Stability-Enhanced Federated Learning with Accelerated Gradient

Tianxiang Chen[1] **Wenjie Hou**[1] **Feng Wang**[1] **Tiantong Wang**[2]
Zhiming Zheng[1] **Shaoting Tang**[1*] **Wei Yang Bryan Lim**[2]

[1]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing,
School of Artificial Intelligence, Beihang University, Beijing, China
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
{txchen, fengwang, houwj}@buaa.edu.cn

Abstract

Federated Learning (FL) is a promising paradigm for distributed machine learning. However, FL often suffers from degraded generalization performance due to the inconsistency between local and global optimization objectives and client-side overfitting. In this paper, we provide a global-update stability analysis framework as an analytical tool to study generalization error and derive the stability bounds of mainstream FL optimization algorithms under non-convex settings. Our analyses reveal how the number of global update steps, data heterogeneity, and update rules influence their stability. We observe that momentum-based FL acceleration methods do not improve stability. To address this issue, we propose FedSEMA, a new FL algorithm that couples global momentum with a gradient-difference corrected Nesterov Accelerated Gradient (NAG) scheme and a hybrid proximal term to enhance stability. This design ensures updates follow a globally consistent descent direction while retaining the benefits of acceleration. Theoretical analysis shows that FedSEMA achieves an improved stability upper bound on non-i.i.d. datasets in the non-convex settings. Extensive experiments on real-world datasets demonstrate that FedSEMA significantly outperforms multiple baseline methods under standard FL settings, achieving faster convergence and state-of-the-art performance. Our code is available at https://github.com/Tenshawn/FedSEMA.

1 Introduction

Federated Learning (FL) enables distributed training without sharing raw data [1], but it struggles with two core challenges in realistic cross-device scenarios: data heterogeneity and limited client participation rate. These factors induce local inconsistency (i.e., divergence between a client's local optimization objective and the global optimization goal) and local overfitting (i.e., client-side models overfit to local data distribution), leading to a phenomenon known as client drift [2, 3]. This hinders both the convergence speed and the generalization performance of the global model [4, 5, 6]. Consequently, the severe client drift caused by local overfitting to inconsistent solutions may lead to the degradation of the global model into merely an average of the client models [2]. This challenge becomes even more pronounced when the participation rate of clients is low during each communication round.

Several methods have attempted to address this challenge by incorporating momentum mechanisms (e.g., FedAdam [7], FedAvgM [8]), variance reduction techniques (e.g., SCAFFOLD [4]) or regularization-based constraints (e.g., FedProx [9], FedDyn [10]). Momentum-based and variance reduction methods mainly focus on reducing optimization error to accelerate convergence, while

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

regularization-based methods aim to improve generalization by constraining the deviation between local and global models during local updates. However, these approaches often fail to significantly enhance generalization in highly non-IID scenarios with low client participation. We attribute this limitation to the absence of explicit attention to the stability of federated optimization, encompassing both its convergence behavior and its sensitivity to data heterogeneity and participation variability.

To study this, we adopt the framework of algorithmic stability [11, 12, 13, 14], which provides a formal link between the sensitivity of an algorithm to data perturbations and its generalization ability. We leverage global-update stability to investigate the sensitivity of FL algorithms to client replacement and establish a formal link to their generalization ability. We analyze the stability of common FL algorithms and show that momentum-based FL acceleration methods do not improve stability. We propose FedSEMA, a new FL algorithm that couples global momentum with a gradient-difference corrected NAG scheme and a hybrid proximal term to enhance stability. This design ensures that the updates always follow a globally consistent descent direction while controlling client drift, improving the stability of global updates and retaining the benefits of acceleration. We prove that FedSEMA achieves a tighter stability bound than existing momentum-based FL algorithms such as FedAdam [7], FedAvgM [8], and FedANAG [15]. We conduct extensive experiments on real-world vision and language datasets, including CIFAR-10/100, EMNIST, Shakespeare, and FEMNIST. Experimental results show that the proposed FedSEMA consistently outperforms existing baselines in terms of test generalization and convergence speed. Our contributions are summarized as follows:

- We provide a global-update stability analysis framework for federated optimization. We establish a formal theoretical link between an algorithm's sensitivity to client replacement and its generalization gap. This perspective reveals that existing FL momentum-base methods often fail to improve (or even degrade) stability, providing a new criterion for evaluating FL algorithms.
- We propose FedSEMA, a federated acceleration algorithm for non-convex FL with partial participation, combining client-side gradient-difference corrected Nesterov momentum and a hybrid proximal term. Compared with previous momentum-based FL methods, FedSEMA achieves improved stability. We provide theoretical analysis of FedSEMA's stability under non-convex settings, showcasing its state-of-the-art performance.
- We empirically validate our theoretical findings by comparing FedSEMA with several established baselines. Extensive experiments confirm that our method consistently outperforms strong baselines in convergence speed, generalization, and stability.

2 Related Work

2.1 Generalization and stability

Generalization and stability in centralized learning. [13] introduced the concept of algorithmic stability and showed that it can be used to derive bounds on the generalization error. On-average stability is proposed by [16] and further studied by [12]. Stability-based generalization analysis was subsequently introduced into stochastic gradient-based methods [11, 17]. [18, 19] extended the results to D-SGD and discussed the relationship between generalization performance and communication topology. Although stability analysis has proven to be an effective tool for characterizing the generalization performance of algorithms in centralized learning, due to issues such as data heterogeneity across clients and client dropout in FL, these conclusions cannot be directly extended to the federated setting.

Generalization and stability in FL. For the generalization bounds in FL, [20] characterized the generalization error for both participating and non-participating clients in the training process. [21] proposed the first algorithm-dependent generalization analysis with uniform stability for the typical personalized FL method. [22] demonstrated that the generalization performance is closely related to the data heterogeneity and the convergence behaviors through on-average stability. [15] focus on the local-update stability of FL, i.e., how client variations affect the model after K local update. To better model the fact that client participation in FL is partial and random, we propose global-update stability, which aims to investigate how variations in the participating client sets affect the difference in models trained by the FL algorithm.

2.2 Federated optimization

[1] first introduced FedAvg to address major challenges such as massively distributed clients and partial client participation. With the rapid development of FL, many optimization methods have been proposed to mitigate the divergence between clients and the global model caused by data heterogeneity. FedProx [9] restricts the offset of local updates by adding a proximal term. FedDyn [10] dynamically modifies the device objective with a penalty term. SCAFFOLD [4] employs variance reduction techniques to correct the client drift in local updates. MimeLite [23] uses a combination of control variates and server-level optimizer states (e.g., momentum) in each client update step. As classic optimization methods, momentum methods can be traced back to Polyak's heavy ball method [24] and Nesterov's accelerated method (NAG) [25]. With the flourishing of momentum mechanisms in machine learning research [26, 27, 28], this motivates researchers to incorporate momentum methods into FL settings. FedAvgM [8] and FedAdam [7] apply momentum to the server side, with the latter utilizing an adaptive optimizer. FedCM [29] introduces global momentum into the local update. FedLNAG [30] employs NAG to accelerate each local iteration. FedACG [31] uses the global model with a lookahead gradient in the penalty term to regularize the local updates. FedANAG [15] incorporates both the global and local analog NAG, while avoiding degrading stability and the communication overhead of uploading additional local momentum. However, these methods only enhance the empirical performance without improving stability relative to FedAvg. Therefore, we aim to design an algorithm with better stability that translates to improved empirical performance.

3 Problem Formulation

3.1 Preliminaries

Throughout this paper, we mainly consider the typical cross-device FL setting, which involves one server and a large number of clients. Let m be the total number of clients, we use $[M]$ to denote the set $\{1, 2, \ldots, m\}$. We denote the set of active clients at round t as $\mathcal{S}_t$, with $|\mathcal{S}_t| = s$. T is the total communication rounds, K represents the number of local updates per communication round. x is the model parameter. $\langle \cdot, \cdot \rangle$ denotes the inner product for two vectors, and $\| \cdot \|$ the Euclidean norm.

We consider minimizing the following optimization problem of the form:

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} F_i(x), \tag{1}$$

where the $F_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{P}_i}[F_i(x, \xi_i)]$ is the local loss function of the client $i \in [M]$, ξ_i represents the random data samples drawn from the distribution $\mathcal{P}_i$ and n_i is the number of local samples. In FL settings, $\mathcal{P}_i$ may differ across the local clients, i.e., for clients i and j, their data distribution might be significantly different. We state some standard assumptions [7, 23, 15] as follows.

Assumption 1. *(L-smooth). F_i is L-smooth for all client i, i.e., $\|\nabla F_i(x) - \nabla F_i(y)\| \leq L\|x - y\|$, $\forall x, y \in R^d$.*

Assumption 2. *(Bounded Variance). The local gradient estimate with randomly sampled data is unbiased, i.e., $\mathbb{E}[g_{t,k}^i] = \nabla F_i(x_{t,k}^i)$. The variance of local gradients is bounded, i.e., $\mathbb{E}[\|g_{t,k}^i - \nabla F_i(x_{t,k}^i)\|^2] \leq \sigma_l^2$. And the variance of local gradients at all clients is bounded, i.e., $\mathbb{E}[\|\nabla F_i(x_t) - \nabla f(x_t)\|^2] \leq \sigma_g^2$, $\forall i \in [M]$ and $\forall t$.*

Assumption 3. *(Z-Lipschitz) F_i is Z-Lipschitz for all client i, i.e., $\|F_i(x) - F_i(y)\| \leq Z\|x - y\|$, $\forall x, y \in R^d$.*

Assumption 1 guarantees a Lipschitz continuity and Assumption 2 guarantees the stochastic gradient is bounded by zero mean and constant variance. Although in practical FL the samples are not i.i.d., they are still sampled from distributions that are not entirely unrelated [9]. Thus, it is reasonable to bound the dissimilarity between local functions in Assumption 2. Assumption 3 is widely used in the stability analysis [11, 32].

3.2 Stability Bounds Analysis

We introduce the notion of algorithmic stability to provide an upper bound on the generalization error. In particular, we formally improve the on-average stability analysis framework proposed

by [22]. While [22] characterizes sensitivity to single data point perturbations within a client, this perspective does not fully capture the dominant source of variability in FL, i.e., partial client participation. In practice, client participation is partial and random, and models must generalize to unseen clients whose data distributions may differ significantly. To better model this scenario, we extend the stability analysis from sample-level perturbations to client-level perturbations, leading to the following definition.

Definition 1. *(Global-update stability for FL). Let $\mathcal{A}(\mathcal{S})$ denote the model output by algorithm $\mathcal{A}$ on client set $\mathcal{S}$. A FL algorithm $\mathcal{A}$ is said to satisfy ϵ-global update stability if given any two client sets $\mathcal{S}$ and $\mathcal{S}^{(i)}$, then for any j*

$$\max_{i \in \mathcal{S}} \mathbb{E}_{\mathcal{A},\mathcal{S}} \left[|F_j(\mathcal{A}(\mathcal{S})) - F_j(\mathcal{A}(\mathcal{S}^{(i)}))| \right] \leq \epsilon,$$

where two neighboring sets $\mathcal{S}$ and $\mathcal{S}^{(i)}$ only differ by one client i. Here, the expectation is also taken over the randomness of the active client set $\mathcal{S}$.

Global-update stability means that any change in the participating FL clients will not cause a significant difference in the model trained by the FL algorithm in expectation. To better capture this partial participation setting, we adopt the concept of semi-empirical risk from [33] and [15] to define generalization error ϵ_{gen}. The unparticipating expected risk as follows

$$R_u = \mathbb{E}_{i \sim \mathcal{C}}[\mathbb{E}_{\xi_i \sim \mathcal{P}_i}[F_i(\mathcal{A}(\mathcal{S}), \xi_i)]], \tag{2}$$

where $\mathcal{C}$ is a meta distribution from which the active clients are drawn. R_u denotes the expected loss of the model over all clients in the meta-distribution. And the semi-empirical risk is defined by

$$R_e = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} [\mathbb{E}_{\xi_i \sim \mathcal{P}_i}[F_i(\mathcal{A}(\mathcal{S}), \xi_i)]]. \tag{3}$$

R_e represents the expected loss of the model over the clients participating in the training. Note that, to characterize the partial and random participation of clients, we adopt semi-empirical risk [33, 15] rather than the traditional empirical risk. This replaces the empirical loss within each client with the expected loss over that client's local data distribution. Traditional generalization error measures the performance of a model on unseen data, while the generalization error we define here measures the performance of the model on the data of unseen clients. To better assess the generalization error for unparticipating clients, we define the generalization error of a model output $\mathcal{A}(\mathcal{S})$ is

$$\epsilon_{gen} := \mathbb{E}_{i \sim \mathcal{C}}[\mathbb{E}_{\xi_i \sim \mathcal{P}_i}[F_i(\mathcal{A}(\mathcal{S}), \xi_i)]] - \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} [\mathbb{E}_{\xi_i \sim \mathcal{P}_i}[F_i(\mathcal{A}(\mathcal{S}), \xi_i)]], \tag{4}$$

a smaller ϵ_{gen} implies the model $\mathcal{A}(\mathcal{S})$ has a better generalization performance on the data of unparticipating clients. The following theorem shows that generalization error of the model can be bounded by global-update stability. The proof is given in Appendix B.

Theorem 1. *Suppose that a FL algorithm $\mathcal{A}$ satisfies the ϵ-global-update stability. Then,*

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{A}}[\epsilon_{gen}] \leq \mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathcal{A}}[|R_u - R_e|] \leq \epsilon.$$

Theorem 1 establishes that if the algorithm exhibits high global update stability (i.e., a small stability upper bound), the generalization gap is guaranteed to be small. This indicates that the model not only fits the participating clients but also generalizes well to unseen clients. Therefore, improving stability becomes a key objective in designing effective FL algorithms. Based on Definition 1, we analyze the stability of existing methods. Following [11], we denote x_T and x'_T as the global models trained by two neighboring active client sets, and, under the Assumption 3, we relate the model difference $\mathbb{E}\|x_T - x'_T\|$ to stability. Supposing the same model initialization, we present in Table 1 the upper bounds of the resulting models for different FL algorithms. See Appendix D for the detailed proof.

As proposed in [22], the model differences of the algorithms increase linearly with the heterogeneity of the client data distributions, are proportional to the sampling variance of the gradients, and also depend on the learning rates chosen during the local training process. Although [15] also analyzed the stability of FL algorithms, which they termed local-update stability, their study only considered the effect of client changes on the model difference after K local training steps, essentially corresponding to stability at the local update level. In contrast, our work investigates global-update stability, namely,

Table 1: Summary of stability bounds of FL algorithms. All results are the ones under non-convex and Assumptions 1-3. We define $\Theta := (\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}$ for simplicity.

