W_1^2 \bar{\eta}}{W_2 \eta_l} \sum_{t=0}^{T-1} \mathbb{E} [\|d_t\|^2] + \frac{\bar{\eta}}{W_2 \eta_l} \sum_{t=0}^{T-1} \mathbb{E} [\|\Delta_t\|^2] + \frac{L\hat{\eta}^2}{W_2 \eta_l} \sum_{t=0}^{T-1} \mathbb{E} [\|\Delta_t\|^2] \\
&+ 10\eta_l^2 L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \right\} \sigma_l^2 + 60\eta_l^2 L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \hat{K}_t^2 \right\} \sigma_g^2
\end{aligned}$$

where $\bar{\eta} = \frac{1}{S} \sum_{s=0}^{S-1} \eta_s$, $\hat{\eta}^2 = \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^2$, and $\hat{\eta}^3 = \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^3$, respectively.

When the following holds,

$$\eta_l \leq \sqrt{\frac{1}{120L^2 C_{\eta\tau} K_{t,\max}^2}}, \quad \forall t \in \{0, \dots, T-1\}$$

we could verify the following inequality,

$$\begin{aligned}
&\frac{60\eta_l^2 L^2}{n} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \sum_{i=1}^n K_{t,i}^2 \mathbb{E} [\|\nabla f(x_{t-\tau_{t,i}})\|^2] \\
&\leq_{(i)} 60\eta_l^2 L^2 \frac{\eta_0}{SW_2} \sum_{t=0}^{T-1} K_{t,\max}^2 \mathbb{E} [\|\nabla f(x_{t-\tau_{t,i}})\|^2] \\
&\leq_{(ii)} 60\eta_l^2 L^2 \tau \frac{\eta_0}{SW_2} \sum_{t=0}^{T-1} K_{t,\max}^2 \mathbb{E} [\|\nabla f(x_t)\|^2] \leq_{(iii)} \frac{\eta_S}{2SW_2} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(x_t)\|^2] \\
&\leq_{(iv)} \frac{1}{2} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2
\end{aligned}$$

where (i) follows from the definition of $K_{t,\max}^2 = \max_{i \in \{1,2,\dots,n\}} K_{t,i}^2$, and $W_2 = \eta_s T_s$ for all $s \in \{1, \dots, S\}$ and $\eta_S \leq \eta_s \leq \eta_0$. (ii) follows from the maximum delay assumption. (iii) holds by plugging in the assumption $\eta_l \leq \sqrt{\frac{\eta_S}{120L^2 \eta_0 \tau K_{t,\max}^2}}$, $\forall t \in \{0, \dots, T-1\}$. (iv) holds as $\frac{\eta_S}{2SW_2} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(x_t)\|^2] \leq \frac{\eta_S}{2} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s \eta_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2$ and $\frac{1}{\eta_S} \leq \frac{1}{\eta_s}$ for all s .

With the maximum delay assumption, we have $\frac{2L^2\tau\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} [\|y_k\|^2] \leq \frac{2L^2\tau^2\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \mathbb{E} [\|y_t\|^2]$. Merging all pieces, we have,

$$\begin{aligned}
& \frac{1}{2} \cdot \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(z_0) - \mathbb{E}[f(z_T)])}{SW_2\eta} \\
& - \frac{\eta_S}{SW_2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + \frac{2L^2\tau^2\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \mathbb{E} [\|y_t\|^2] \\
& + \frac{L^2W_1^2\bar{\eta}}{W_2\eta} \sum_{t=0}^{T-1} \mathbb{E} [\|d_t\|^2] + \frac{\bar{\eta}}{W_2\eta} \sum_{t=0}^{T-1} \mathbb{E} [\|\Delta_t\|^2] + \frac{L\hat{\eta}^2}{W_2\eta} \sum_{t=0}^{T-1} \mathbb{E} [\|\Delta_t\|^2] \\
& + 10\eta_l^2L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \right\} \sigma_l^2 + 60\eta_l^2L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \hat{K}_t \right\} \sigma_g^2
\end{aligned}$$

We define ϕ_1, ϕ_2 , and ϕ_3 for ease of notation.

$$\phi_1 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \bar{K}_t, \quad \text{and} \quad \phi_2 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \hat{K}_t, \quad \text{and} \quad \phi_3 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K_t}$$

We could verify,

$$\frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \stackrel{(i)}{\leq} \frac{1}{W_2} \frac{1}{S} \sum_{s=0}^{S-1} \eta_s \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \stackrel{(ii)}{\leq} \frac{\bar{\eta}}{W_2} \sum_{t=0}^{T-1} \bar{K}_t = \frac{T\bar{\eta}}{W_2} \phi_1$$

(i) holds due to $T_s\eta_s = W_2$ by assumption, (ii) holds due to $\frac{1}{S} \sum_{s=0}^{S-1} \eta_s \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \leq \left(\frac{1}{S} \sum_{s=0}^{S-1} \eta_s \right) \cdot \left(\sum_{s=0}^{S-1} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \right) = \bar{\eta} \sum_{t=0}^{T-1} \bar{K}_t$.

Similarly, we have,

$$\frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \hat{K}_t \leq \frac{T\bar{\eta}}{W_2} \phi_2, \quad \text{and} \quad \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \frac{1}{K_t} \leq \frac{T\bar{\eta}}{W_2} \phi_3$$

Plugging in the bounds for $\sum_{t=0}^{T-1} \mathbb{E} [\|\Delta_t\|^2]$, $\sum_{t=0}^{T-1} \mathbb{E} [\|y_t\|^2]$, and $\sum_{t=0}^{T-1} \mathbb{E} [\|d_t\|^2]$,

$$\begin{aligned}
& \frac{1}{2} \cdot \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(z_0) - \mathbb{E}[f(z_T)])}{SW_2\eta} \\
& + \frac{60\eta_l^2L^2T\bar{\eta}\phi_2}{W_2} \sigma_g^2 - \frac{\eta_S}{SW_2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
& + \left(\frac{10\eta_l^2L^2T\bar{\eta}}{W_2} \phi_1 + \frac{2L^2\tau^2\hat{\eta}^3\eta_l^2T}{mW_2} \phi_3 + \frac{L^2W_1^2\bar{\eta}\eta_lT}{mW_2} \phi_3 + \frac{\bar{\eta}\eta_lT}{mW_2} \phi_3 + \frac{L\hat{\eta}^2\eta_l}{mW_2} \phi_3 \right) \sigma_l^2 \\
& + \frac{\eta_l^2}{m^2} \left(\frac{\bar{\eta}}{W_2\eta_l} + \frac{L\hat{\eta}^2}{W_2\eta_l} + \frac{L^2W_1^2\bar{\eta}C_\beta}{W_2\eta_l} + \frac{2L^2\tau^2\hat{\eta}^3C_\beta}{W_2} \right) \\
& \cdot \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

We now bound $\mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]$,

$$\begin{aligned}
\mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] &= \sum_{i=1}^n \mathbb{P}\{i \in \mathcal{S}_t\} \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \\
&+ \sum_{i \neq j} \mathbb{P}\{i, j \in \mathcal{S}_t\} \left\langle \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i), \frac{1}{K_{t,j}} \sum_{k=0}^{K_{t,j}-1} \nabla f_j(x_{t-\tau_{t,j},k}^j) \right\rangle \\
&\stackrel{(i)}{=} \frac{m}{n} \sum_{i=1}^n \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \\
&+ \frac{m(m-1)}{n(n-1)} \sum_{i \neq j} \left\langle \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i), \frac{1}{K_{t,j}} \sum_{k=0}^{K_{t,j}-1} \nabla f_j(x_{t-\tau_{t,j},k}^j) \right\rangle \\
&\stackrel{(ii)}{=} \frac{m^2}{n} \sum_{i=1}^n \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \\
&- \frac{m(m-1)}{2n(n-1)} \sum_{i \neq j} \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) - \frac{1}{K_{t,j}} \sum_{k=0}^{K_{t,j}-1} \nabla f_j(x_{t-\tau_{t,j},k}^j) \right\|^2
\end{aligned}$$

where (i) follows from uniform arrival assumption, i.e. $\mathbb{P}\{i, j \in \mathcal{S}_t\} = \frac{m(m-1)}{n(n-1)}$ and $\mathbb{P}\{i \in \mathcal{S}_t\} = \frac{m}{n}$. (ii) follows from $\langle a, b \rangle = \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|^2 - \frac{1}{2} \|a - b\|^2$.