Method	Non-convex	Historical gradient[a]	Proximal term	Communication[b]
FedAvg[1]	$2\eta_l K\Theta T$	-	-	$1\times$
FedProx[9] [c]	$2\lambda^{-1}\Theta[1-(1-\lambda\eta_l)^K]T$	-	✓	$1\times$
SCAFFOLD[4]	$6\eta_l K\Theta T$	Variance reduction	-	$2\times$
FedDyn[10]	$4\lambda^{-1}\Theta[1-(1-\lambda\eta_l)^K]T$	-	✓	$1\times$
SCAFFOLD-M[35]	$(2+4\beta)\eta_l K\Theta T$	Client side-momentum	-	$2\times$
FedLNAG[30]	$2\eta_l K\Theta T$	Client side-momentum	-	$2\times$
FedAdam[7]	$2\tau^{-1}\eta_l K\Theta T$	Server side-momentum	-	$1\times$
FedCM[29]	$2\Theta T$	Client side-momentum	-	$1\sim 1.5\times$
MimeLite[23]	$2\eta_l K\Theta T$	-	-	$2\times$
FedANAG[15]	$2(1-\beta)^{-1}\eta_l K\Theta T$	Client side-momentum	-	$1\sim 1.5\times$
FedACG[31]	$(4-2\lambda)(1-\lambda)^{-1}\eta_l K\Theta T$	Server side-momentum	✓	$1\times$
FedSEMA(ours)	$(1-(1-\Gamma)^T)\Gamma^{-1}\zeta\Theta(2+2\beta+4\beta_0)$	Client side-momentum	✓	$1\sim 1.5\times$

[a] This column refers to the utilization of historical gradients $g_{r,\tau}^i$ for $\tau < k$ and $r \leq t$ during the k-th local update.

[b] Here, communication refers to amount of data transmitted w.r.t. to FedAvg.

[c] λ, μ, β are the hyper-parameters of different algorithms.

the impact of changes in client participation on the model difference of FL algorithms over T communication rounds. Moreover, the conclusion on local momentum stability in [15] is limited to the setting of quadratic functions. As shown in Table 1, our stability results demonstrate that, in the non-convex setting, the global stability of existing momentum-based algorithms grows linearly with T. Furthermore, applying Nesterov momentum locally does not significantly increase the upper bound of model discrepancy, and relying solely on global momentum is also insufficient to improve the stability of FL algorithms.

Analysis reveals that FedProx and FedDyn exhibit better stability due to the incorporation of proximal terms in the local update process. Based on the discussion in Appendix D.2, after introducing the proximal term, the local-update stability of FedProx and FedDyn is reduced from the linear upper bound $\mathcal{O}(K)$ to a finite upper bound $\mathcal{O}(1-(1-\lambda\eta_l)^K)$. However, this cannot improve the global-update stability of the FL algorithm, which remains at $\mathcal{O}(T)$. To address this limitation, we propose a hybrid proximal term: $\frac{1}{2(\mu+\lambda)}\left\|(\mu+\lambda)\,x_{t,K} - \lambda x_t\right\|^2$. Then we can get the following conclusion.

Theorem 2. *Suppose Assumption 1-3 hold and consider FedSEMA. Let x_T and x'_T be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}|F_i(x_T) - F_i(x'_T)| \leq Z\mathbb{E}\|x_T - x'_T\| \leq \frac{1-(1-\Gamma)^T}{\Gamma}\zeta(2+2\beta+4\beta_0)(\sigma_l^2+2\sigma_g^2+2Z^2)^{\frac{1}{2}}Z,$$

where $\zeta := \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}$ and $\Gamma := 1-(\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}(1-\eta_l(\lambda+\mu))^K) \in (0,1)$. By combining the proximal term with weight decay, this design transforms the accumulated historical gradients over T rounds into an exponential moving average, thereby reducing the global stability of the algorithm from a linear upper bound $\mathcal{O}(T)$ to a finite upper bound $\mathcal{O}((1-(1-\Gamma)^T)\Gamma^{-1}\zeta)$. The detailed proof can be found in Appendix D.12.

Learning Rate and Convergence. The stability analysis of various algorithms demonstrates that, under non-convex objective functions, stability is influenced by the learning rate. As the number of global iterations T increases, the bias of the FL model in non-convex settings accumulates and becomes progressively larger. To guarantee convergence in such scenarios, the learning rate η_l is typically set to be inversely proportional to the square root of the global rounds T, that is, $\eta_l = \mathcal{O}\left(\frac{1}{LK\sqrt{T}}\right)$. This choice also effectively compensates for the stability discrepancies in non-convex settings. Indeed, this convergence requirement has been validated in many existing studies [34, 30, 15]. Even with this compensation introduced by learning rate decay, the stability of existing algorithms remains at $\mathcal{O}(\sqrt{T})$. Our algorithm achieves a stability of $\mathcal{O}(\frac{1-(1-\Gamma)^T}{\Gamma}\zeta)$, where both Γ and ζ involve the η_l. By substituting η_l, our algorithm obtain stability of $\mathcal{O}(1)$. Detailed calculations are provided in Appendix D.13.

4 Methodology

4.1 FedSEMA algorithm

We introduce FedSEMA, as shown in Algorithm 1, to improve stability and generalization while accelerating update, by coupling global momentum with client-side acceleration and a hybrid proximal term. These components enable FedSEMA to reduce the negative effects of heterogeneity while improving generalization performance with fewer communication rounds.

At the beginning of each round t, a subset of clients $\mathcal{S}_t$ is randomly selected to participate in the current training process (line 4). The global server will communicate the parameter x_t and global momentum $v_t(v_0 = 0)$ to the active clients for local training (line 6). Each active client performs three stages: (1) computing the unbiased stochastic gradient $g_{t,k}^i = \nabla F_i(x_{t,k}^i, \xi_i)$ with local data ξ_i (line 8); (2) computing the local momentum $m_{t,k}^i$ through combining v_t and $g_{t,k}^i$ (line 9); (3) executing the gradient descent step with the local momentum $m_{t,k}^i$ and the hybrid proximal term (line 10). After K iterations, local updates are completed, and the clients communicate the offset of the model parameters $x_{t,K}^i - x_{t,0}^i$ to the server for aggregation (line 12). On the global server, the global momentum v_t will be updated as the weighted sum of the current average local offset Δx_t and the historical global momentum (line 15). Then, the server performs a simple update to derive the global model x_{t+1} (line 16).

Momentum update. Momentum is typically an exponential average of historical gradients, used to correct the direction of the current gradient update, mitigate oscillations, and thereby converge to the global optimum. However, in FL settings, using local momentum can lead the update direction toward local optima rather than the global optimum, thus hindering global convergence. This issue becomes particularly pronounced in the presence of data heterogeneity. Therefore, we use global momentum locally to stabilize and accelerate convergence. Nevertheless, completely ignoring local historical gradient information can also limit the convergence speed of FL. To leverage local historical gradient information without compromising algorithmic stability, we combine the gradient difference $g_{t,k}^i - g_{t,k-1}^i$ with global momentum to make $m_{t,k}^i$ equivalent to Nesterov momentum. In fact, we can demonstrate that the following expression is equivalent to NAG.

$$\begin{cases} v_k = \beta v_{k-1} + \nabla f(x_k) + \beta_0(\nabla f(x_k) - \nabla f(x_{k-1})) \\ x_{k+1} = x_k - \eta_l v_k. \end{cases} \tag{5}$$

The proof details can be referred to in Appendix C. The gradient difference and the exponential average of historical gradients have different interpretations. As stated in [36] and [37], the former proposed a first-order procedure (i.e., difference of gradients) to achieve the negative curvature of a Hessian matrix. Accordingly, our algorithm follows a similar procedure that contains second-order information. Furthermore, we use the gradient difference to correct the $g_{t,k}^i$.

Hybrid proximal term. In traditional optimization problems, penalized proximal terms are typically added to the objective function to improve the model's generalization ability and prevent overfitting [38, 39]. In FedProx [9], the prox term is used to enhance local consistency in FL, while FedDyn [10] utilizes a dynamic regularization term to align the global and local optima. Our analysis shows that the proximal term can improve the local-update stability $\mathcal{O}(1-(1-\eta_l\lambda)^K)$, but the global-update stability remains $\mathcal{O}(T)$. Therefore, we propose a hybrid proximal term: $\frac{1}{2(\mu+\lambda)} \left\|(\mu+\lambda)\, x_{t,K} - \lambda x_t\right\|^2$, which further improves the stability of global updates to $\mathcal{O}((1-(1-\Gamma)^T)\Gamma^{-1}\zeta)$. As shown in the local update in Algorithm 1 (Line.10), for all $i, j \in [M], i \neq j$ we have:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq (1-\Gamma)\mathbb{E}\|x_t - x_t'\| + \zeta(2+2\beta)\Theta. \tag{6}$$

where $\zeta = \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}$. Proof details can be referred to in Appendix D.12. From (6), it can be observed that the hybrid proximal term asymmetrically reweights the local and global models, thereby introducing an additional contraction effect during the global update and attenuating the discrepancy between global updates across neighboring client sets, which differs fundamentally from FedProx, as shown in Fig 6b. This also explains why our method achieves improved stability.

Communication costs. The communication cost of the FedSEMA consists of two parts: the downlink cost of broadcasting the global model x_t and global momentum v_t from the server, and the uplink cost of unicasting s local model updates from the clients. In wireless communication, broadcasting a

Algorithm 1 FedSEMA Algorithm Framework

Input: model parameters x_0, total communication rounds T, number of local iterations K, momentum coefficient β,proximal term λ, μ, local learning rate η_l, global learning rate η.
Output:model parameters x_T.
for $t = 0, 1, 2, \ldots, T-1$ **do**
  select active clients-sets $\mathcal{S}_t$ at round t
  for client $i \in \mathcal{S}_t$ in parallel **do**
    communicate x_t and v_t to local client i and initialize $x^i_{t,0} = x_t$
    for $k = 0, 1, 2, \ldots, K-1$ **do**
      compute unbiased stochastic gradient: $g^i_{t,k} = \nabla F_i(x^i_{t,k}, \xi_i)$
      gradient correction: $\hat{g}^i_{t,k} = g^i_{t,k} + \beta_0(g^i_{t,k} - g^i_{t,k-1})$
      update the momentum: $m^i_{t,k} = \beta v_t + \hat{g}^i_{t,k}$
      update the gradient step: $x^i_{t,k+1} = x^i_{t,k} - \eta_l(m^i_{t,k} + ((\lambda + \mu)x^i_{t,k} - \lambda x_t))$
    end for
    $\Delta x^i_t = x^i_{t,K} - x^i_{t,0}$ and communicate Δx^i_t to server
  end for
  $\Delta x_t = \frac{1}{s}\sum_{i \in \mathcal{S}_t} \Delta x^i_t$
  update global momentum: $v_{t+1} = \frac{\beta}{1+\beta} v_t - \frac{\Delta x_t}{(1+\beta)K\eta_l}$
  update global model: $x_{t+1} = x_t + \eta \Delta x_t$
end for

file to s destinations is significantly more efficient than unicasting s copies of the same file to each destination. Therefore, the communication overhead of FedSEMA is approximately $\frac{2+s}{1+s}$ times that of FedAvg, where s typically represents a large number in cross-device FL [40].

4.2 Convergence analysis

We rigorously prove the convergence of FedSEMA for non-convex settings, assuming partial participation, i.e.$|\mathcal{S}_t| < m, \forall t \in [T]$. Our analysis can be directly extended to full participation. For the full proof, please refer to Appendix E.

Theorem 3. *Under Assumption 1-2, suppose the loss function is non-convex and* $0 < \beta_0 < 0.1$. *We define* $D_0 := \mathbb{E}(\hat{f}(z_0) - \hat{f}(z_T))$. *When the learning rates satisfy* $\eta_l \leq \frac{1}{16KL}$ *and*

$$\frac{\eta\eta_l L}{\beta'} \leq \frac{m(s-1)}{s(m-1)} \min\left\{\frac{\beta'\sqrt{K}}{4\alpha\beta\sqrt{3\alpha}}, \frac{\eta}{6\beta\sqrt{2\alpha}}, \frac{1}{2\alpha}\right\},$$

where $\beta' = 1 - \alpha'\beta$, *then the auxiliary sequence* z_t *generated by executing FedSEMA satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_t\|\nabla \hat{f}(x_t)\|^2 \leq \mathcal{O}\bigg(\frac{LD_0}{\sqrt{sKT}} + \frac{\sigma_l^2}{\sqrt{sKT}} + \frac{1}{T}\left(\frac{\sigma_l^2}{K^2} + K\sigma_l^2 + \frac{\sigma_l^2}{sK} + \frac{\sigma_g^2}{K}\right) + \Psi_l + \Psi_g\bigg),$$

where $\Psi_l = \big(\frac{1}{T} + \frac{K}{sT}\big)\frac{(m-s)}{m-1}\big(\frac{1}{T} + \frac{1}{sK^2T} + \frac{1}{\sqrt{sKT}}\big)\sigma_l^2, \Psi_g = \big(\frac{K}{T} + K\big)\frac{(m-s)}{m-1}\big(\frac{1}{T} + \frac{1}{sK^2T} + \frac{1}{\sqrt{sKT}}\big)\sigma_g^2$.

Remark 1. *To obtain the above results, it is also necessary to suppose that* $\eta = \mathcal{O}(\sqrt{sK})$ *and* $\eta_l = \mathcal{O}(\frac{1}{LK\sqrt{T}})$. *This is consistent with the assumptions on the convergence rate in [7, 34]. However, we do not make the assumption of bounded gradients as in [7]. When* T *is sufficiently large, the dominant term of convergence bound of algorithm is* $\mathcal{O}(\frac{1}{\sqrt{mKT}})$ *for full participation and* $\mathcal{O}(\frac{\sqrt{K}}{\sqrt{sT}})$ *for partial participation.*

Remark 2. *The terms* Ψ_l *and* Ψ_g *capture the additional contributions of stochastic gradient noise and client heterogeneity under partial participation, respectively. Both terms vanish under full participation* ($s = m$).