The following equality with respect to $\mathbb{E} \left[\left\| \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]$ is straightforward to verify,

$$\begin{aligned}
\mathbb{E} \left[\left\| \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] &= n \sum_{i=1}^n \mathbb{E} \left[\left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&- \frac{1}{2} \sum_{i \neq j} \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) - \frac{1}{K_{t,j}} \sum_{k=0}^{K_{t,j}-1} \nabla f_j(x_{t-\tau_{t,j},k}^j) \right\|^2
\end{aligned}$$

If the following condition holds,

$$2L^2\tau^2\hat{\eta}^2C_\eta^2S\eta_l^2 + (L^2W_1^2C_\eta^2 + L\bar{\eta}C_\eta + C_\eta)S\eta_l \leq \frac{m}{n}$$

we have,

$$\begin{aligned}
& -\frac{\eta_S}{SW_2} \mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
& + \frac{\eta_l^2}{m^2} \left(\frac{\eta_0}{SW_2 \eta_l} + \frac{L\eta_0^2}{SW_2 \eta_l} + \frac{L^2 W_1^2 \eta_0 C_\beta}{SW_2 \eta_l} + \frac{2L^2 \tau^2 \eta_0^3 C_\beta}{SW_2} \right) \cdot \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
& \stackrel{(i)}{\leq} -\frac{\eta_S}{SW_2 n^2} \mathbb{E} \left[\left\| \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
& + \frac{\eta_S}{mn SW_2} \mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
& \stackrel{(ii)}{=} -\frac{\eta_S}{SW_2 n} \sum_{i=1}^n \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 + \frac{m\eta_S}{n^2 SW_2} \sum_{i=1}^n \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \\
& + \frac{(n-m)\eta_S}{2n^2 SW_2 (n-1)} \sum_{i \neq j} \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) - \frac{1}{K_{t,j}} \sum_{k=0}^{K_{t,j}-1} \nabla f_j(x_{t-\tau_{t,j},k}^j) \right\|^2 \\
& \stackrel{(iii)}{\leq} \left(-\frac{\eta_S}{SW_2 n} + \frac{m\eta_S}{n^2 SW_2} + \frac{(n-m)\eta_S}{n^2 SW_2} \right) \sum_{i=1}^n \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \stackrel{(iv)}{=} 0
\end{aligned}$$

where (i) holds by plugging in the assumption, $2L^2 \tau^2 \hat{\eta}^2 C_\eta^2 S \eta_l^2 + (L^2 W_1^2 C_\eta^2 + L\bar{\eta} C_\eta + C_\eta) S \eta_l \leq \frac{m}{n}$. (ii) holds by plugging in the equality for $\mathbb{E} \left[\left\| \sum_{i=1}^n \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]$ and $\mathbb{E} \left[\left\| \sum_{i=1}^n \mathbf{1}\{i \in \mathcal{S}_t\} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]$. (iii) holds as $\sum_{i \neq j} \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) - \frac{1}{K_{t,j}} \sum_{k=0}^{K_{t,j}-1} \nabla f_j(x_{t-\tau_{t,j},k}^j) \right\|^2 = 2(n-1) \sum_{i=1}^n \left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2$. (iv) holds as $-\frac{\eta_S}{SW_2 n} + \frac{m\eta_S}{n^2 SW_2} + \frac{(n-m)\eta_S}{n^2 SW_2} = 0$.

Merging all pieces together,

$$\begin{aligned}
& \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{4(f(x_0) - f^*)}{SW_2 \eta_l} + \frac{120\eta_l^2 L^2 T \bar{\eta} \phi_2}{W_2} \sigma_g^2 \\
& + \left(\frac{20\eta_l^2 L^2 T \bar{\eta}}{W_2} \phi_1 + \frac{4L^2 \tau^2 \hat{\eta}^3 \eta_l^2 T}{mW_2} \phi_3 + \frac{2L^2 W_1^2 \bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2\bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2L\hat{\eta}^2 \eta_l}{mW_2} \phi_3 \right) \sigma_l^2
\end{aligned}$$

Suppose $S = 1$, i.e. the typical constant hyperparameter regime, and further suppose local updating number as K , the total number of rounds as T , $\eta_0 = \bar{\eta} = \Theta(\sqrt{mK})$ and $\eta_l = \Theta(\frac{1}{\sqrt{T}})$. In this case, $\phi_1 = K$, $\phi_2 = K^2$, $\phi_3 = \frac{1}{K}$, $W_2 = \Theta(T\sqrt{mK})$. Assume $W_1^2 = \mathcal{O}(\sqrt{mK})$. We have the bound as,

We have the bounds as,

$$\begin{aligned}
& \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) (f(z_0) - f^*) + \\
& + \left(\mathcal{O}\left(\frac{K}{T}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{1}{mK\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) \right) \sigma_l^2 + \mathcal{O}\left(\frac{K^2}{T}\right) \sigma_g^2
\end{aligned}$$

Only keep the dominant terms, we could get,

$$\begin{aligned} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 &\leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) (f(z_0) - f^*) + \\ &+ \left(\mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right)\right) \sigma_l^2 + \mathcal{O}\left(\frac{K^2}{T}\right) \sigma_g^2 \end{aligned}$$

Suppose $S = \Theta(1)$, i.e. the multistage regime, the total number of rounds are T , $\bar{\eta} = \Theta(\sqrt{mK})$, $\hat{\eta}^2 = \Theta(mK)$, $\hat{\eta}^3 = \Theta\left(m^{\frac{3}{2}}K^{\frac{3}{2}}\right)$, and $\eta_l = \Theta\left(\frac{1}{\sqrt{T}}\right)$, $W_2 = \Theta\left(\frac{T\sqrt{mK}}{S}\right)$, i.e. $T\bar{\eta}$ is equally divided into S stages. Assume $W_1^2 = \mathcal{O}(\sqrt{mK})$. We have the bound as,

$$\begin{aligned} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 &\leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) (f(z_0) - f^*) + \\ &+ \left(\mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right)\right) \sigma_l^2 + \mathcal{O}\left(\frac{K^2}{T}\right) \sigma_g^2 \end{aligned}$$

In both cases, the rate is $\mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{K^2}{T}\right)$. □

G Proof of Corollary 5.4

Corollary G.1 (Formal Statement of Corollary 5.4). *We optimize $f(x)$ using Algorithm 4 (General Arrival) and $\{f_i\}_{i=1}^n$ fulfills Assumptions 1-3. Suppose bounded maximum delay, i.e. $\tau_{t,i} \leq \tau < \infty$ for any $i \in \mathcal{S}_t$ and $t \in \{1, \dots, T\}$. Denote $C_\eta \triangleq \frac{\eta_l}{\eta_s}$.*