Table 2: Number of communication rounds required to achieve the preset target accuracy (presented in the first two lines on four datasets), and the top validation accuracy averaged over five random seeds for recent state-of-the-art methods (presented in the third line on four datasets), with the bold number representing the best result. Using CIFAR-10 as an example, the first two lines present the required communication rounds to reach 81% and 75% test accuracy, respectively. The third line shows the highest accuracy achieved by each method within 2000 rounds. '2000+' indicates that 81% accuracy was not reached even after 2000 rounds.

Method	CIFAR-10			CIFAR-100			EMNIST			Shakespeare		
	75	81	top(%)	40	45	top(%)	78	80	top(%)	45	50	top(%)
FedAvg	572	2000+	78.64	2000+	2000+	40.79	102	271	81.01	176	516	51.61
FedProx	572	2000+	78.88	1207	2000+	41.74	46	154	81.83	194	522	51.54
FedDyn	354	2000+	80.14	1512	2000+	41.89	65	146	81.02	231	489	51.77
SCAFFOLD	364	1535	81.34	1308	2000+	42.24	60	141	81.55	180	465	52.04
MimeLite	439	1959	81.02	578	1050	47.59	57	129	81.91	177	477	51.71
SCAFFOLD-M	436	1495	81.94	763	1225	48.69	56	148	81.96	97	342	52.26
FedAvgM	489	1522	81.86	714	1197	48.31	50	151	81.92	84	313	52.32
FedAdam	1295	2000+	78.29	1458	1877	44.51	61	162	81.58	99	359	52.30
FedCM	530	1891	81.09	1265	1718	45.02	277	1101	80.16	2000+	2000+	39.14
FedLNAG	497	1000+	79.60	1031	1602	42.31	41	196	81.92	93	296	52.23
FedANAG	338	882	83.51	673	1701	45.32	38	181	82.01	68	183	52.46
FedACG	338	891	83.52	582	1139	50.59	36	155	82.02	72	214	52.52
FedSEMA(ours)	**283**	**846**	**84.09**	**532**	**847**	**54.58**	**30**	**131**	**82.25**	**53**	**144**	**53.08**

5 Experiments

5.1 Setup

Datasets and models. We adopt standard FL benchmarks following the same train/test splits as prior works [1, 9], including CIFAR-10 and CIFAR-100 [41], a subset of EMNIST [42] referred to as EMNIST-L, and the Shakespeare dataset [43]. To comprehensively evaluate performance across diverse federated scenarios, the experiments adopt three settings similar to those in previous work [15]: Setting I (CIFAR-10 and CIFAR-100): 200 clients with a 2% participation rate per round; Setting II (EMNIST-L): 500 clients with a 1% participation rate; Setting III (Shakespeare): 100 clients with a 3% participation rate. To simulate data heterogeneity, we follow the approach of [8], sampling label distributions from a Dirichlet distribution with concentration parameter 0.3. For the IID setting, training data is randomly assigned to clients. In the non-IID setting, label distributions follow the sampled Dirichlet proportions. The model architectures include multinomial logistic regression, convolutional neural networks (CNNs), and recurrent neural networks (RNNs), standard ResNet-18 network, aligned with configurations used in previous studies [2, 10, 31, 15]. Due to space limitations, further details on datasets and models are provided in Appendix A.

Baseline methods. We compare against a broad set of classical and efficient methods designed to address local inconsistency and client drift in FL, including FedAvg [1], FedProx [9], FedAdam [7], SCAFFOLD [4], FedCM [29], SCAFFOLD-M [35] FedDyn [10], FedACG [31], FedANAG [15], FedAvgM [8], MimeLite [23], and FedLNAG [30]. FedAdam improves global updates via adaptive optimization. FedCM, FedACG, FedANAG, and FedAVG-M leverage global momentum to correct local updates. SCAFFOLD and FedDyn mitigate client heterogeneity by aligning local and global loss functions using control variates or dual variables, respectively. SCAFFOLD-M extends SCAFFOLD by incorporating momentum. FedProx addresses local inconsistency through the addition of a proximal term. To ensure a fair evaluation under practical federated settings, we report the test accuracy after 2000 communication rounds across all experiments, following the baseline settings in [10, 15]. Further implementation details and hyperparameter configurations are provided in Appendix A.

Evaluation metrics. Different algorithms are evaluated with two metrics: the convergence rate (i.e., number of rounds required to reach a target test accuracy) and the generalization performance (i.e., final global model performance on a validation set throughout training). These two metrics are widely adopted by many prior works such as [10, 31, 15].

Table 3: Accuracy (%) vs. Non-iid-ness, With PR = 1%.

Dirichlet	0.3	0.6	0.8	iid
FedSEMA	**82.88**	**84.11**	**84.51**	**85.40**
FedACG	81.52	82.34	83.12	85.03
FedANAG	82.37	83.11	84.27	85.13
SCAFFOLD	80.88	81.88	82.31	83.17
FedDyn	78.45	81.95	83.01	83.10
FedLNAG	79.05	80.53	80.93	81.99
FedAvgM	81.34	82.15	83.12	84.70
FedAdam	78.81	81.10	82.52	84.50
FedCM	79.44	80.46	81.62	82.07

Table 4: Accuracy (%) vs. PR, With Dirichlet = 0.3.

PR(%)	1	2	5	10
FedSEMA	**82.88**	**84.09**	**83.53**	**83.04**
FedACG	81.52	83.52	83.14	82.63
FedANAG	82.37	83.51	83.04	82.50
SCAFFOLD	80.88	81.34	81.97	82.12
FedDyn	78.45	80.14	80.45	82.46
FedLNAG	79.05	79.60	80.10	79.54
FedAvgM	81.34	81.86	82.71	82.08
FedAdam	78.81	78.29	81.77	81.73
FedCM	79.44	81.09	82.00	81.89

5.2 Main results

We evaluate the aforementioned FL algorithms under a variety of challenging settings, including scenarios characterized by extreme data heterogeneity and very low client participation rates. The peak test accuracies reported in Table 2 are obtained by averaging results over five independent runs with different random seeds to ensure statistical reliability. Table 2 demonstrates the test accuracies of FedSEMA and baseline methods on CIFAR-10, CIFAR-100, EMNIST, and Shakespeare under different settings, and the corresponding convergence curves are shown in Appendix A.11. We further investigate the sensitivity of performance to the hyperparameters β, β_0, μ and λ under varying configurations in Appendix A.6, and present an ablation study in Appendix A.12. Appendix A.10 presents the variance of test accuracy across different seeds to complement the robustness analysis. In addition, we verified the stability and empirical generalization error of most algorithms as shown in Appendix A.4. We also conducted experiments under feature-skewed settings in Appendix A.9.

As shown in Table 2, FedSEMA consistently and significantly improves both convergence speed and accuracy across almost all scenarios. This remarkable performance is attributed to the simultaneous utilization of global momentum and gradient differences at the client side to approximate NAG and the hybrid proximal term. This momentum-based acceleration mechanism not only leverages gradient differences and global momentum to accelerate both local and global training, but also provides a globally consistent update direction, thereby enhancing both local and global convergence. Furthermore, aligning local updates with the trajectory of global gradients improves inter-client model consistency, resulting in increased stability. In contrast, other momentum-based methods such as FedLNAG can accelerate local convergence toward local optima, but due to the misalignment between local and global optima, the performance of the aggregated global model remains suboptimal. On the other hand, approaches that rely solely on global momentum such as FedCM, FedACG, and FedANAG may improve global convergence, but have limited ability to improve inter-client consistency, which in turn leads to degraded generalization performance. This observation is also theoretically supported by our stability analysis. In addition, compared with methods that rely solely on global momentum and the current gradient, FedSEMA leverages the hybrid proximal term to correct the gradient, thereby further improving the consistency of client update directions. This process reduces model discrepancies across clients and further enhances training efficiency.

Tables 3 and 4 demonstrate that FedSEMA exhibits the highest robustness under conditions of severe data heterogeneity and low client participation rates. To further evaluate performance, we examine the peak validation accuracy achieved by each method on the CIFAR-10 dataset under varying degrees of non-i.i.d. distributions and client participation rates (PR). Across both i.i.d. and non-i.i.d. scenarios, FedSEMA consistently outperforms all baselines under various participation settings in terms of generalization. Moreover, increases in data heterogeneity and decreases in client participation exert only a minimal adverse impact on the performance of FedSEMA. Results under more heterogeneous settings are provided in Appendix A.8.

6 Conclusion

We propose a global-update stability analysis framework and establish a formal theoretical link between an algorithm's sensitivity to client replacement and its generalization gap. Our stability

analysis shows that existing acceleration schemes do not yield notable advantages in terms of stability. Therefore, we propose FedSEMA, a novel and practical FL algorithm designed to address the issue of unstable client updates in existing momentum-based and variance-reduction-based methods. FedSEMA employs a gradient-difference corrected NAG mechanism to guide client updates and introduces a hybrid proximal term to further enhance the consistency among client models. We provide theoretical analysis demonstrating that FedSEMA achieves improved stability in non-convex settings. Extensive experiments conducted on multiple datasets validate the superiority of FedSEMA in both training efficiency and final performance.

Acknowledgments and Disclosure of Funding

This work is supported by Program of National Natural Science Foundation of China (12425114, 62441617), Taishan Scholars Program of Shandong Province (tstp20230635), Key R&D Program of Shandong Province (2026CXPT283), the Fundamental Research Funds for the Central Universities, the Opening Project of the State Key Laboratory of General Artificial Intelligence(Project NSKLAGI2025OP16), the special Funding for the "Case-by-Case-Introduction of Top Talent (Teams)", and Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

References

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. In *International Conference on Learning Representations*, 2023.

[3] Yan Sun, Li Shen, and Dacheng Tao. Understanding how consistency works in federated learning via stage-wise relaxed initialization. *Advances in Neural Information Processing Systems*, 36:80543–80574, 2023.

[4] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

[5] Zachary Charles and Jakub Konečnỳ. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.

[6] Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pages 6692–6701. PMLR, 2020.

[7] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

[8] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

[9] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

[10] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

[11] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[12] Ilja Kuzborskij and Christoph Lampert. Data-dependent stability of stochastic gradient descent. In *international conference on machine learning*, pages 2815–2824. PMLR, 2018.

[13] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[14] Yikai Zhang, Wenjia Zhang, Sammy Bald, Vamsi Pingali, Chao Chen, and Mayank Goswami. Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in artificial intelligence*, pages 2364–2373. PMLR, 2022.

[15] Hao Zhang, Chenglin Li, Wenrui Dai, Ziyang Zheng, Junni Zou, and Hongkai Xiong. Stabilizing and accelerating federated learning on heterogeneous data with partial client participation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):67–83, 2025.

[16] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[17] Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34):1–44, 2020.

[18] Tao Sun, Dongsheng Li, and Bao Wang. Stability and generalization of decentralized stochastic gradient descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9756–9764, 2021.

[19] Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized sgd. In *International Conference on Machine Learning*, pages 27479–27503. PMLR, 2022.

[20] Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *The Eleventh International Conference on Learning Representations*, 2023.

[21] Yingqi Liu, Qinglun Li, Jie Tan, Yifan Shi, Li Shen, and Xiaochun Cao. Understanding the stability-based generalization of personalized federated learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[22] Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 676–684. PMLR, 2024.

[23] Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Breaking the centralized barrier for cross-device federated learning. *Advances in Neural Information Processing Systems*, 34:28663–28676, 2021.

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

[25] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[26] Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. In *International Conference on Learning Representations*, 2020.

[27] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020.

[28] Mahmoud Assran and Michael Rabbat. On the convergence of nesterov's accelerated gradient method in stochastic settings. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[29] Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. Fedcm: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

[30] Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[31] Geeho Kim, Jinkyu Kim, and Bohyung Han. Communication-efficient federated learning with accelerated client gradient. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12385–12394, June 2024.

[32] Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pages 5809–5819. PMLR, 2020.

[33] Honglin Yuan, Warren Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? 2022.

[34] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

[35] Ziheng Cheng, Xinmeng Huang, Pengfei Wu, and Kun Yuan. Momentum benefits non-iid federated learning simply and provably. In *International Conference on Learning Representations*, 2024.

[36] Yi Xu, Rong Jin, and Tianbao Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[37] Yuanyuan Liu, Fanhua Shang, Weixin An, Junhao Liu, Hongying Liu, and Zhouchen Lin. A single-loop accelerated extra-gradient difference algorithm with improved complexity bounds for constrained minimax optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[38] Anders Krogh and John Hertz. A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4, 1991.

[39] Benyamin Ghojogh and Mark Crowley. The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*, 2019.

[40] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[41] Alex Krizhevsky. Learning multiple layers of features from tiny images. Jan 2009.

[42] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[43] William Shakespeare. *The complete works of William Shakespeare*, volume 10. Harper, 1907.

[44] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International conference on machine learning*, pages 7252–7261. PMLR, 2019.

[45] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[47] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020.

[48] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

A Experiment details

A.1 Real data

We evaluate our models on four benchmark datasets: EMNIST-L, CIFAR-10, CIFAR-100, and Shakespeare. EMNIST-L is a subset of the EMNIST dataset, comprising the first 10 characters from the alphabet split, following prior work such as FedProx [9] and FedDyn [10]. For image classification tasks, EMNIST-L, CIFAR-10, and CIFAR-100 are used, with image dimensions of $1 \times 28 \times 28$ for EMNIST-L and $3 \times 32 \times 32$ for both CIFAR-10 and CIFAR-100. These datasets contain 10, 10, and 100 classes, respectively. The Shakespeare dataset is used for next-character prediction, with 80 distinct characters as inputs and the subsequent character as the label, resulting in 80 output classes. We adopt the standard training/testing splits for all datasets. A summary of the number of training and test samples for each benchmark is provided in Table 7.

To generate Non-IID data partitions for EMNIST, CIFAR-10, and CIFAR-100, we allocate training samples to clients based on class labels. Specifically, following prior works [44], we use the Dirichlet distribution to create federated heterogeneous datasets by sampling a class-probability vector for each client, where each vector is drawn from a Dirichlet distribution with a concentration parameter that controls the degree of data heterogeneity. For each client, labels are sampled according to this probability vector, and corresponding images are drawn without replacement. This process is repeated until all data points are allocated. As a result, each client's label distribution follows the Dirichlet distribution, with the concentration parameter governing the level of statistical heterogeneity across devices. For instance, when the Dirichlet factor is set to 0.3, approximately 80% of each client's data is concentrated in 3–4 dominant classes. In the IID setting, data is randomly shuffled and evenly distributed across clients.

For the Shakespeare dataset, we adopt the LEAF framework [45] to generate Non-IID data, capping each client's dataset to 2000 samples, consistent with FedDyn. In this setting, data is naturally partitioned by role, where each device corresponds to a character in the script and holds all lines spoken by that character. In the IID setting, all lines are merged and then randomly distributed across clients.