Under the condition $\eta_l \leq \min\left\{\frac{1}{8K_{t,\max}L}, \sqrt{\frac{1}{180L^2C_\eta\tau K_{t,\max}^2}}\right\}$. We would have:

$$\bar{g} \leq \frac{4(f(z_0) - f^*)}{SW_2\eta_l} + \Phi_l\sigma_l^2 + \Phi_g\sigma_g^2$$

where we define $\bar{\eta} \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \eta_s$ (average server learning rate), $\hat{\eta}^2 \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^2$, $\hat{\eta}^3 \triangleq \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^3$, $\frac{1}{K_t} = \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}}$, $\bar{K}_t \triangleq \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}$, $\hat{K}_t^2 \triangleq \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}^2$, $\phi_1 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \bar{K}_t$, $\phi_2 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \hat{K}_t^2$, and $\phi_3 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{K_t}$, for ease of notation. And $\Phi_l \triangleq \frac{30\eta_l^2 L^2 \phi_1 T \bar{\eta}}{W_2} + \frac{6L^2 \tau^2 \hat{\eta}^3 \eta_l^2 T}{mW_2} \phi_3 + \frac{2L^2 W_1^2 \bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2\bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2L \hat{\eta}^2 \eta_l T}{mW_2} \phi_3$ and $\Phi_g \triangleq 6 + \frac{180\eta_l^2 L^2 T \bar{\eta} \phi_2}{W_2}$.

$$\begin{aligned} \Phi_l &\triangleq \frac{30\eta_l^2 L^2 \phi_1 T \bar{\eta}}{W_2} + \frac{6L^2 \tau^2 \hat{\eta}^3 \eta_l^2 T}{mW_2} \phi_3 + \frac{2L^2 W_1^2 \bar{\eta} \eta_l T}{mW_2} \phi_3 \\ &\quad + \frac{2\bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2L \hat{\eta}^2 \eta_l T}{mW_2} \phi_3 \\ \Phi_g &\triangleq 6 + \frac{180\eta_l^2 L^2 T \bar{\eta} \phi_2}{W_2} \end{aligned}$$

Suppose $S = \Theta(1)$, i.e. the multistage regime, the total number of rounds are T , $\bar{\eta} = \Theta(\sqrt{mK})$, $\hat{\eta}^2 = \Theta(mK)$, $\hat{\eta}^3 = \Theta\left(m^{\frac{3}{2}}K^{\frac{3}{2}}\right)$, and $\eta_l = \Theta\left(\frac{1}{\sqrt{T}}\right)$, $W_2 = \Theta\left(\frac{T\sqrt{mK}}{S}\right)$, i.e. $T\bar{\eta}$ is equally divided into S stages. Assume $W_1^2 = \mathcal{O}(\sqrt{mK})$. We have the bound as,

$$\bar{g} \leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{K^2}{T}\right) + \mathcal{O}(\sigma_g^2)$$

Proof of Multistage GM with General Arrival. We introduce a Lyapunov sequence $\{z_t\}_{t=0}^{T-1}$ which is devised as follows:

$$z_t = x_t - \frac{\eta_t \beta_t \nu_t}{1 - \beta_t} d_t \tag{9}$$

where $d_0 = 0$.

We could easily verify $z_{t+1} - z_t = -\eta_t \Delta_t$. Let $-\eta y_t = x_{t+1} - x_t$, we first bound $\mathbb{E} \left[\|\Delta_t\|^2 \right]$, $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$, and $\sum_{t=0}^{T-1} \mathbb{E} \left[\|y_t\|^2 \right]$.

Bounding $\mathbb{E} \left[\|\Delta_t\|^2 \right]$:

$$\begin{aligned}
\mathbb{E} \left[\|\Delta_t\|^2 \right] &= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \Delta_{t-\tau_t, i}^i \right\|^2 \right] \\
&\stackrel{(i)}{=} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} g_{t-\tau_t, i, k}^i \right\|^2 \right] \\
&= \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \left(g_{t-\tau_t, i, k}^i - \nabla f_i(x_{t-\tau_t, i, k}^i) + \nabla f_i(x_{t-\tau_t, i, k}^i) \right) \right\|^2 \right] \\
&\stackrel{(ii)}{=} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \left\{ g_{t-\tau_t, i, k}^i - \nabla f_i(x_{t-\tau_t, i, k}^i) \right\} \right\|^2 \right] + \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_t, i, k}^i) \right\|^2 \right] \\
&\stackrel{(iii)}{\leq} \frac{1}{m^2} \sum_{i \in \mathcal{S}_t} \frac{\eta_l^2}{K_{t,i}^2} \sum_{k=0}^{K_{t,i}-1} \sigma_l^2 + \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{\eta_l}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_t, i, k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_t, i, k}^i) \right\|^2 \right]
\end{aligned}$$

where $\frac{1}{K_t} = \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}}$. (i) follows from the definition of $\Delta_{t-\tau_t, i}^i$, (ii) and (iii) hold as $\mathbb{E} \left[\left\| \sum_{i=1}^n x_i \right\|^2 \right] = \sum_{i=1}^n \mathbb{E} \left[\|x_i\|^2 \right]$ when $\mathbb{E} [x_i] = 0$, and we know $\mathbb{E} \left[g_{t-\tau_t, i, k}^i - \nabla f_i(x_{t-\tau_t, i, k}^i) \right] = 0$.

Bounding $\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right]$:

We could verify:

$$d_t = \sum_{p=0}^t a_{t,p} \Delta_p, \quad \text{where } a_{t,p} = (1 - \beta_p) \prod_{q=p+1}^t \beta_q$$

We further get,

$$\begin{aligned}
\mathbb{E} \left[\|d_t\|^2 \right] &= \mathbb{E} \left[\left\| \sum_{p=0}^t a_{t,p} \Delta_p \right\|^2 \right] \\
&\leq \sum_{e=1}^d \mathbb{E} \left[\sum_{p=0}^t a_{t,p} \Delta_{p,e} \right]^2 \leq \sum_{e=1}^d \mathbb{E} \left[\left(\sum_{p=0}^t a_{t,p} \right) \left(\sum_{p=0}^t a_{t,p} \Delta_{p,e}^2 \right) \right] \leq \left(1 - \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t a_{t,p} \mathbb{E} \left[\|\Delta_p\|^2 \right] \\
&\leq \left(1 - \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t a_{t,p} \left\{ \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_p, i, k}^i) \right\|^2 \right] \right\} \\
&\leq \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^t a_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_p, i, k}^i) \right\|^2 \right]
\end{aligned}$$

Summing over $t \in \{0, 1, \dots, T-1\}$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] &\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{t=0}^{T-1} \sum_{p=0}^t a_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^{T-1} \left(\sum_{t=p}^{T-1} a_{t,p} \right) \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Bounding $\sum_{t=0}^{T-1} \mathbb{E} \left[\|y_t\|^2 \right]$:

We could verify:

$$y_t = \sum_{p=0}^t b_{t,p} \Delta_p,$$

where $b_{t,p}$ is defined as follows,

$$b_{t,p} = \begin{cases} 1 - \beta_t \nu_t & p = t \\ \nu_t (1 - \beta_p) \prod_{q=p+1}^t \beta_q & p < t \end{cases}$$

We further get,

$$\begin{aligned}
\mathbb{E} \left[\|y_t\|^2 \right] &= \mathbb{E} \left[\left\| \sum_{p=0}^t b_{t,p} \Delta_p \right\|^2 \right] \\
&\leq \sum_{e=1}^d \mathbb{E} \left[\sum_{p=0}^t b_{t,p} \Delta_{p,e} \right]^2 \leq \sum_{e=1}^d \mathbb{E} \left[\left(\sum_{p=0}^t b_{t,p} \right) \left(\sum_{p=0}^t b_{t,p} \Delta_{p,e}^2 \right) \right] \leq \left(1 - \nu_t \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t b_{t,p} \mathbb{E} \left[\|\Delta_p\|^2 \right] \\
&\leq \left(1 - \nu_t \prod_{q=0}^t \beta_q \right) \sum_{p=0}^t b_{t,p} \left\{ \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \right\} \\
&\leq \frac{\eta_l^2}{m} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^t b_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Summing over $t \in \{0, 1, \dots, T-1\}$,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{E} \left[\|y_t\|^2 \right] &\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{t=0}^{T-1} \sum_{p=0}^t b_{t,p} \cdot \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} \sum_{p=0}^{T-1} \left(\sum_{t=p}^{T-1} b_{t,p} \right) \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{p,i}} \sum_{k=0}^{K_{p,i}-1} \nabla f_i(x_{p-\tau_{p,i},k}^i) \right\|^2 \right] \\
&\leq \frac{\eta_l^2}{m} \sum_{t=0}^{T-1} \frac{1}{K_t} \sigma_l^2 + \frac{\eta_l^2}{m^2} C_\beta \sum_{t=0}^{T-1} \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