A.2 Models

For CIFAR-10 dataset, we adopt a CNN architecture similar to those used in FedAvg and FedDyn, which consists of two convolutional layers with 64 filters of size 5×5, followed by two 2×2 max-pooling layers, two fully connected layers with 384 and 192 neurons respectively, and a final softmax output layer (see Table 5 for complete model details). This CNN model is identical to the one used in FedAvg, except that batch normalization layers are excluded. For the CIFAR-100 dataset, we use standard ResNet-18-GN model [46] backbone with the 7×7 filter size in the first convolution layer as implemented in the previous works, e.g. for [4, 10]. We follow the [47, 2] to replace the batch normalization layer with group normalization layer [48], which can be aggregated directly by averaging. These are all common setups in many previous works. For the EMNIST-L dataset, we use a multi-class logistic regression model with a cross-entropy loss function. For the next-character prediction task on the Shakespeare dataset, we follow the configurations used in FedProx and FedDyn, employing an RNN model that first embeds the input character sequence into an 8-dimensional space, then passes it through a two-layer LSTM with 100 units, and finally outputs predictions via a softmax layer (see Table 6 for full specifications).

Table 5: CNN Architecture for CIFAR-10

Layer Type	Size
Convolution + ReLU	$5 \times 5 \times 64$
Max Pooling	2×2
Convolution + ReLU	$5 \times 5 \times 64$
Max Pooling	2×2
Fully Connected + ReLU	1600×384
Fully Connected + ReLU	384×192
Fully Connected	192×10 & 192×100

Table 6: Shakespeare model architecture

Layer Type	Size
Embedding	(80, 8)
LSTM	(80, 100)
LSTM	(80, 100)
Fully Connected	(100, 80)

Table 7: Train and test splits

Dataset	No. Train	No. Test	No. clients	sampling rate	No. Train per client	Batch size	Rounds
EMNIST-L	48000	8000	500	1%	96	48	2000
CIFAR-10	50000	10000	200	2%	250	50	2000
CIFAR-100	50000	10000	200	2%	250	50	2000
Shakespeare	200000	40000	100	3%	2000	100	2000

A.3 Hyper-parameters

In our experiments, all algorithms are implemented using PyTorch 2.0.0 with CUDA 11.8 on a GEFORCE RTX 4090 GPU. We consider different hyperparameter configurations for various setups and datasets. For all experiments, we fix the batch size as 48 for EMNIST-L, 50 for CIFAR-10 and CIFAR-100, and 100 for Shakespeare. For each dataset, we compare the performance of different methods under various hyperparameter configurations. The client learning rate η_l and the learning rate decay factor are individually tuned via grid search.

EMNIST-L. As for the 500 devices, balanced data, full participation setup, hyperparameters are searched for all algorithms in all IID and Dirichlet settings for a fixed 100 communication rounds. The search space consists of learning rates in $[0.1, 0.01, 0.001]$, λs in $[1, 0.01, 0.001]$ and αs in $[0.001, 0.01, 0.02, 0.1]$, epochs is 20. The learning rate decay is selected from the range of $[0.995, 0.998, 0.999, 1.0]$. The selected configuration for FedAvg is 0.1 learning rate; for FedProx is 0.001 learning rate and 0.001 μ; for FedDyn is 0.001 learning rate and 0.01α; and for SCAFFOLD and SCAFFOLD-M is 0.001 learning rate; and for FedSEMA, FedANAG, FedACG, FedLNAG, FedAvgM are 0.001 learning rate and 0.9β; and for FedCM is 0.001 learning rate and 0.1β; and for MimeLite is 0.001 learning rate and 0.2β; and for FedAdam is 0.001 learning rate and 0.1 global learning rate for all IID and Dirichlet settings. These configurations are fixed and their performances are obtained for 500 communication rounds.

CIFAR-10. We used similar hyperparameters as in EMNIST-L dataset. The configuration includes: 0.1 for learning rate, 5 for epochs. The learning rate decay is selected from the range of $[0.995, 0.998, 0.999, 1.0]$. The α value is selected from the range of $[0.001, 0.01, 0.1]$ for FedDyn. The μ value is selected from the range of $[0.1, 0.01, 0.001, 0.0001]$.

CIFAR-100. The same hyperparameters are applied to the CIFAR-100 experiments with 200 devices including: 0.1 for learning rate, 5 for epochs. The learning rate decay is selected from the range of $[0.995, 0.998, 0.999, 1.0]$. The α value is selected from the range of $[0.001, 0.01, 0.1]$ for FedDyn. The μ value is selected from the range of $[0.1, 0.01, 0.001, 0.0001]$.

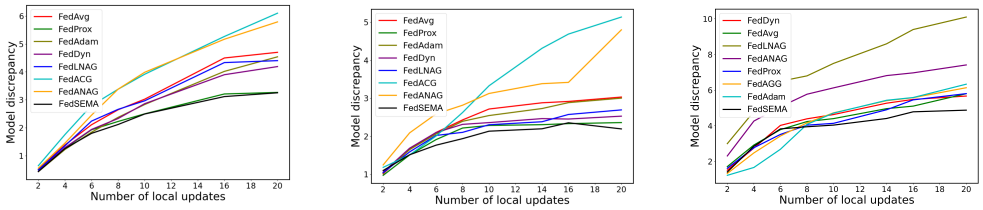
Shakespeare. As for 100 devices, balanced data, full participation setup, the hyperparameters are searched with all combinations of learning rate in 1, epochs in $[1, 5]$, λs in $[0.01, 0.001]$ and αs in $[0.001, 0.009, 0.01, 0.015]$. The learning rate decay is selected from the range of $[0.995, 0.998, 0.999, 1.0]$. The selected configuration for FedSEMA, FedANAG, FedACG, FedLNAG, FedAvgM are 0.9β; and for FedCM is 0.1β; and for MimeLite is 0.2β; The α value is selected from the range of $[0.001, 0.01, 0.1]$ for FedDyn. The μ value is selected from the range of $[0.1, 0.01, 0.001, 0.0001]$.

A.4 Validation of stability and empirical generalization error

As stated in Section 3.2, since we can establish a connection between the difference in loss functions and the difference in models, we measure stability by computing the maximum discrepancy among client models after two times of FL training. We evaluate the global update stability by measuring the model discrepancy for models including multiclass logistic classification and CNN. We use EMNIST on multi-class logistic classification model with 500 clients and 1% client sampling, and CIFAR10/CIFAR100 on CNN with 200 clients and 2% client sampling. Empirical evaluations on logistic regression and CNN models are shown in Fig 1 and Fig 2. The model discrepancy here is measured using the L_2 norm. For the empirical generalization error, we measure it using the absolute difference between the training and test losses, as well as the absolute difference between the training and test accuracies. We evaluate this metric on CIFAR-10 using a CNN model with 200 clients and a 2% client sampling rate. The results are shown in Fig. 3.

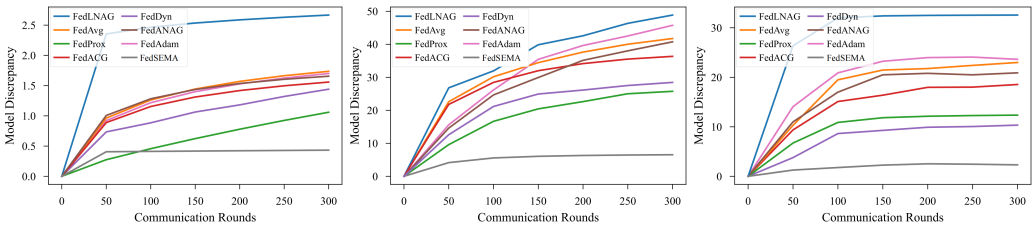
As shown in Fig. 1, the experimental results have verified local-update stability of our method. Existing momentum-based algorithms, due to data heterogeneity, may actually impair local-update stability when local momentum acceleration is used (e.g., FedLNAG); using global momentum (e.g., FedAdam) performs slightly better, but it still does not improve compared to FedAvg. Besides, while FedDyn's stability is enhanced compared to FedAvg, its local-update stability is inferior to that of FedSEMA and FedProx due to low participation rates.

As show in Fig. 2, the experimental results have verified global-update stability of our method. The FedLNAG algorithm exhibits significantly poorer stability than other methods across all tasks. The stability of other momentum-based methods is largely comparable to that of FedAvg. In terms of global update stability, the improvements achieved by FedDyn and FedProx over FedAvg and related methods are rather limited, and their performance remains far inferior to that of FedSEMA. In contrast, FedSEMA demonstrates substantially better global stability than all other methods, further validating the effectiveness of our approach.



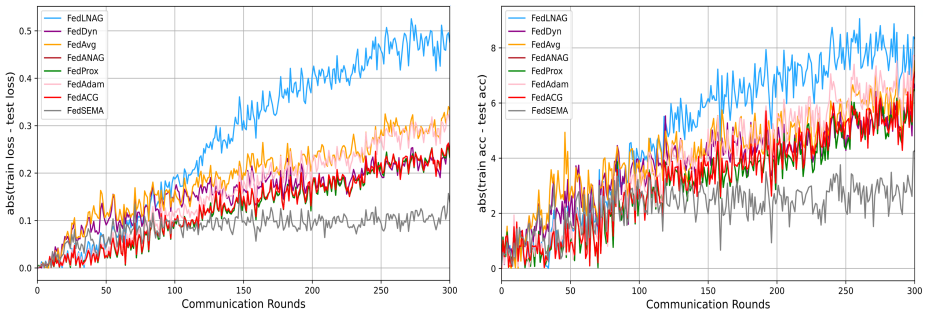
(a) Emnist (b) Cifar10 (c) Cifar100

Figure 1: Local model discrepancy (local-update stability) vs. index of local updates K.



(a) Emnist (b) Cifar10 (c) Cifar100

Figure 2: Global model discrepancy (global-update stability) vs. index of communication rounds T.



(a) loss (b) accuracy

Figure 3: Empirical generalization error vs. communication rounds T.

As show in Fig. 3, the results are highly consistent with the stability results, further confirming the reliability of our theoretical findings.

Table 8: Accuracy (%) vs. Non-iid-ness, With PR = 1%, on CIFAR-100.

Dirichlet	0.3	0.6	0.8	iid
FedSEMA (ours)	**52.91**	**53.69**	**54.89**	**55.81**
FedANAG	44.08	45.59	46.16	47.33
FedACG	50.07	50.08	51.11	52.23
FedAvgM	47.76	48.69	49.71	50.36
FedProx	41.62	42.09	42.95	43.46
SCAFFOLD	41.78	42.69	44.12	45.83
FedDyn	41.23	43.78	45.11	45.57
FedAdam	44.23	46.02	47.47	49.38

Table 9: Accuracy (%) vs. PR, With Dirichlet = 0.3, on CIFAR-100.

PR(%)	1	2	5	10
FedSEMA (ours)	**52.91**	**54.58**	**54.44**	**53.95**
FedANAG	44.08	45.32	45.04	44.57
FedACG	50.07	50.59	50.34	50.18
FedAvgM	47.76	48.31	48.62	48.74
FedProx	41.62	41.74	42.58	42.98
SCAFFOLD	41.78	42.24	42.54	44.03
FedDyn	41.23	41.89	42.16	43.96
FedAdam	44.23	44.51	46.30	46.14

A.5 Validation on a large-scale dataset

We further evaluate FedSEMA on CIFAR-100 under different data heterogeneity levels and client participation rates. Specifically, we vary the Dirichlet parameter in $\{0.3, 0.6, 0.8, \text{iid}\}$ with the participation rate fixed to 1%, and vary the participation rate in $\{1\%, 2\%, 5\%, 10\%\}$ with the Dirichlet parameter fixed to 0.3.

As shown in Tables 8 and 9, FedSEMA consistently achieves the best performance across all settings. These results further demonstrate the robustness of FedSEMA under challenging data heterogeneity and varying client participation rates.

A.6 Hyperparameter sensitivity

Follow previous work [10, 2], We evaluate the performance of our algorithm under different settings of β, β_0, μ and λ on the CIFAR-10 and CIFAR-100 datasets, using 500 communication rounds and 10% client participation among 100 clients. The Dirichlet parameter is set to $\alpha = 0.3$.

Hyperparameter sensitivity on CIFAR-10. Momentum coefficent β for global momentum. In the analysis, β can be selected as a proper value which has no impact on the convergence complexity. Table 10 shows that performance of FedSEMA remains stable across β values ranging from 0.1 to 0.99. Despite a slight decrease when β is too small ($\beta = 0.1$), the accuracy remains consistently high, peaking at $\beta = 0.9$.

Table 10: Performance of different β with $\mu = 0.001$, $\lambda = 0.01$ and $\beta_0 = 0.1$.

β	0.99	0.95	0.9	0.8	0.7	0.6	0.1
Acc.	80.45	81.41	81.58	81.27	81.19	80.88	78.89

Momentum Coefficent β_0 for gradient difference. As shown in Table 11, the performance of FedSEMA remains stable across β_0 values ranging from 0.001 to 0.3. Despite slight fluctuations, the accuracy stays consistently high, peaking at $\beta_0 = 0.1$.

Table 11: Performance of different β_0 with $\beta = 0.9$, $\mu = 0.001$ and $\lambda = 0.01$.

β_0	0.3	0.2	0.1	0.05	0.02	0.01	0.001
Acc.	81.08	81.27	81.58	81.44	81.21	80.96	80.76

Coefficent μ for the weight of the proximal term. As shown in Table 12, the performance of FedSEMA remains stable across μ values ranging from 0.0001 to 0.01. Despite slight fluctuations, the accuracy stays consistently high, peaking at $\mu = 0.001$.

Table 12: Performance of different μ with $\beta = 0.9$, $\beta_0 = 0.1$ and $\lambda = 0.01$.

μ	0.1	0.01	0.005	0.001	0.0005	0.0001	0.00001
Acc.	77.46	80.33	81.16	81.58	81.35	80.33	78.86

Coefficent λ for the weight of the proximal term. As shown in Table 13, the performance of FedSEMA remains stable across λ values ranging from 0.0001 to 0.1. Despite slight fluctuations, the accuracy stays consistently high, peaking at $\lambda = 0.01$.

Table 13: Performance of different λ with $\beta = 0.9$, $\mu = 0.001$ and $\beta_0 = 0.1$.

λ	1	0.1	0.05	0.01	0.005	0.001	0.0001
Acc.	76.52	80.88	81.18	81.58	81.49	81.40	81.29

The performance of FedSEMA remains stable across a wide range of values for the parameters β, β_0, μ, and λ. Specifically, for β values ranging from 0.6 to 0.95, FedSEMA maintains consistently high accuracy with slight fluctuations, peaking at $\beta = 0.9$. For β_0 values ranging from 0.001 to 0.3, the accuracy remains high and stable, with the highest performance observed at $\beta_0 = 0.1$. For μ values ranging from 0.01 to 0.0001, the accuracy remains high and stable, with the highest performance observed at $\mu = 0.001$. Also, for λ values ranging from 0.1 to 0.0001, FedSEMA shows stable performance with consistent accuracy, reaching its peak at $\lambda = 0.01$. These results demonstrate that FedSEMA is robust to variations in these hyperparameters, even under the condition that of high heterogeneity and low participation rates. Furthermore, we also conducted a hyperparameter sensitivity analysis under a more extreme heterogeneity setting with $\alpha = 0.1$. The conclusions are consistent with the above.

Table 14: Performance of different β with $\mu = 0.001$, $\lambda = 0.01$ and $\beta_0 = 0.1$.

β	0.99	0.95	0.9	0.8	0.7	0.6	0.1
Acc.	78.21	79.23	79.41	79.08	78.91	78.70	76.66

Table 15: Performance of different β_0 with $\mu = 0.001$, $\lambda = 0.01$ and $\beta = 0.9$.

β_0	0.3	0.2	0.1	0.05	0.02	0.01	0.001
Acc.	78.87	79.09	79.41	78.96	78.69	78.45	78.22

Table 16: Performance of different μ with $\beta = 0.9$, $\lambda = 0.01$ and $\beta_0 = 0.1$.

μ	0.1	0.01	0.005	0.001	0.0005	0.0001	0.00001
Acc.	75.38	78.16	79.02	79.41	79.21	78.14	76.79

Table 17: Performance of different λ with $\beta = 0.9$, $\mu = 0.001$ and $\beta_0 = 0.1$.

λ	1	0.1	0.05	0.01	0.005	0.001	0.0001
Acc.	74.88	78.76	79.04	79.41	79.32	79.22	79.10

Hyperparameter sensitivity on CIFAR-100. We further conduct hyperparameter sensitivity experiments on CIFAR-100. As shown in Table 19, 20 and 21, FedSEMA achieves the best performance at $\beta = 0.9$, $\beta_0 = 0.1$, $\mu = 0.001$, and $\lambda = 0.01$, and maintains competitive performance within neighboring values. These results suggest that the selected hyperparameter ranges are appropriate and that FedSEMA is not overly sensitive to small perturbations around the default configuration.

Table 18: Performance of different β on CIFAR-100 with $\mu = 0.001$, $\beta_0 = 0.1$ and $\lambda = 0.01$.

β	0.99	0.95	0.9	0.8	0.7	0.6	0.1
Acc.	45.33	46.34	47.06	45.82	45.36	45.31	41.55

Table 19: Performance of different β_0 on CIFAR-100 with $\mu = 0.001$, $\beta = 0.9$ and $\lambda = 0.01$.

β_0	0.3	0.2	0.1	0.05	0.02	0.01	0.001
Acc.	46.01	46.23	47.06	46.36	45.68	45.32	45.14

Table 20: Performance of different μ on CIFAR-100 with $\beta = 0.9$, $\beta_0 = 0.1$ and $\lambda = 0.01$.

μ	0.1	0.01	0.005	0.001	0.0005	0.0001	0.00001
Acc.	40.51	44.33	45.77	47.06	46.05	43.89	40.54

Table 21: Performance of different λ on CIFAR-100 with $\mu = 0.001$, $\beta_0 = 0.1$ and $\beta = 0.9$.

λ	1	0.1	0.05	0.01	0.005	0.001	0.0001
Acc.	40.44	45.40	45.81	47.06	46.54	45.96	44.91

A.7 Wall-clock Time Analysis

We further provide a wall-clock time analysis on CIFAR-100. In FL, convergence speed is commonly measured by the number of communication rounds required to reach a target accuracy, as communication is often the dominant bottleneck. Following prior work [31, 15], we report both the average time per communication round and the total wall-clock time required to reach a target accuracy.

Table 22 reports the average time per round of different algorithms on CIFAR-100. FedSEMA has a comparable per-round cost to most baselines. Table 23 further reports the wall-clock time required to reach 45% accuracy. FedSEMA achieves the shortest time among all methods that reach the target accuracy, demonstrating that its faster convergence in communication rounds also translates into favorable wall-clock efficiency. Here, "–" indicates that the method does not reach 45% accuracy within the considered training budget. Although FedSEMA introduces a slightly higher per-round cost than FedAvg and FedAdam, it reaches the target accuracy with substantially less total time due to faster convergence.

Table 22: Average time per communication round on CIFAR-100.

Method	**FedAvg**	**FedProx**	**SCAFFOLD**	**FedDyn**	**FedLNAG**	**FedAdam**	**FedCM**	**MimeLite**	**FedACG**	**FedANAG**	**FedSEMA**
Time (s)	0.358	0.401	0.424	0.467	0.521	0.366	0.415	0.485	0.404	0.416	0.446

Table 23: Wall-clock time to reach 45% accuracy on CIFAR-100.

Method	**FedAvg**	**FedProx**	**SCAFFOLD**	**FedDyn**	**FedLNAG**	**FedAdam**	**FedCM**	**MimeLite**	**FedACG**	**FedANAG**	**FedSEMA**
Time (s)	–	–	–	–	834.64	686.98	712.97	509.25	460.16	707.62	**377.76**

A.8 Different heterogeneity

To further examine the behavior of all algorithms under more extreme levels of data heterogeneity, we additionally evaluate the setting with smaller heterogeneity parameters, specifically $\alpha = 0.1$ and $\alpha = 0.2$, while keeping all other experimental configurations identical to those in Table 3. The results in Table 24 show that FedSEMA exhibits the highest robustness under conditions of severe data heterogeneity.

Table 24: Performance under more extreme heterogeneity.

α	**FedSEMA**	**FedACG**	**FedANAG**	**SCAFFOLD**	**FedDyn**	**SCAFFOLD-M**	**FedLNAG**	**FedAvgM**	**FedAdam**	**FedCM**
0.1	**81.25**	76.64	80.56	79.18	76.65	78.91	76.52	78.7	74.3	73.39
0.2	**82.59**	80.97	81.86	80.35	77.51	80.16	78.05	80.09	78.31	78.82

A.9 Feature skew

We additionally evaluate our method under feature-skew settings, following the experimental setup in [31]. We conduct the evaluation on the real-world LEAF dataset FEMNIST, which naturally contains feature-skewed client distributions. In this experiment, we set the number of clients to 2000, use user-level partitioning, randomly sample 5 clients per communication round, and adopt a two-layer CNN architecture as in [45]. The experimental results in Table 25 demonstrate that our method also outperforms the baselines under feature-skew settings, highlighting its robustness to feature-skew heterogeneity.

Table 25: Performance comparison under feature skew

Method	**Acc. (%, ↑) 250R**	**Acc. (%, ↑) 500R**	**Rounds (↓) 75%**	**Rounds (↓) 80%**
FedSEMA(ours)	**79.91**	**81.26**	**79**	**264**
FedACG	79.5	80.62	82	285
FedANAG	79.37	80.54	86	288
FedAdam	76.79	79.01	180	500+
SCAFFOLD	78.58	79.86	175	500+
FedLNAG	74.46	78.31	346	500+
FedAvg	77.29	78.74	192	500+
SCAFFOLD-M	78.81	80.13	156	435
MimeLite	78.78	80.18	179	372
FedAvgM	78.73	80.07	166	467
FedCM	75.58	79.05	226	500+
FedProx	77.79	79.33	187	500+
FedDyn	77.31	79.7	172	487

A.10 Variance measurements of top accuracy with different seeds

We re-run the experiments with different random seeds to measure the variance of the top accuracy and eliminate the impact of randomness, as reported in Table 26.

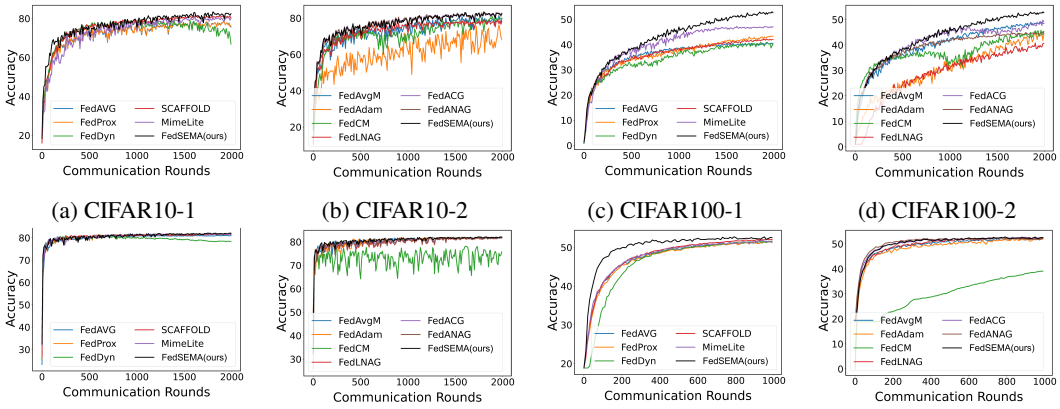
Table 26: The variance measurement of top validation accuracy that can be achieved, with 5 random seeds.

Dataset	**FedSEMA**	**FedANAG**	**FedACG**	**FedLNAG**	**FedCM**	**FedAdam**	**SCAFFOLD-M**
EMNIST	**82.25** $\pm$ 0.08	82.01 $\pm$ 0.18	82.02 $\pm$ 0.16	81.92 $\pm$ 0.35	80.16 $\pm$ 0.82	81.58 $\pm$ 0.60	81.96 $\pm$ 0.14
CIFAR-10	**84.09** $\pm$ 0.15	83.51 $\pm$ 0.17	83.52 $\pm$ 0.16	79.60 $\pm$ 0.47	81.09 $\pm$ 0.81	78.29 $\pm$ 0.34	81.94 $\pm$ 0.35
CIFAR-100	**54.58** $\pm$ 0.76	45.32 $\pm$ 0.85	50.59 $\pm$ 0.92	42.31 $\pm$ 1.64	45.02 $\pm$ 1.92	44.51 $\pm$ 1.34	48.69 $\pm$ 1.21
Shakespeare	**53.08** $\pm$ 0.14	52.46 $\pm$ 0.14	52.52 $\pm$ 0.13	52.23 $\pm$ 0.11	39.14 $\pm$ 1.89	52.30 $\pm$ 0.18	52.26 $\pm$ 0.13

Dataset	**FedAvgM**	**MimeLite**	**SCAFFOLD**	**FedDyn**	**FedProx**	**FedAvg**
EMNIST	81.92 $\pm$ 0.19	81.91 $\pm$ 0.14	81.55 $\pm$ 0.15	81.02 $\pm$ 0.65	81.83 $\pm$ 0.56	81.01 $\pm$ 0.53
CIFAR-10	81.86 $\pm$ 0.37	81.02 $\pm$ 0.23	81.34 $\pm$ 0.34	80.14 $\pm$ 0.61	78.88 $\pm$ 0.51	78.64 $\pm$ 0.39
CIFAR-100	48.31 $\pm$ 1.67	47.59 $\pm$ 0.80	42.24 $\pm$ 1.18	41.89 $\pm$ 1.19	41.74 $\pm$ 1.12	40.79 $\pm$ 1.34
Shakespeare	52.32 $\pm$ 0.08	51.71 $\pm$ 0.31	52.04 $\pm$ 0.25	51.77 $\pm$ 0.23	51.54 $\pm$ 0.21	51.61 $\pm$ 0.19

A.11 Convergence and top validation accuracy

The experimental results demonstrate that our algorithm consistently outperforms the majority of existing optimization algorithms proposed to address data heterogeneity as shown in Fig 4 and Fig 5.



(a) CIFAR10-1 (b) CIFAR10-2 (c) CIFAR100-1 (d) CIFAR100-2

(e) emnist-1 (f) emnist-2 (g) shakespeare-1 (h) shakespeare-2

Figure 4: Convergence, top validation accuracy for various FL methods in the different datasets.

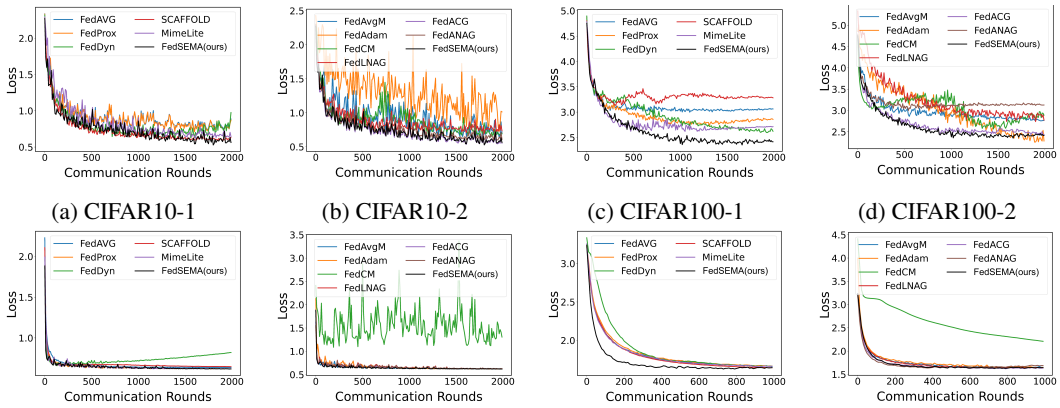
A.12 Ablation Studies

Compared with the FedAvg, FedSEMA adds two main modules: (1) global momentum, (2) hybrid proximal term. We test the performance of 2000 communication rounds of the different combination of the modules above on the CIFAR-10 and CIFAR-100 with the settings of 2% participating ratio of total 200 clients. The Table 27 shows their performance.

A.13 Conceptual illustration of stability

Fig. 6 provides a conceptual illustration of stability. In Fig. 6a, each curve represents the local update trajectory of a client. Stronger local-update stability implies that variations in the set of clients selected for local training have a smaller impact on the aggregated model. In Fig. 6b, each curve denotes the global update trajectory of an FL task. Stronger global-update stability indicates that changes in the participating client set exert less influence on the final global model.