Since f is L -smooth, taking conditional expectation with respect to all randomness prior to step t , we have

$$\begin{aligned}
\mathbb{E}[f(z_{t+1})] &\leq f(z_t) + \mathbb{E}[\langle \nabla f(z_t), z_{t+1} - z_t \rangle] + \frac{L}{2} \mathbb{E}[\|z_{t+1} - z_t\|^2] \\
&\leq f(z_t) + \mathbb{E}[\langle \nabla f(z_t), -\eta_t \Delta_t \rangle] + \frac{L}{2} \eta_t^2 \mathbb{E}[\|\Delta_t\|^2] \\
&\leq f(z_t) + \underbrace{\mathbb{E}[\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle]}_{A_1} + \underbrace{\mathbb{E}[\langle \nabla f(x_t), -\eta_t \Delta_t \rangle]}_{A_2} + \underbrace{\frac{L}{2} \eta_t^2 \mathbb{E}[\|\Delta_t\|^2]}_{A_3}
\end{aligned}$$

Bounding A_1 :

$$\begin{aligned}
A_1 &= \mathbb{E}[\langle \sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t)), -\sqrt{\eta_t} \Delta_t \rangle] \\
&\stackrel{(i)}{\leq} \mathbb{E}[\|\sqrt{\eta_t} (\nabla f(z_t) - \nabla f(x_t))\| \cdot \|\sqrt{\eta_t} \Delta_t\|] \\
&\stackrel{(ii)}{\leq} \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E}[\|d_t\|^2] + \frac{1}{2} \eta_t \mathbb{E}[\|\Delta_t\|^2]
\end{aligned}$$

where (i) holds by applying Cauchy-Schwarz inequality, and (ii) follows from Young's inequality and f is L -smooth.

Bounding A_2 :

$$\begin{aligned}
A_2 &= \mathbb{E}[\langle \nabla f(x_t), -\eta_t \Delta_t \rangle] \\
&= \eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t - \eta_t \nabla f(x_t) \rangle] \\
&= -\eta_t \mathbb{E}[\|\nabla f(x_t)\|^2] + \eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t \rangle]
\end{aligned}$$

where we further bound $\eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t \rangle]$,

$$\begin{aligned}
\eta_t \mathbb{E}[\langle \nabla f(x_t), \eta_t \nabla f(x_t) - \Delta_t \rangle] &= \eta_t \mathbb{E} \left[\left\langle \sqrt{\eta_t} \nabla f(x_t), \frac{\sqrt{\eta_t}}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f(x_t) - g_{t-\tau_{t,i},k}^i) \right\rangle \right] \\
&\stackrel{(i)}{=} \eta_t \mathbb{E} \left[\left\langle \sqrt{\eta_t} \nabla f(x_t), \frac{\sqrt{\eta_t}}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f(x_t) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\rangle \right] \\
&\stackrel{(ii)}{=} \frac{\eta_t \eta_t}{2} \mathbb{E}[\|\nabla f(x_t)\|^2] - \frac{\eta_t \eta_t}{2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&\quad + \frac{\eta_t \eta_t}{2} \mathbb{E} \left[\left\| \nabla f(x_t) - \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]
\end{aligned}$$

where (i) holds as we take conditional expectation with respect to all randomness prior to step t . (ii) holds as $\langle a, b \rangle = \frac{1}{2} \|a\|^2 + \frac{1}{2} \|b\|^2 - \frac{1}{2} \|a - b\|^2$.

We further have,

$$\begin{aligned}
& \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \nabla f(x_t) - \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&= \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f(x_t) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\|^2 \right] \\
&\stackrel{(i)}{\leq} \frac{3}{2} \eta_t \eta_l \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} (\nabla f(x_t) - \nabla f_i(x_t)) \right\|^2 \right] + \frac{3}{2} \eta_t \eta_l \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} (\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i}})) \right\|^2 \right] \\
&\quad + \frac{3}{2} \eta_t \eta_l \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} (\nabla f_i(x_{t-\tau_{t,i}}) - \nabla f_i(x_{t-\tau_{t,i},k}^i)) \right\|^2 \right] \\
&\stackrel{(ii)}{\leq} \frac{3}{2} \eta_t \eta_l \frac{1}{m} \sum_{i \in \mathcal{S}_t} \mathbb{E} [\|\nabla f(x_t) - \nabla f_i(x_t)\|^2] + \frac{3}{2} \eta_t \eta_l \frac{1}{m} \sum_{i \in \mathcal{S}_t} \mathbb{E} [\|\nabla f_i(x_t) - \nabla f_i(x_{t-\tau_{t,i}})\|^2] \\
&\quad + \frac{3}{2} \eta_t \eta_l \frac{1}{m} \sum_{i \in \mathcal{S}_t} \mathbb{E} \left[\left\| \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i}}) - \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&\stackrel{(iii)}{\leq} \frac{3}{2} \eta_t \eta_l \sigma_g^2 + \frac{3\eta_t \eta_l L^2}{2m} \sum_{i \in \mathcal{S}_t} \mathbb{E} [\|x_t - x_{t-\tau_{t,i}}\|^2] + \frac{3\eta_t \eta_l L^2}{2m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \mathbb{E} [\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2]
\end{aligned}$$

where (i) and (ii) hold as $\|\sum_{i=1}^n x_i\|^2 \leq n \sum_{i=1}^n \|x_i\|^2$, (iii) holds as f_i is L -smooth. Thus, we have,

$$\begin{aligned}
A_2 &\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} [\|\nabla f(x_t)\|^2] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&\quad + \frac{3}{2} \eta_t \eta_l \sigma_g^2 + \frac{3\eta_t \eta_l L^2}{2m} \sum_{i \in \mathcal{S}_t} \mathbb{E} [\|x_t - x_{t-\tau_{t,i}}\|^2] + \frac{3\eta_t \eta_l L^2}{2m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \mathbb{E} [\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2]
\end{aligned}$$

When $\eta_l \leq \frac{1}{8K_{t,i}L}$, we have,

$$\mathbb{E} \left[\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2 \right] \leq 5K_{t,i}\eta_l^2 (\sigma_l^2 + 6K_{t,i}\sigma_g^2) + 30K_{t,i}^2\eta_l^2 \mathbb{E} [\|\nabla f(x_{t-\tau_{t,i}})\|^2]$$

We can further bound $\frac{1}{m} \sum_{i \in \mathcal{S}_t} \mathbb{E} [\|x_t - x_{t-\tau_{t,i}}\|^2]$

$$\begin{aligned}
\frac{1}{m} \sum_{i \in \mathcal{S}_t} \mathbb{E} [\|x_t - x_{t-\tau_{t,i}}\|^2] &\stackrel{(i)}{\leq} \mathbb{E} [\|x_t - x_{t-\tau_{t,u}}\|^2] = \mathbb{E} \left[\left\| \sum_{k=t-\tau_{t,u}}^{t-1} (x_{k+1} - x_k) \right\|^2 \right] \\
&\stackrel{(ii)}{\leq} \mathbb{E} \left[\left\| \sum_{k=t-\tau_{t,u}}^{t-1} \eta_k y_k \right\|^2 \right] \stackrel{(iii)}{\leq} \tau \eta_{t-\tau_{t,u}}^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} [\|y_k\|^2]
\end{aligned}$$

where (i) holds as we define $u = \arg \max_{i \in \{1,2,\dots,n\}} \mathbb{E} [\|x_t - x_{t-\tau_{t,i}}\|^2]$, (ii) follows from the definition of y_k , (iii) holds as bounded maximum delay assumption, i.e. $\tau_{t,i} \leq \tau$ for any t and i , and learning rate is decaying, i.e. $\eta_t \leq \eta_{t-\tau_{t,u}}$.