For local-update stability, under the same learning rate, FedSEMA achieves a smaller maximum local model discrepancy than FedAvg and remains comparable to FedProx. FedSEMA (blue solid line)



(a) CIFAR10-1 (b) CIFAR10-2 (c) CIFAR100-1 (d) CIFAR100-2

(e) emnist-1 (f) emnist-2 (g) shakespeare-1 (h) shakespeare-2

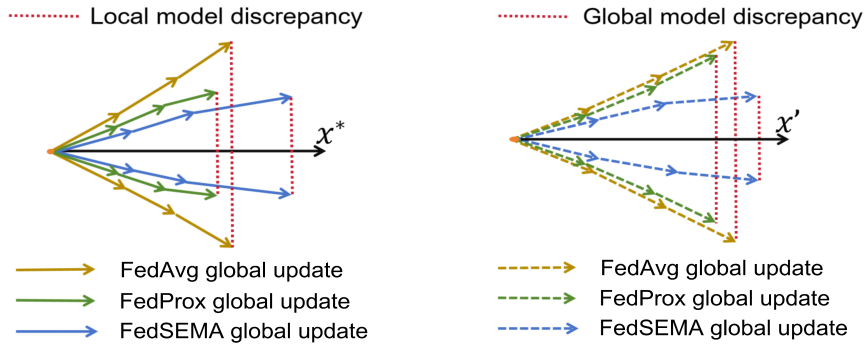
Figure 5: Convergence, loss for various FL methods in the different datasets.

Table 27: Ablation studies on different modules.

Dataset	Momentum	Hybrid proximal term	Accuracy
CIFAR-10	✓	×	83.05
	×	✓	80.54
	✓	✓	84.09
CIFAR-100	✓	×	50.16
	×	✓	44.73
	✓	✓	54.58

consistently updates with a larger step size than FedProx (green solid line), thereby moving closer to the global optimum of the participating clients. This enables superior performance while maintaining the same maximum model discrepancy. As illustrated, the updates of FedSEMA are closest to X^* (black solid line), followed by FedProx, while FedAvg remains the farthest. When the learning rate is set to $\eta_l = \mathcal{O}\left(\frac{1}{LK\sqrt{T}}\right)$, FedProx, FedAvg, and FedSEMA achieve the same maximum model discrepancy; however, due to momentum acceleration, FedSEMA still outperforms FedProx.

For global-update stability, regardless of whether the learning rate is fixed or set as $\eta_l = \mathcal{O}\left(\frac{1}{LK\sqrt{T}}\right)$, FedSEMA consistently yields a smaller global model discrepancy compared to both FedAvg and FedProx. FedSEMA (blue dashed line) updates with a larger step size than FedProx (green dashed line), and thus, while maintaining a smaller maximum model discrepancy, it approaches the global optimum of all clients more closely.



(a) Conceptual illustration of the local-update stability.

(b) Conceptual illustration of the global-update stability.

Figure 6: Conceptual illustration of stability. X^* is the the global optimum of the participating clients; X' is the global optimum of all clients.

B Proof of theorem 1

In this section, we provide the proof of Theorem 1.
Given that $\mathcal{S}$ and $\mathcal{S}'$ only differ by one client. According to the definition of R_e and R_e, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}\left[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}),\xi_i)]\right] \\
&=\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}\left[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}),\xi_i)]\right] \\
&=\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}\left[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}(\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}),\xi_i)]-\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}'),\xi_i)])\right] \\
&\quad+\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}\left[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}'),\xi_i)]\right] \\
&=\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}\left[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}(\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}),\xi_i)]-\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}'),\xi_i)])\right] \\
&\quad+\mathbb{E}_{\mathcal{S}'}\mathbb{E}_{\mathcal{A}}\left[\mathbb{E}_{i\sim\mathcal{C}}[\mathbb{E}_{\xi_i\sim\mathcal{P}_i}[F_i(A(\mathcal{S}'),\xi_i)]]\right].
\end{aligned}$$

Since $\mathcal{A}$ satisfies ϵ-global update stability, rearranging the above equality we can get

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{A}}[|R_u(\mathcal{A}(\mathcal{S}))-R_e(\mathcal{A}(\mathcal{S}))|]\leq\epsilon.$$

C The equivalent form of the NAG.

Recall the original form of NAG:

$$\begin{aligned}
g_k &= \nabla f(x_k-\eta_l\beta v_{k-1}) \\
v_k &= \beta v_{k-1}+g_k \\
x_{k+1} &= x_k-\eta_l v_k.
\end{aligned}$$

From the second and third equations, we can obtain

$$\begin{aligned}
x_{k+1}-\eta_l\beta v_k &= x_k-\eta_l(1+\beta)v_k \\
&= x_k-\eta_l(1+\beta)(\beta v_{k-1}+g_k) \\
&= x_k-\eta_l\beta v_{k-1}-\eta_l\beta^2 v_{k-1}-\eta_l(1+\beta)\nabla f(x_k-\eta_l\beta v_{k-1}).
\end{aligned}$$

Denote $\hat{x}_k := x_k-\eta_l\beta v_{k-1}$ and $\hat{v}_k=\beta^2 v_{k-1}+(1+\beta)\nabla f(x_k-\eta_l\beta v_{k-1})$, then we have

$$\begin{aligned}
\hat{v}_k &= \beta^2 v_{k-1}+(1+\beta)\nabla f(\hat{x}_k) \\
\hat{x}_{k+1} &= \hat{x}_k-\eta_l\hat{v}_k.
\end{aligned}$$

Note that $v_k=\beta v_{k-1}+g_k$, we get

$$\begin{aligned}
\hat{v}_k-\beta\hat{v}_{k-1} &= \beta^2 v_{k-1}+(1+\beta)\nabla f(\hat{x}_k)-\beta^3 v_{k-2}-\beta(1+\beta)\nabla f(\hat{x}_{k-1}) \\
&= \beta^3 v_{k-2}+(1+\beta)\nabla f(\hat{x}_k)-\beta^3 v_{k-2}-\beta\nabla f(\hat{x}_{k-1}) \\
&= \nabla f(\hat{x}_k)+\beta(\nabla f(\hat{x}_k)-\nabla f(\hat{x}_{k-1})).
\end{aligned}$$

Now we get

$$\begin{aligned}
\hat{v}_k &= \beta\hat{v}_{k-1}+\nabla f(\hat{x}_k)+\beta(\nabla f(\hat{x}_k)-\nabla f(\hat{x}_{k-1})) \\
\hat{x}_{k+1} &= \hat{x}_k-\eta_l\hat{v}_k.
\end{aligned}$$

We substitute v_{k-1} with the global momentum and change the coefficient in front of the gradient difference to β_0, and we obtain our local update rule.

D Stability upper bound for non-convex losses

Lemma 1. *Suppose Assumption 1-3 hold. Then*

$$\mathbb{E}\|\nabla F_i(x_{t,k}^i;\xi_i)\| \leq (\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}, \forall i \in [M], \forall 0 \leq k \leq K-1.$$

Proof.

$$\begin{aligned}\mathbb{E}\|\nabla F_i(x_{t,k}^i)\|^2 &= \mathbb{E}\|\nabla F_i(x_{t,k}^i) - \nabla f(x_{t,k}^i) + \nabla f(x_{t,k}^i)\|^2 \\ &\leq 2\mathbb{E}\|\nabla F_i(x_{t,k}^i) - \nabla f(x_{t,k}^i)\|^2 + 2\mathbb{E}\|\nabla f(x_{t,k}^i)\|^2.\end{aligned}$$

Since

$$\begin{aligned}&\mathbb{E}\|\nabla F_i(x_{t,k}^i,\xi_i) - \nabla F_i(x_{t,k}^i)\|^2 \\ &= \mathbb{E}\|\nabla F_i(x_{t,k}^i,\xi_i)\|^2 + \mathbb{E}\|\nabla F_i(x_{t,k}^i)\|^2 - 2\langle \mathbb{E}\nabla F_i(x_{t,k}^i,\xi_i), \nabla F_i(x_{t,k}^i)\rangle \\ &= \mathbb{E}\|\nabla F_i(x_{t,k}^i,\xi_i)\|^2 - \mathbb{E}\|\nabla F_i(x_{t,k}^i)\|^2,\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E}\|\nabla F_i(x_{t,k}^i,\xi_i)\|^2 \\ &\leq \mathbb{E}\|\nabla F_i(x_{t,k}^i,\xi_i) - \nabla F_i(x_{t,k}^i)\|^2 + 2\mathbb{E}\|\nabla F_i(x_{t,k}^i) - \nabla f(x_{t,k}^i)\|^2 + 2\mathbb{E}\|\nabla f(x_{t,k}^i)\|^2 \\ &\leq \sigma_l^2 + 2\sigma_g^2 + 2Z^2.\end{aligned}$$

The last inequality follows from Assumption 3, by the Lipschitz continuity, we can obtain that $\nabla f \leq Z$. □

D.1 Analysis for FedAvg under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l \nabla F_i(x_{t,k}^i,\xi_i).$$

Theorem 4. *(global update stability of FedAvg). Suppose Assumption 1-3 hold and consider FedAvg. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update formula, we have

$$\begin{aligned}&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\ &= \mathbb{E}\|x_{t,k-1}^i - \eta_l \nabla F_i(x_{t,k-1}^i,\xi_i) - (x_{t,k-1}^j - \eta_l \nabla F_j(x_{t,k-1}^j,\xi_j))\| \\ &\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + \eta_l \mathbb{E}\|\nabla F_i(x_{t,k-1}^i,\xi_i)\| + \eta_l \mathbb{E}\|\nabla F_j(x_{t,k-1}^j,\xi_j)\| \\ &\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l,\end{aligned}$$

where the last inequality we use Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq \mathbb{E}\|x_t - x_t'\| + 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l K.$$

Note that in the local stability analysis of [15], it holds that $\mathbb{E}\|x_t - x_t'\| = 0$. Therefore, the local-update stability of the FedAvg algorithm is given by $\mathcal{O}\Big(2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l K\Big)$. Similarly, the local-update stability of subsequent algorithms can be derived from $\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\|$. Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x_{t,K}^i - x_t^i)$, we have

$$\begin{aligned}\mathbb{E}\|x_{t+1} - x_{t+1}'\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x_{t,K}^i - \sum_{j\in\mathcal{S}'} x_{t,K}^j\| \\ &\leq \mathbb{E}\|x_t - x_t'\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.\end{aligned}$$

Note that $x_0 = x_0'$, then unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.2 Analysis for FedProx under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l(\nabla F_i(x_{t,k}^i, \xi_i) + \lambda(x_{t,k}^i - x_t)).$$

Theorem 5. *(global update stability of of FedProx). Suppose Assumption 1-3 hold and consider FedProx. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{2}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1 - (1 - \lambda\eta_l)^K]T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\
&= \mathbb{E}\|(1 - \eta_l\lambda)(x_{t,k-1}^i - x_{t,k-1}^j) - \eta_l\nabla F_i(x_{t,k-1}^i, \xi_i) + \eta_l\nabla F_j(x_{t,k-1}^j, \xi_j) + \lambda\eta_l(x_t - x_t')\| \\
&\leq \mathbb{E}\|(1 - \eta_l\lambda)(x_{t,k-1}^i - x_{t,k-1}^j)\| + \eta_l\mathbb{E}\|\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)\| + \lambda\eta_l\mathbb{E}\|x_t - x_t'\| \\
&\leq (1 - \lambda\eta_l)\mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l + \lambda\eta_l\mathbb{E}\|x_t - x_t'\|,
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives:

$$\begin{aligned}
\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| &\leq (1 - \lambda\eta_l)^K\mathbb{E}\|x_t - x_t'\| + \sum_{k=0}^{K-1}(1 - \lambda\eta_l)^k[2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l \\
&\quad + \lambda\eta_l\mathbb{E}\|x_t - x_t'\|] \\
&= (1 - \lambda\eta_l)^K\mathbb{E}\|x_t - x_t'\| + \frac{1 - (1 - \lambda\eta_l)^K}{\lambda\eta_l}[2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l \\
&\quad + \lambda\eta_l\mathbb{E}\|x_t - x_t'\|] \\
&= \mathbb{E}\|x_t - x_t'\| + \frac{2}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1 - (1 - \lambda\eta_l)^K].
\end{aligned}$$

Since $x_{t+1} = \frac{1}{s}\sum_{i\in\mathcal{S}} x_{t,K}^i$, we obtain

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_{t+1}'\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x_{t,K}^i - \sum_{j\in\mathcal{S}'} x_{t,K}^j\| \\
&\leq \mathbb{E}\|x_t - x_t'\| + \frac{2}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1 - (1 - \lambda\eta_l)^K].
\end{aligned}$$

Note that $x_0 = x_0'$, then unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{2}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1 - (1 - \lambda\eta_l)^K]T.$$

□

D.3 Analysis for SCAFFOLD under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l(\nabla F_i(x_{t,k}^i, \xi_i) - c_i + c),$$

where $c_i = \frac{1}{K}\sum_{k=0}^{K-1}\nabla F(x_{t-\tau,k}^i, \xi_i)$ is the control variate of the client i whose last participation in training was in the $t - \tau$ round. Note that, unlike other algorithms, the local update here involves gradient information from $t - \tau$ rounds.

Theorem 6. *(global update stability of SCAFFOLD). Suppose Assumption 1-3 hold and consider SCAFFOLD. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq 6\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\
&= \mathbb{E}\|x_{t,k-1}^i - \eta_l(\nabla F_i(x_{t,k-1}^i, \xi_i) - c_i + c) - x_{t,k-1}^j + \eta_l(\nabla F_j(x_{t,k-1}^j, \xi_j) - c_j + c')\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + \eta_l \mathbb{E}\|c_i - c_j\| + \eta_l E\|\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)\| \\
&\quad + \eta_l \mathbb{E}\|c - c'\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 6\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq \mathbb{E}\|x_t - x_t'\| + 6(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}\eta_l K.$$

Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x_{t,K}^i - x_t^i)$, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_{t+1}'\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x_{t,K}^i - \sum_{j\in\mathcal{S}'} x_{t,K}^j\| \\
&\leq \mathbb{E}\|x_t - x_t'\| + 6\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

Note that $x_0 = x_0'$, then unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq 6\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.4 Analysis for FedDyn under convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l(\nabla F_i(x_{t,k}^i, \xi_i) - \nabla F_i(x_{t-\tau,K}^i, \xi_i) + \lambda(x_{t,k}^i - x_t)),$$

where the $t - \tau$ is because clients do not participate in training every round and recently participated in round $t - \tau$.

Theorem 7. *(global update stability of FedDyn). Suppose Assumption 1-3 hold and consider FedDyn. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{4}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1 - (1 - \lambda\eta)^K]T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\
&= \mathbb{E}\|x_{t,k-1}^i - \eta_l(\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_i(x_{t-\tau,K}^i, \xi_i) + \lambda(x_{t,k-1}^i - x_t)) - x_{t,k-1}^j \\
&\quad + \eta_l(\nabla F_j(x_{t,k-1}^j, \xi_j) - \nabla F_j(x_{t-\tau,K}^j, \xi_j) + \lambda(x_{t,k-1}^j - x_t'))\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j - \lambda\eta_l(x_{t,k-1}^i - x_{t,k-1}^j)\| + \eta_l \mathbb{E}\|\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)\| \\
&\quad + \eta_l \mathbb{E}\|\nabla F_i(x_{t-\tau,K}^i, \xi_i) - \nabla F_j(x_{t-\tau,K}^j, \xi_j)\| + \lambda\eta_l \mathbb{E}\|x_t - x_t'\| \\
&\leq (1 - \lambda\eta_l)\mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 4\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} + \lambda\eta_l \mathbb{E}\|x_t - x_t'\|,
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq \mathbb{E}\|x_t - x_t'\| + \frac{4}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1 - (1 - \lambda\eta_l)^K].$$

Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x^i_{t,K} - x^i_{t,0})$, we obtain

$$\begin{aligned}\mathbb{E}\|x_{t+1} - x'_{t+1}\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x^i_{t,K} - \sum_{j\in\mathcal{S}'} x^j_{t,K}\| \\ &\leq \mathbb{E}\|x_t - x'_t\| + \frac{4}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1-(1-\lambda\eta_l)^K].\end{aligned}$$

Note that $x_0 = x'_0$, then unrolling it gives

$$\mathbb{E}\|x_T - x'_T\| \leq \frac{4}{\lambda}(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}[1-(1-\lambda\eta_l)^K]T.$$

□

D.5 Analysis for FedLNAG under non-convex loss

Consider the orignial update rule of NAG:

$$\begin{aligned} y^i_{t,k} &= x^i_{t,k} - \beta m^i_{t,k} \\ m^i_{t,k+1} &= \beta m^i_{t,k} + (1-\beta)\nabla F_i(y^i_{t,k}, \xi_i) \\ x^i_{t,k+1} &= x^i_{t,k} - \eta_l m^i_{t,k+1}. \end{aligned}$$

Theorem 8. *(global update stability of FedLNAG). Suppose Assumption 1-3 hold and consider FedLNAG. Let x_T and x'_T be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x'_T\| \leq 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. For the momentum term $m_{t,k}$, by recursion we have

$$m_{t,k} = (1-\beta)\sum_{\tau=0}^{k-1}\beta^\tau \nabla F(y_{t,k-\tau-1}, \xi).$$

Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x^i_{t,k+1} - x^j_{t,k+1}\| \\
&= \mathbb{E}\|x^i_{t,k} - \eta_l m^i_{t,k+1} - (x^j_{t,k} - \eta_l m^j_{t,k+1})\| \\
&\leq \mathbb{E}\|x^i_{t,k} - x^j_{t,k}\| + \eta_l\|m^i_{t,k+1}\| + \eta_l\|m^j_{t,k+1}\| \\
&\leq \mathbb{E}\|x^i_{t,k} - x^j_{t,k}\| + \eta_l(1-\beta)\sum_{\tau=0}^{k}\beta^\tau\|\nabla F_i(y^i_{t,k-\tau}, \xi_i)\| + \eta_l(1-\beta)\sum_{\tau=0}^{k}\beta^\tau\|\nabla F_j(y^j_{t,k-\tau}, \xi_j)\| \\
&\leq \mathbb{E}\|x^i_{t,k} - x^j_{t,k}\| + 2\eta_l(1-\beta^{k+1})(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives

$$\mathbb{E}\|x^i_{t,K} - x^j_{t,K}\| \leq \mathbb{E}\|x_t - x'_t\| + 2\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}K.$$

Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x^i_{t,K} - x^i_{t,0})$, we obtain

$$\begin{aligned}\mathbb{E}\|x_{t+1} - x'_{t+1}\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x^i_{t,K} - \sum_{j\in\mathcal{S}'} x^j_{t,K}\| \\ &\leq \mathbb{E}\|x_t - x'_t\| + 2\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}K.\end{aligned}$$

Note that $x_0 = x'_0$, then unrolling it gives

$$\mathbb{E}\|x_T - x'_T\| \leq 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.6 Analysis for FedAdam under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l \nabla F_i(x_{t,k}^i, \xi_i).$$

Theorem 9. *(global update stability of FedAdam). Suppose Assumption 1-3 hold and consider FedAdam. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{2}{\tau}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Firstly,

$$\begin{aligned}
\Delta_t^i &= x_{t,K}^i - x_t \\
&= x_{t,K-1}^i - \eta_l \nabla F_i(x_{t,K-1}^i, \xi_i) - x_t \\
&= \sum_{k=0}^{K-1} -\eta_l \nabla F_i(x_{t,k}^i, \xi_i),
\end{aligned}$$

By using Lemma 1 we get

$$\mathbb{E}\|\Delta_t^i\| \leq \eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

Note that $\Delta_t = \beta_1 \Delta_{t-1} + (1-\beta_1)(\frac{1}{s}\sum_{i\in\mathcal{S}} \Delta_t^i)$, so we can deduce that

$$\mathbb{E}\|\Delta_t\| \leq \eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

since $x_{t+1} = x_t + \frac{\Delta_t}{\sqrt{v_t}+\tau}$, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_{t+1}'\| &= \mathbb{E}\|x_t + \frac{\Delta_t}{\sqrt{v_t}+\tau} - (x_t' + \frac{\Delta_t'}{\sqrt{v_t'}+\tau})\| \\
&\leq \mathbb{E}\|x_t - x_t'\| + \mathbb{E}\|\frac{\Delta_t}{\sqrt{v_t}+\tau}\| + \mathbb{E}\|\frac{\Delta_t'}{\sqrt{v_t'}+\tau}\| \\
&\leq \mathbb{E}\|x_t - x_t'\| + \mathbb{E}\|\frac{\Delta_t}{\tau}\| + \mathbb{E}\|\frac{\Delta_t'}{\tau}\| \\
&\leq \mathbb{E}\|x_t - x_t'\| + \frac{2}{\tau}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

Note that $x_0 = x_0'$, then unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{2}{\tau}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.7 Analysis for FedCM under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l(\alpha \nabla F_i(x_{t,k}^i, \xi_i) + (1-\alpha)m_t),$$

where $\alpha \in [0, 1]$ is momentum parameter.

Theorem 10. *(global update stability of FedCM). Suppose Assumption 1-3 hold and consider FedCM. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. Accoring to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\
&= \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j - \eta_l\alpha(\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)) - \eta_l(1-\alpha)(m_t - m_t')\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + \eta_l\alpha\mathbb{E}\|\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)\| + \eta_l(1-\alpha)\mathbb{E}\|m_t - m_t'\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 2\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq \mathbb{E}\|x_t - x_t'\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

Since $x_{t+1} = x_t + \frac{1}{\eta_l K s}\sum_{i\in\mathcal{S}}(x_{t,K}^i - x_t)$, we obtain

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_{t+1}'\| &\leq \mathbb{E}\|(1 - \frac{1}{\eta_l K})(x_t - x_t')\| + \frac{1}{\eta_l K s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x_{t,K}^i - \sum_{j\in\mathcal{S}'} x_{t,K}^j\| \\
&\leq (1 - \frac{1}{\eta_l K})\mathbb{E}\|x_t - x_t'\| + \frac{1}{\eta_l K}\mathbb{E}\|x_t - x_t'\| + 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} \\
&= \mathbb{E}\|x_t - x_t'\| + 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

Note that $x_0 = x_0'$, unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq 2(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.8 Analysis for MimeLite under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l((1-\beta)\nabla F_i(x_{t,k}^i, \xi_i) + \beta m),$$

where $\beta \in [0, 1]$ is momentum parameter.

Theorem 11. *(global update stability of MimeLite). Suppose Assumption 1-3 hold and consider MimeLite. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\
&= \mathbb{E}\|x_{t,k-1}^i - \eta_l(1-\beta)\nabla F_i(x_{t,k-1}^i, \xi_i) - (x_{t,k-1}^j - \eta_l(1-\beta)\nabla F_j(x_{t,k-1}^j, \xi_j))\| \\
&\quad + \beta\eta_l\mathbb{E}\|m - m'\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + \eta_l(1-\beta)\mathbb{E}\|\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)\| + \eta_l\beta\mathbb{E}\|m - m'\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 2\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq \mathbb{E}\|x_t - x_t'\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

Since $x_{t+1} = \frac{1}{s}\sum_{i\in\mathcal{S}} x_{t,K}^i$, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_{t+1}'\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x_{t,K}^i - \sum_{j\in\mathcal{S}'} x_{t,K}^j\| \\
&\leq \mathbb{E}\|x_t - x_t'\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

Note that $x_0 = x_0'$, unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.9 Analysis for FedACG under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l \nabla F_i(x_{t,k}^i, \xi_i).$$

Theorem 12. *(global update stability of FedACG). Suppose Assumption 1-3 hold and consider FedACG. Let x_T and x_T' be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{4-2\lambda}{1-\lambda}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x_{t,k}^i - x_{t,k}^j\| \\
&= \mathbb{E}\|x_{t,k-1}^i - \eta_l \nabla F_i(x_{t,k-1}^i, \xi_i) - (x_{t,k-1}^j - \eta_l \nabla F_j(x_{t,k-1}^j, \xi_j))\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + \eta_l \mathbb{E}\|\nabla F_i(x_{t,k-1}^i, \xi_i) - \nabla F_j(x_{t,k-1}^j, \xi_j)\| \\
&\leq \mathbb{E}\|x_{t,k-1}^i - x_{t,k-1}^j\| + 2\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},
\end{aligned}$$

where the last inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x_{t,K}^i - x_{t,K}^j\| \leq \mathbb{E}\|x_{t,0}^i - x_{t,0}^j\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

Since $x_{t,0}^i = x_{t-1} + \lambda m_{t-1}$ and $m_t = \lambda m_{t-1} + \frac{1}{s}\sum_{i\in\mathcal{S}}(x_{t,K}^i - x_{t,0}^i)$, then

$$\begin{aligned}
x_t &= x_{t-1} + m_t \\
&= x_{t-1} + \lambda m_{t-1} + \frac{1}{s}\sum_{i\in\mathcal{S}}(x_{t,K}^i - x_{t,0}^i) \\
&= \frac{1}{s}\sum_{i\in\mathcal{S}} x_{t,K}^i,
\end{aligned}$$

so we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x_{t+1}'\| &= \frac{1}{s}\mathbb{E}]\|\sum_{i\in\mathcal{S}} x_{t+1,K}^i - \sum_{j\in\mathcal{S}'} x_{t+1,K}^j\| \\
&\leq \mathbb{E}\|x_t + \lambda m_t - (x_t' + \lambda m_t')\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} \\
&\leq \mathbb{E}\|x_t - x_t'\| + \lambda\mathbb{E}\|m_t - m_t'\| + 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} \\
&\leq \mathbb{E}\|x_t - x_t'\| + \left(\frac{1-\lambda^t}{1-\lambda} + 1\right) 2\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} \\
&\leq \mathbb{E}\|x_t - x_t'\| + \frac{4-2\lambda}{1-\lambda}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

Note that $x_0 = x_0'$, then unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{4-2\lambda}{1-\lambda}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.10 Analysis for SCAFFOLD-M under non-convex losses

The local update at iteration k is described as follows:

$$x_{t,k+1}^i = x_{t,k}^i - \eta_l(\beta(\nabla F_i(x_{t,k}^i, \xi_i) - c_t^i + c_t) + (1-\beta)v_t),$$

where $\beta \in (0, 1)$ is the momentum coefficients.

Theorem 13. *(global updaye stability of SCAFFOLD-M). Suppose Assumption 1-3 hold and consider SCAFFOLD-M. Let x_T and x'_T be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L- smooth function, then*

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x^i_{t,k} - x^j_{t,k}\| \\
&= \mathbb{E}\|x^i_{t,k-1} - \eta_l(\beta(\nabla F_i(x^i_{t,k-1}, \xi_i) - c^i_t + c_t) + (1-\beta)v_t) - x^j_{t,k-1} \\
&\quad + \eta_l(\beta(\nabla F_j(x^j_{t,k-1}, \xi_j) - c^j_t + c'_t) + (1-\beta)v'_t)\| \\
&\leq \mathbb{E}\|x^i_{t,k-1} - x^j_{t,k-1}\| + \eta_l\beta\mathbb{E}\|\nabla F_i(x^i_{t,k-1}, \xi_i) - \nabla F_j(x^j_{t,k-1}, \xi_j)\| + \eta_l\beta\mathbb{E}\|c^i_t - c^j_t\| \\
&\quad + \eta_l\beta\mathbb{E}\|c_t - c'_t\| + \eta_l(1-\beta)\mathbb{E}\|v_t - v'_t\| \\
&\leq \mathbb{E}\|x^i_{t,k-1} - x^j_{t,k-1}\| + (2+4\beta)\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},
\end{aligned}$$

where the second inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x^i_{t,K} - x^j_{t,K}\| \leq \mathbb{E}\|x_t - x'_t\| + (2+4\beta)\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x^i_{t,K} - x^i_{t,0})$, we have

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x'_{t+1}\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x^i_{t,K} - \sum_{j\in\mathcal{S}'} x^j_{t,K}\| \\
&\leq \mathbb{E}\|x_t - x'_t\| + (2+4\beta)\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.
\end{aligned}$$

Note that $x_0 = x'_0$, then unrolling it gives

$$\mathbb{E}\|x_T - x'_T\| \leq (2+4\beta)\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.11 Analysis for FedANAG under non-convex losses

The local update at iteration k is described as follows:

$$x^i_{t,k+1} = x^i_{t,k} - \eta_l((1+\beta)\nabla F_i(x^i_{t,k}, \xi_i) + \beta^2 m_t).$$