Merging all pieces together, we have the following,

$$\begin{aligned}
A_2 &\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
&\frac{3}{2} \eta_t \eta_l \sigma_g^2 + \frac{3\eta_t \eta_l L^2}{2m} \sum_{i \in \mathcal{S}_t} \mathbb{E} \left[\|x_t - x_{t-\tau_{t,i}}\|^2 \right] + \frac{3\eta_t \eta_l L^2}{2m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \mathbb{E} \left[\|x_{t-\tau_{t,i}} - x_{t-\tau_{t,i},k}^i\|^2 \right] \\
&\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + \frac{3}{2} \eta_t \eta_l \sigma_g^2 \\
&+ \frac{15}{2} L^2 \eta_t \eta_l^3 \bar{K}_t \sigma_i^2 + 45 L^2 \eta_t \eta_l^3 \hat{K}_t^2 \sigma_g^2 + 45 L^2 \eta_t \eta_l^3 \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] \\
&\quad + \frac{3}{2} \tau L^2 \eta_t \eta_{t-\tau_{t,u}}^2 \eta_l \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right]
\end{aligned}$$

where $\bar{K}_t \triangleq \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}$ and $\hat{K}_t^2 \triangleq \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}^2$.

Plug all pieces back in $\mathbb{E}[f(z_{t+1})] \leq f(z_t) + A_1 + A_2 + A_3$,

$$\begin{aligned}
\mathbb{E}[f(z_{t+1})] - f(z_t) &\leq -\frac{\eta_t \eta_l}{2} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] - \frac{\eta_t \eta_l}{2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + \frac{3}{2} \eta_t \eta_l \sigma_g^2 \\
&+ \frac{15}{2} L^2 \eta_t \eta_l^3 \bar{K}_t \sigma_i^2 + 45 L^2 \eta_t \eta_l^3 \hat{K}_t^2 \sigma_g^2 + 45 L^2 \eta_t \eta_l^3 \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] \\
&\quad + \frac{3}{2} \tau L^2 \eta_t \eta_{t-\tau_{t,u}}^2 \eta_l \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] \\
&\quad + \frac{1}{2} \eta_t^3 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E} \left[\|d_t\|^2 \right] + \frac{1}{2} \eta_t \mathbb{E} \left[\|\Delta_t\|^2 \right] + \frac{L}{2} \eta_t^2 \mathbb{E} \left[\|\Delta_t\|^2 \right]
\end{aligned}$$

Reorganizing terms and we have,

$$\begin{aligned}
\mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] &\leq \frac{2(f(z_t) - \mathbb{E}[f(z_{t+1})])}{\eta_t \eta_l} - \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + 3\sigma_g^2 \\
&+ 15 L^2 \eta_l^2 \bar{K}_t \sigma_i^2 + 90 L^2 \eta_l^2 \hat{K}_t^2 \sigma_g^2 + 90 L^2 \eta_l^2 \frac{1}{m} \sum_{i \in \mathcal{S}_t} K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] + 3\tau L^2 \eta_{t-\tau_{t,u}}^2 \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] \\
&\quad + \frac{1}{\eta_l} \eta_t^2 L^2 \left(\frac{\beta_t \nu_t}{1 - \beta_t} \right)^2 \mathbb{E} \left[\|d_t\|^2 \right] + \frac{1}{\eta_l} \mathbb{E} \left[\|\Delta_t\|^2 \right] + \frac{L}{\eta_l} \eta_t \mathbb{E} \left[\|\Delta_t\|^2 \right]
\end{aligned}$$

Sum over all S stages and take average, by some algebraic transformations, we get,

$$\begin{aligned}
\bar{G} &\triangleq \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(z_0) - \mathbb{E}[f(z_T)])}{SW_2\eta_l} \\
&- \frac{\eta_S}{SW_2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + 3\sigma_g^2 + \frac{3L^2\tau\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] \\
&\quad + \frac{90\eta_l^2 L^2}{m} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \sum_{i \in \mathcal{S}_t} K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] \\
&\quad + \frac{L^2 W_1^2 \bar{\eta}}{W_2 \eta_l} \sum_{t=0}^{T-1} \mathbb{E} \left[\|d_t\|^2 \right] + \frac{\bar{\eta}}{W_2 \eta_l} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\Delta_t\|^2 \right] + \frac{L\hat{\eta}^2}{SW_2\eta_l} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\Delta_t\|^2 \right] \\
&\quad + 15\eta_l^2 L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \right\} \sigma_l^2 + 90\eta_l^2 L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \hat{K}_t \right\} \sigma_g^2
\end{aligned}$$

where $\bar{\eta} = \frac{1}{S} \sum_{s=0}^{S-1} \eta_s$, $\hat{\eta}^2 = \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^2$, and $\hat{\eta}^3 = \frac{1}{S} \sum_{s=0}^{S-1} \eta_s^3$, respectively.

When the following holds,

$$\eta \leq \sqrt{\frac{1}{180L^2 C_\eta \tau K_{t,\max}^2}}, \quad \forall t \in \{0, \dots, T-1\}$$

where $C_\eta = \frac{\eta_0}{\eta_S}$.

we could verify the following inequality,

$$\begin{aligned}
&\frac{90\eta_l^2 L^2}{m} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \sum_{i \in \mathcal{S}_t} K_{t,i}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] \\
&\leq_{(i)} 90\eta_l^2 L^2 \frac{\eta_0}{SW_2} \sum_{t=0}^{T-1} K_{t,\max}^2 \mathbb{E} \left[\|\nabla f(x_{t-\tau_{t,i}})\|^2 \right] \\
&\leq_{(ii)} 90\eta_l^2 L^2 \tau \frac{\eta_0}{SW_2} \sum_{t=0}^{T-1} K_{t,\max}^2 \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq_{(iii)} \frac{\eta_S}{2SW_2} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \\
&\leq_{(iv)} \frac{1}{2} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2
\end{aligned}$$

where (i) follows from the definition of $K_{t,\max}^2 = \max_{i \in \{1,2,\dots,n\}} K_{t,i}^2$, and $W_2 = \eta_s T_s$ for all $s \in \{1, \dots, S\}$ and $\eta_S \leq \eta_s \leq \eta_0$. (ii) follows from the maximum delay assumption. (iii) holds by plugging in the assumption $\eta \leq \sqrt{\frac{\eta_S}{180L^2 \eta_0 \tau K_{t,\max}^2}}$, $\forall t \in \{0, \dots, T-1\}$. (iv) holds as $\frac{\eta_S}{2SW_2} \sum_{t=0}^{T-1} \mathbb{E} \left[\|\nabla f(x_t)\|^2 \right] \leq \frac{\eta_S}{2} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s \eta_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2$ and $\frac{1}{\eta_s} \leq \frac{1}{\eta_S}$ for all s .

With the maximum delay assumption, we have $\frac{3L^2\tau\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \sum_{k=t-\tau_{t,u}}^{t-1} \mathbb{E} \left[\|y_k\|^2 \right] \leq \frac{3L^2\tau\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \mathbb{E} \left[\|y_t\|^2 \right]$. Merging all pieces, we have,

$$\begin{aligned}
& \frac{1}{2} \cdot \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(z_0) - \mathbb{E}[f(z_T)])}{SW_2\eta_l} \\
& - \frac{\eta_S}{SW_2} \mathbb{E} \left[\left\| \frac{1}{m} \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] + 3\sigma_g^2 + \frac{3L^2\tau^2\hat{\eta}^3}{W_2} \sum_{t=0}^{T-1} \mathbb{E}[\|y_t\|^2] \\
& + \frac{L^2W_1^2\bar{\eta}}{W_2\eta_l} \sum_{t=0}^{T-1} \mathbb{E}[\|d_t\|^2] + \frac{\bar{\eta}}{W_2\eta_l} \sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] + \frac{L\hat{\eta}^2}{W_2\eta_l} \sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2] \\
& + 15\eta_l^2L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \right\} \sigma_l^2 + 90\eta_l^2L^2 \left\{ \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \hat{K}_t^2 \right\} \sigma_g^2
\end{aligned}$$

We define ϕ_1 , ϕ_2 , and ϕ_3 for ease of notation.

$$\phi_1 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \bar{K}_t, \quad \text{and} \quad \phi_2 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \hat{K}_t^2, \quad \text{and} \quad \phi_3 \triangleq \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{\bar{K}_t}$$

We could verify,

$$\frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \stackrel{(i)}{\leq} \frac{1}{W_2} \frac{1}{S} \sum_{s=0}^{S-1} \eta_s \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \stackrel{(ii)}{\leq} \frac{\bar{\eta}}{W_2} \sum_{t=0}^{T-1} \bar{K}_t = \frac{T\bar{\eta}}{W_2} \phi_1$$

(i) holds due to $T_s\eta_s = W_2$ by assumption, (ii) holds due to $\frac{1}{S} \sum_{s=0}^{S-1} \eta_s \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t \leq \left(\frac{1}{S} \sum_{s=0}^{S-1} \eta_s\right) \cdot \left(\sum_{s=0}^{S-1} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \bar{K}_t\right) = \bar{\eta} \sum_{t=0}^{T-1} \bar{K}_t$.

Similarly, we have,

$$\frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \hat{K}_t^2 \leq \frac{T\bar{\eta}}{W_2} \phi_2, \quad \text{and} \quad \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \frac{1}{\bar{K}_t} \leq \frac{T\bar{\eta}}{W_2} \phi_3$$

Plugging in the bounds for $\sum_{t=0}^{T-1} \mathbb{E}[\|\Delta_t\|^2]$, $\sum_{t=0}^{T-1} \mathbb{E}[\|y_t\|^2]$, and $\sum_{t=0}^{T-1} \mathbb{E}[\|d_t\|^2]$,

$$\begin{aligned}
& \frac{1}{2} \cdot \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(z_0) - \mathbb{E}[f(z_T)])}{SW_2\eta_l} + \\
& \left(-\frac{\eta_S}{SW_2m^2} + \frac{3L^2\tau^2\hat{\eta}^3\eta_l^2C_\beta}{W_2m^2} + \frac{L^2W_1^2\bar{\eta}\eta_lC_\beta}{W_2m^2} + \frac{\bar{\eta}\eta_l}{W_2m^2} + \frac{L\hat{\eta}^2\eta_l}{W_2m^2} \right) \\
& \cdot \mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right] \\
& + \left(\frac{15\eta_l^2L^2\phi_1T\bar{\eta}}{W_2} + \frac{3L^2\tau^2\hat{\eta}^3\eta_l^2T}{mW_2} \phi_3 + \frac{L^2W_1^2\bar{\eta}\eta_lT}{mW_2} \phi_3 + \frac{\bar{\eta}\eta_lT}{mW_2} \phi_3 + \frac{L\hat{\eta}^2\eta_lT}{mW_2} \phi_3 \right) \sigma_l^2 \\
& \left(3 + \frac{90\eta_l^2L^2T\bar{\eta}\phi_2}{W_2} \right) \sigma_g^2
\end{aligned}$$

We could verify, when the following condition holds,

$$3L^2S\tau^2\hat{\eta}_0^2C_\eta^2\eta_l^2 + C_\eta S(L^2W_1^2C_\eta + 1 + L\bar{\eta})\eta_l \leq 1$$

we have the coefficient for $\mathbb{E} \left[\left\| \sum_{i \in \mathcal{S}_t} \frac{1}{K_{t,i}} \sum_{k=0}^{K_{t,i}-1} \nabla f_i(x_{t-\tau_{t,i},k}^i) \right\|^2 \right]$, i.e.,

$$-\frac{\eta_S}{SW_2m^2} + \frac{3L^2\tau^2\hat{\eta}^3\eta_l^2C_\beta}{W_2m^2} + \frac{L^2W_1^2\bar{\eta}\eta_lC_\beta}{W_2m^2} + \frac{\bar{\eta}\eta_l}{W_2m^2} + \frac{L\hat{\eta}^2\eta_l}{W_2m^2} \leq 0$$

Therefore, we have,

$$\begin{aligned} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 &\leq \frac{4(f(z_0) - f^*)}{SW_2\eta_l} + \\ &+ \left(\frac{30\eta_l^2 L^2 \phi_1 T \bar{\eta}}{W_2} + \frac{6L^2 \tau^2 \hat{\eta}^3 \eta_l^2 T}{mW_2} \phi_3 + \frac{2L^2 W_1^2 \bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2\bar{\eta} \eta_l T}{mW_2} \phi_3 + \frac{2L\hat{\eta}^2 \eta_l T}{mW_2} \phi_3 \right) \sigma_l^2 \\ &\quad \left(6 + \frac{180\eta_l^2 L^2 T \bar{\eta} \phi_2}{W_2} \right) \sigma_g^2 \end{aligned}$$

Suppose $S = 1$, i.e. the typical constant hyperparameter regime, and further suppose local updating number as K , the total number of rounds as T , $\eta_0 = \bar{\eta} = \Theta(\sqrt{mK})$ and $\eta_l = \Theta(\frac{1}{\sqrt{T}})$. In this case, $\phi_1 = K$, $\phi_2 = K^2$, $\phi_3 = \frac{1}{K}$, $W_2 = \Theta(T\sqrt{mK})$. Suppose $W_1^2 = \mathcal{O}(\sqrt{mK})$. We have the bound as,

We have the bounds as,

$$\begin{aligned} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 &\leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) (f(z_0) - f^*) + \\ &+ \left(\mathcal{O}\left(\frac{K}{T}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{1}{mK\sqrt{T}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) \right) \sigma_l^2 + \left(6 + \mathcal{O}\left(\frac{K^2}{T}\right) \right) \sigma_g^2 \end{aligned}$$

Only keep the dominant terms, we could get,

$$\begin{aligned} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 &\leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) (f(z_0) - f^*) + \\ &+ \left(\mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) \right) \sigma_l^2 + \left(6 + \mathcal{O}\left(\frac{K^2}{T}\right) \right) \sigma_g^2 \end{aligned}$$

Suppose $S = \Theta(1)$, i.e. the multistage regime, the total number of rounds are T , $\bar{\eta} = \Theta(\sqrt{mK})$, $\hat{\eta}^2 = \Theta(mK)$, $\hat{\eta}^3 = \Theta(m^{\frac{3}{2}}K^{\frac{3}{2}})$, and $\eta_l = \Theta(\frac{1}{\sqrt{T}})$, $W_2 = \Theta(\frac{T\sqrt{mK}}{S})$, i.e. $T\bar{\eta}$ is equally divided into S stages, suppose $W_1^2 = \mathcal{O}(\sqrt{mK})$, we have the bound as,

$$\begin{aligned} \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{T_s} \sum_{t=T_0+\dots+T_{s-1}}^{T_0+\dots+T_s-1} \|\nabla f(x_t)\|^2 &\leq \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) (f(z_0) - f^*) + \\ &+ \left(\mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) \right) \sigma_l^2 + \left(6 + \mathcal{O}\left(\frac{K^2}{T}\right) \right) \sigma_g^2 \end{aligned}$$

The bound is thus, $\mathcal{O}\left(\frac{1}{\sqrt{mKT}}\right) + \mathcal{O}\left(\frac{\tau^2}{T}\right) + \mathcal{O}\left(\frac{K^2}{T}\right) + \mathcal{O}(\sigma_g^2)$.

□

H Experiments

H.1 Experimental Settings

We test how the performances of our proposed algorithms compared to FedAvg baseline in different settings. We train ResNet (He et al. 2016) and VGG (Simonyan and Zisserman 2015) on CIFAR10 (Krizhevsky 2009). To simulate data heterogeneity in CIFAR-10, we impose label imbalance across clients, i.e. each client is allocated a proportion of the samples of each label according to a Dirichlet distribution. Same procedure has been taken by (Hsu, Qi, and Brown 2019; Yurochkin et al. 2019; Wang et al. 2020; Li et al. 2022). The concentration parameter $\alpha > 0$ indicates the level of *non-i.i.d.*, with a smaller α implies higher heterogeneity, and $\alpha \rightarrow \infty$ implies *i.i.d.* setting.

How Asynchrony and Heterogeneous Local Epochs Are Implemented in Autonomous FedGM?

To simulate the asynchrony, we allow each worker to select one global model randomly from the last recent 5 global models instead of only using the current round's model in vanilla FedAvg. To simulate the heterogeneous local epochs, we allow each worker to randomly select local epoch number from $\{1, 2, \dots, 6\}$ at each round so that each worker has a time-varying, device dependent local epoch. Note in vanilla FedAvg, we fix the local epoch as 3.

Unless specified otherwise, we have the following default experimental settings,

Table 1: Default Experimental Settings

Number of Clients: 100	Participation Ratio: 0.05
Concentration Parameter: $\alpha = 0.5$	Local Epoch: 3
Local Learning Rate: $\eta_l = 0.01$	Total Number of Rounds: 500
η Grid: $\{0.5, 1.0, 1.5, \dots, 5.0\}$	β Grid: $\{0.7, 0.9, 0.95\}$
ν Grid: $\{0.7, 0.9, 0.95\}$	Local Momentum: Disabled

H.2 More Experiments in Section 6.1

Different Model Architecture and Levels of Heterogeneity

Figure 3 shows the results for VGG on CIFAR-10 with FedGM, FedAvgM, and FedAvg. We perform grid search over $\eta \in \{0.5, 1.0, 1.5, \dots, 5.0\}$, $\beta \in \{0.7, 0.9, 0.95\}$, and $\nu \in \{0.7, 0.9, 0.95\}$. We report the curves with best final test accuracy after 500 rounds. We could observe FedGM outperforms FedAvgM and FedAvg in both training and testing, which again verifies our claim that general momentum is a more capable algorithm compared to FedAvgM.

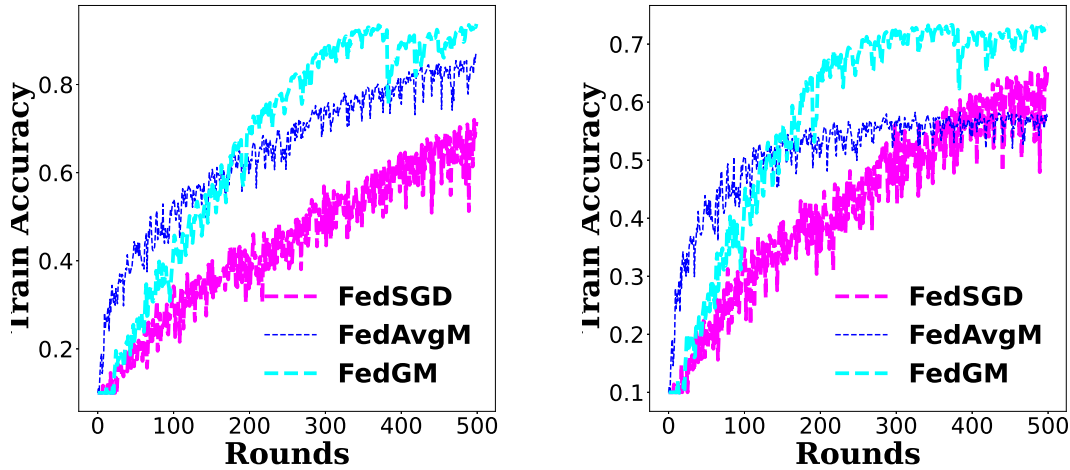


Figure 3: 3(a) Training and 3(b) Testing Curve for VGG on CIFAR-10.

Figure 4 shows the results for ResNet on CIFAR-10 with FedGM and FedAvg with different concentration parameters $\alpha = 0.3$ and $\alpha = 0.5$ (i.e. *non-i.i.d.*). We perform a similar grid search as in Section 6.1. We could observe the superiority of FedGM compared to FedAvg is consistent with different levels of *non i.i.d.*.

Verifying Remark 4.5

Remark 4.5 hypothesizes FedGM could converge with a large η while FedAvg would diverge easily with an only moderately large server learning rate. The reason is that η acts like a multiplier to client learning rate η_l in FedAvg, while in FedGM, β and ν act as a buffer that could absorb the impact from a large η . We verify this remark here.

Figure 5 shows the results for ResNet on CIFAR-10 with FedAvg but different learning rates $\eta = 1.0$, $\eta = 2.0$, and $\eta = 3.0$. We could see FedAvg experiences an unstable convergence even when $\eta = 2.0$ and completely divergent when $\eta = 3.0$.

Figure 6 shows the results for FedGM but different learning rates $\eta = 1.0$, $\eta = 3.0$, and $\eta = 5.0$. All experimental settings are identical to Figure 5 except for the difference between FedAvg and FedGM. We could see FedGM sustains a much larger η compared to FedAvg. It could converge and even accelerate with $\eta = 5.0$ compared to FedAvg baseline.

H.3 More Experiments in Section 6.2

Figure 7 motivates our multistage FedGM. We run ResNet on CIFAR-10 with FedGM but different learning rates $\eta = 1.0$, $\eta = 2.0$, and $\eta = 5.0$ for 2000 rounds. We fix $\beta = \nu = 0.95$ for expository purpose. We could see in early rounds (i.e. the first 500 rounds), $\eta = 5.0$ has advantages that it converges faster than small $\eta = 1.0$. However, $\eta = 1.0$ is much more stable than $\eta = 5.0$ in the last 500 rounds when they all get nearly perfect training accuracy. This is consistent with the motivation of multistage FedGM, i.e. large initial η benefits exploration, while small later η benefits exploitation, and multistage scheduler obtains a balance.

Figure 8 presents the results of running multistage FedGM for 2000 rounds, to see whether the advantage of multistage disappears with a longer training time. The two black vertical lines at round 286 and 857 mark the end of 1st/2nd stage. As we

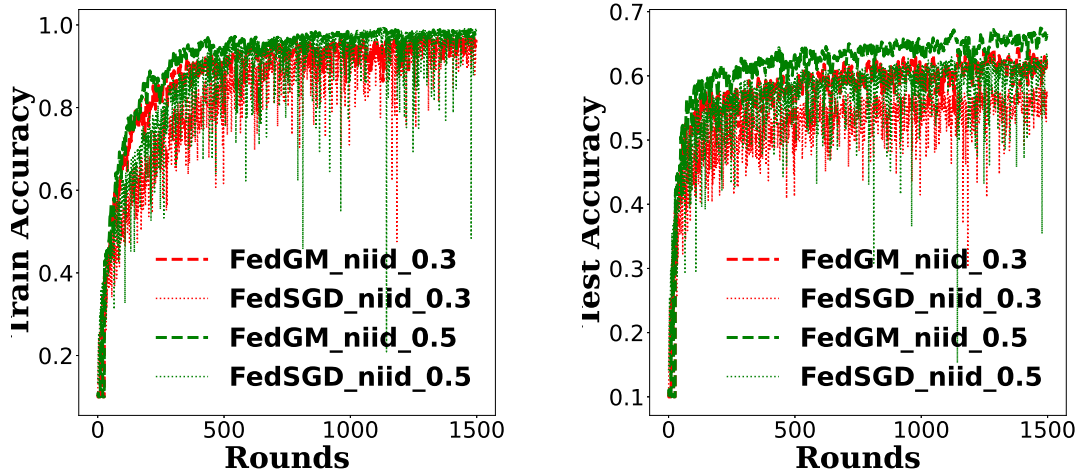


Figure 4: 4(a) Training and 4(b) Testing Curve for ResNet on CIFAR-10 with Various Levels of Heterogeneity

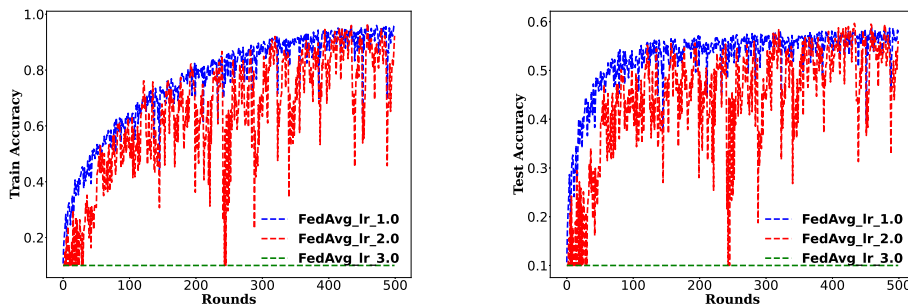


Figure 5: 5(a) Training and 5(b) Testing Curve for FedAvg with various server learning rates η .

could observe from Figure 8, the superiority of multistage FedGM is consistent with longer training time.

H.4 More Experiments in Section 6.3

Figure 9 shows the results for ResNet on CIFAR-10 with Autonomous FedGM and Autonomous FedAvg. The experimental settings are exactly same as Figure 1 except the random delay is 10 instead of 5. Specifically, in Figure ?? we allow each worker to select one global model randomly from the last recent 5 global models, while in Figure 1 we allow each worker to select one global model randomly from the last recent 10 global models. The objective is to mimic different levels of asynchrony. We report the curves with best final test accuracy. We plot a FedGM (i.e. no random delay and identical local epochs) as a baseline. Similarly as Figure 1, we observe momentum is crucial as Autonomous FedGM outperforms Autonomous FedAvg with system heterogeneity. Autonomous FedGM does experience a slowdown compared to the ideal FedGM, but the difference is within a manageable margin, which validates the effectiveness of our proposed Autonomous FedGM.

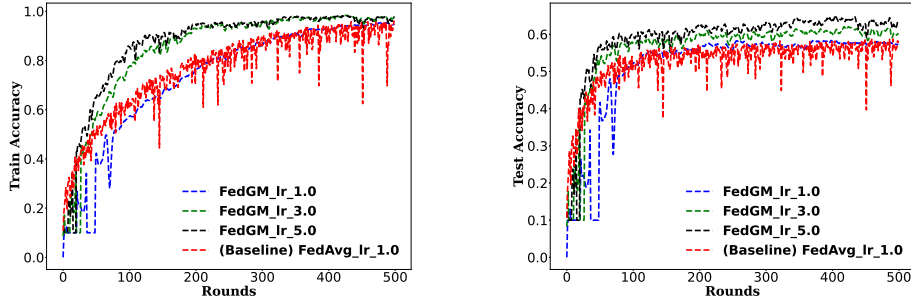


Figure 6: 6(a) Training and 6(b) Testing Curve for FedGM with various server learning rates η .

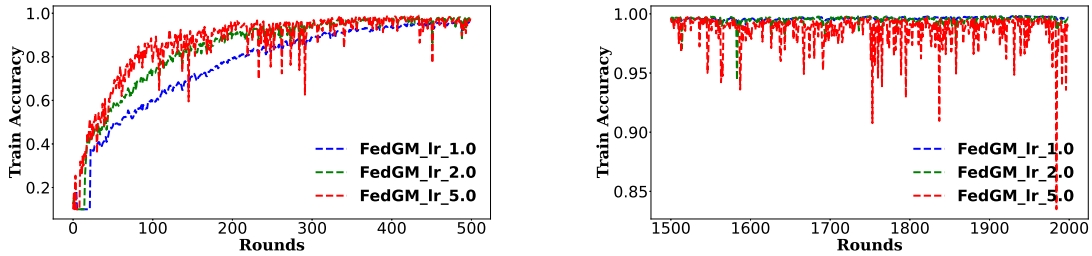


Figure 7: Training Curves for FedGM with various server learning rates η . 7(a) the first 500 rounds; 7(b) the last 500 rounds.

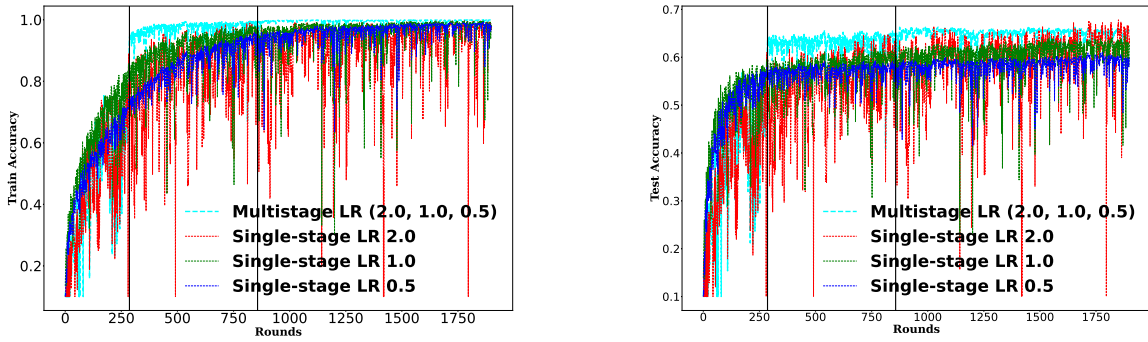


Figure 8: 8(a) Training and 8(b) Testing Curves for Multistage FedGM vs. Single-stage FedGM for 2000 rounds.

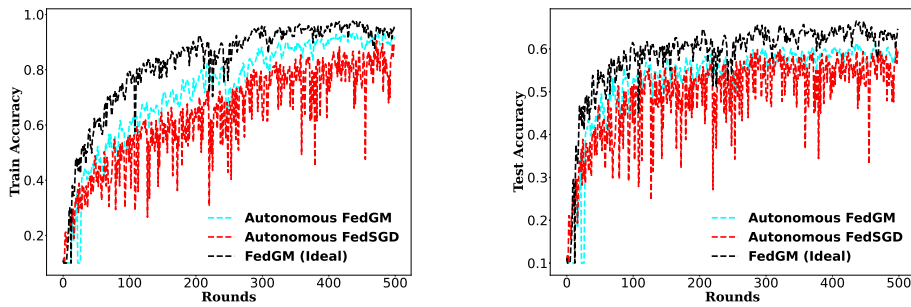


Figure 9: 9(a) Training and 9(b) Testing Curve for ResNet on CIFAR-10 with Random Delay = 10.