Theorem 14. *(global update stability of FedANAG). Suppose Assumption 1-3 hold and consider FedANAG. Let x_T and x'_T be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x'_T\| \leq \frac{2}{1-\beta}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}\|x^i_{t,k} - x^j_{t,k}\| \\
&= \mathbb{E}\|x^i_{t,k-1} - \eta_l(1+\beta)\nabla F_i(x^i_{t,k-1}, \xi_i) - (x^j_{t,k-1} - \eta_l(1+\beta)\nabla F_j(x^j_{t,k-1}, \xi_j)) \\
&\quad - \eta_l\beta^2(m_t - m'_t)\| \\
&\leq \mathbb{E}\|x^i_{t,k-1} - x^j_{t,k-1}\| + \eta_l(1+\beta)\mathbb{E}\|\nabla F_i(x^i_{t,k-1}, \xi_i) - \nabla F_j(x^j_{t,k-1}, \xi_j)\| \\
&\quad + \eta_l\beta^2\mathbb{E}\|m_t - m'_t\| \\
&\leq \mathbb{E}\|x^i_{t,k-1} - x^j_{t,k-1}\| + \left(1 + \beta + \frac{\beta^2(1-\beta^t)}{1-\beta}\right)2\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} \\
&\leq \mathbb{E}\|x^i_{t,k-1} - x^j_{t,k-1}\| + \frac{2}{1-\beta}\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},
\end{aligned}$$

where the second inequality follows Lemma 1. Unrolling it gives:

$$\mathbb{E}\|x^i_{t,K} - x^j_{t,K}\| \leq \mathbb{E}\|x_t - x'_t\| + \frac{2}{1-\beta}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.$$

Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x^i_{t,K} - x^i_{t,0})$, we have

$$\begin{aligned}\mathbb{E}\|x_{t+1} - x'_{t+1}\| &= \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x^i_{t,K} - \sum_{j\in\mathcal{S}'} x^j_{t,K}\| \\ &\leq \mathbb{E}\|x_t - x'_t\| + \frac{2}{1-\beta}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.\end{aligned}$$

Note that $x_0 = x'_0$, then unrolling it gives

$$\mathbb{E}\|x_T - x'_T\| \leq \frac{2}{1-\beta}\eta_l K(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}T.$$

□

D.12 Analysis for FedSEMA under non-convex losses

The local update at iteration k is described as follows:

$$x^i_{t,k+1} = x^i_{t,k} - \eta_l(\nabla F_i(x^i_{t,k}, \xi_i) + \beta_0\delta_{g_t} + \beta v_t + \lambda(x^i_{t,k} - x^i_{t,0}) + \mu x^i_{t,k}),$$

where $\delta_{g_k} = \nabla F_i(x^i_{t,k}, \xi_i) - \nabla F_i(x^i_{t,k-1}, \xi_i)$ and $\beta \in (0,1)$ is the momentum coefficients.

Theorem 15. *(global update stability of FedSEMA). Suppose Assumption 1-3 hold and consider FedSEMA. Let x_T and x'_T be two model results obtained by neighboring active client-sets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Under the assumption that F_i is a non-convex and L-smooth function, then*

$$\mathbb{E}\|x_T - x'_T\| \leq \frac{1-(1-\Gamma)^T}{\Gamma}\zeta(2 + 2\beta + 4\beta_0)(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},$$

where $\Gamma = 1 - \left[\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}(1-\eta_l(\lambda+\mu))^K\right] \in (0,1)$ *and* $\zeta = \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}$.

Proof. Consider two clients $i \in \mathcal{S}$ and $j \in \mathcal{S}'$. According to the local update formula, we have

$$\begin{aligned}&\mathbb{E}\|x^i_{t,k} - x^j_{t,k}\| \\ &= \mathbb{E}\|x^i_{t,k-1} - \eta_l\nabla F_i(x^i_{t,k-1}, \xi_i) - \eta_l(\lambda+\mu)x^i_{t,k-1} - \beta\eta_l v_t - (x^j_{t,k} - \eta_l\nabla F_j(x^j_{t,k-1}, \xi_j) \\ &\quad - \eta_l(\lambda+\mu)x^j_{t,k-1} - \beta\eta_l v'_t) + \eta_l\beta_0(\delta_{g_{k-1}} - \delta'_{g_{k-1}}) + \eta_l\lambda(x_t - x'_t)\| \\ &\leq \mathbb{E}\|(1-\eta_l(\lambda+\mu))(x^i_{t,k-1} - x^j_{t,k-1})\| + (1+\beta_0)\eta_l\mathbb{E}\|\nabla F_i(x^i_{t,k-1}, \xi_i) - \nabla F_j(x^j_{t,k-1}, \xi_j)\| \\ &\quad + \beta_0\eta_l\mathbb{E}\|\nabla F_i(x^i_{t,k-2}, \xi_i) - \nabla F_i(x^j_{t,k-2}, \xi_j)\| + \beta\eta_l\mathbb{E}\|v_t - v'_t\| + \eta_l\lambda\mathbb{E}\|x_t - x'_t\| \\ &\leq (1-\eta_l(\lambda+\mu))\mathbb{E}\|x^i_{t,k-1} - x^j_{t,k-1}\| + \eta_l\lambda\mathbb{E}\|x_t - x'_t\| + (2 + 2\beta + 4\beta_0)\eta_l(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}},\end{aligned}$$

where the last inequality we use Lemma 1. Unrolling it gives:

$$\begin{aligned}\mathbb{E}\|x^i_{t,K} - x^j_{t,K}\| &\leq (1-\eta_l(\lambda+\mu))^K\mathbb{E}\|x_t - x'_t\| + \frac{\lambda - \lambda(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}\mathbb{E}\|x_t - x'_t\| \\ &\quad + \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}(2 + 2\beta + 4\beta_0)(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}} \\ &= \left[\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}(1-\eta_l(\lambda+\mu))^K\right]\mathbb{E}\|x_t - x'_t\| \\ &\quad + \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}(2 + 2\beta + 4\beta_0)(\sigma_l^2 + 2\sigma_g^2 + 2Z^2)^{\frac{1}{2}}.\end{aligned}$$

We define $\zeta = \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}$. Since $x_{t+1} = x_t + \frac{1}{s}\sum_{i\in\mathcal{S}}(x^i_{t,K} - x^i_{t,0})$, we have

$$\mathbb{E}\|x_{t+1} - x'_{t+1}\| = \frac{1}{s}\mathbb{E}\|\sum_{i\in\mathcal{S}} x^i_{t,K} - \sum_{j\in\mathcal{S}'} x^j_{t,K}\|$$

$$\leq (1-\Gamma)\mathbb{E}\|x_t - x_t'\| + (2+2\beta+4\beta_0)\zeta(\sigma_l^2+2\sigma_g^2+2Z^2)^{\frac{1}{2}},$$

where $\Gamma = 1 - \left[\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}(1-\eta_l(\lambda+\mu))^K\right] \in (0,1)$. Note that $x_0 = x_0'$, then unrolling it gives

$$\mathbb{E}\|x_T - x_T'\| \leq \frac{1-(1-\Gamma)^T}{\Gamma}\zeta(2+2\beta+4\beta_0)(\sigma_l^2+2\sigma_g^2+2Z^2)^{\frac{1}{2}}.$$

D.13 Discussion of stability from the perspective of the learning rate in convergence analysis.

Our full stability bound is $\mathcal{O}\left(\frac{1-(1-\Gamma)^T}{\Gamma}\zeta\right)$. Note that

$$\begin{aligned}\Gamma &= 1 - \left(\frac{\lambda}{\lambda+\mu} + \frac{\mu}{\lambda+\mu}(1-\eta_l(\lambda+\mu))^K\right) \\ &= \frac{\mu}{\lambda+\mu}\left(1-(1-\eta_l(\lambda+\mu))^K\right).\end{aligned}$$

Since $\eta_l = \mathcal{O}(\frac{1}{LK\sqrt{T}})$, for T large enough we have

$$\Gamma \approx \frac{\mu}{\lambda+\mu}(K\eta_l(\lambda+\mu)) \approx \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Then we have

$$\frac{1-(1-\Gamma)^T}{\Gamma} \approx \mathcal{O}(\sqrt{T}).$$

Similarly, we can obtain that

$$\zeta = \frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu} \approx \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

So we have

$$\mathcal{O}\left(\frac{1-(1-\Gamma)^T}{\Gamma}\zeta\right) \approx \mathcal{O}(1).$$

□

E Convergence of proposed algorithm

E.1 Preliminary lemmas

We will use the following foundational lemma for our proof.

Lemma 2. *For $v_1,\ v_2 \in \mathbb{R}^d$, we have*

$$\|v_1+v_2\|^2 \leq (1+a)\|v_1\|^2 + (1+\frac{1}{a})\|v_2\|^2.$$

Lemma 3. *For L-smooth function f, and x, y in its domain, the following is true:*

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

Lemma 4. *For independent, mean 0 random variables $X_1, ..., X_n$, we have*

$$\mathbb{E}[\|X_1+...+X_n\|^2] = \mathbb{E}[\|X_1\|^2 + ... + \|X_n\|^2]. \tag{7}$$

Lemma 5. *Given vector $X_1, \cdots, X_m \in \mathbb{R}^d$, if we sample $\mathcal{S} \subset \{1, \cdots, m\}$ uniformly randomly such that $|\mathcal{S}| = s$, then it holds that*

$$\mathbb{E}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}} X_i\right\|^2 = \frac{m-s}{ms(m-1)}\sum_{i\in[M]}\|X_i\|^2 + \frac{s-1}{ms(m-1)}\left\|\sum_{i\in[M]} X_i\right\|^2$$

Proof. Let $\mathbb{I}\{i \in \mathcal{S}\}$ be the indicator for the event $i \in \mathcal{S}$, we prove this lemma by calculating as follows:

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{s}\sum_{i\in\mathcal{S}} X_i\right\|^2 &= \mathbb{E}\left\|\frac{1}{s}\sum_{i\in[M]} X_i\mathbb{I}\{i\in\mathcal{S}\}\right\|^2 \\
&= \frac{1}{s^2}\mathbb{E}\bigg(\sum_i \|X_i\|^2\mathbb{I}\{i\in\mathcal{S}\} + 2\sum_{i<j}\langle X_i, X_j\rangle\mathbb{I}\{i,j\in\mathcal{S}\}\bigg) \\
&= \frac{1}{sm}\sum_{i\in[M]}\|X_i\|^2 + \frac{1}{s^2}\frac{s(s-1)}{m(m-1)}2\sum_{i<j}\langle X_i, X_j\rangle \\
&= \frac{1}{sm}\sum_{i\in[M]}\|X_i\|^2 + \frac{1}{s^2}\frac{s(s-1)}{m(m-1)}\left(\left\|\sum_{i\in[M]} X_i\right\|^2 - \sum_{i\in[M]}\|X_i\|^2\right) \\
&= \frac{m-s}{sm(m-1)}\sum_{i=1}^{m}\|X_i\|^2 + \frac{s-1}{sm(m-1)}\left\|\sum_{i=1}^{m} X_i\right\|^2.
\end{aligned}$$

□

E.2 Convergence of FedSEMA

In this subsection, we present the proofs for the FedSEMA algorithm. We denote $\nabla\hat{f}(x)$ as $\nabla f(x) + \mu x$.

Lemma 6. *For $x^i_{t,k} \in R^d$, $\forall i \in \mathcal{S}_t$, we denote $\delta^i_{t,k} = x^i_{t,k} - x^i_{t,k-1}$ with setting $\delta^i_{t,0} = 0$, and $\Delta^i_{t,k} = \sum_{\tau=0}^{k}\delta^i_{t,\tau} = x^i_{t,k} - x^i_{t,0}$. Under the local update rule in Algorithm 1, we have:*

$$\Delta^i_{t,K} = -\eta_l\sum_{k=0}^{K-1}(1-\eta_l(\lambda+\mu))^{K-1-k}\left(\hat{g}^i_{t,k} + \mu x_t\right) - \beta\frac{1-(1-\eta_l(\lambda+\mu))^K}{\lambda+\mu}v_t.$$

Proof. According to the update rule, we have:

$$\begin{aligned}
\delta^i_{t,k} &= \Delta^i_{t,k} - \Delta^i_{t,k-1} \\
&= x^i_{t,k} - x^i_{t,k-1} \\
&= -\eta_l\left(\beta v_t + \hat{g}^i_{t,k-1} + (\lambda+\mu)x^i_{t,k-1} - \lambda x_t\right) \\
&= -\eta_l\left(\beta v_t + \hat{g}^i_{t,k-1} + \mu x_t\right) - \eta_l(\lambda+\mu)(x^i_{t,k-1} - x_t) \\
&= -\eta_l\left(\beta v_t + \hat{g}^i_{t,k-1} + \mu x_t\right) - \eta_l(\lambda+\mu)\Delta^i_{t,k-1}.
\end{aligned}$$

Then we can obtain

$$\begin{aligned}
\Delta^i_{t,k} &= (1-\eta_l(\lambda+\mu))\Delta^i_{t,k-1} - \eta_l\left(\beta v_t + \hat{g}^i_{t,k-1} + \mu x_t\right) \\
&= -\eta_l\sum_{\tau=0}^{k-1}(1-\eta_l(\lambda+\mu))^{k-1-\tau}\left(\beta v_t + \hat{g}^i_{t,\tau} + \mu x_t\right) \\
&= -\eta_l\sum_{\tau=0}^{k-1}(1-\eta_l(\lambda+\mu))^{k-1-\tau}\left(\hat{g}^i_{t,\tau} + \mu x_t\right) - \beta\frac{1-(1-\eta_l(\lambda+\mu))^k}{\lambda+\mu}v_t.
\end{aligned}$$

Let

$$\alpha_k = (1-\eta_l(\lambda+\mu))^{K-1-k}, \qquad \alpha = \frac{1-(1-\eta_l(\lambda+\mu))^K}{\eta_l(\lambda+\mu)} = \sum_{k=0}^{K-1}(1-\eta_l(\lambda+\mu))^{K-1-k},$$

so we have

$$\Delta^i_{t,K} = -\eta_l\sum_{k=0}^{K-1}\alpha_k\left(\hat{g}^i_{t,k} + \mu x_t\right) - \alpha\beta\eta_l v_t.$$

□

Lemma 7. *Under the update rule of Algorithm 1, we have*

$$v_{t+1} = \frac{\gamma\beta}{1+\beta} v_t + \frac{1}{(1+\beta)sK} \sum_{i \in \mathcal{S}_t} G_t^i,$$

where $\gamma = 1 + \frac{\alpha}{K}$ *and* $G_t^i = \sum_{k=0}^{K-1} \alpha_k \left(\hat{g}_{t,k}^i + \mu x_t \right)$.

Proof.

$$\begin{aligned}
v_{t+1} &= \frac{\beta}{1+\beta} v_t - \frac{\Delta x_t}{(1+\beta)\eta_l K} \\
&= \frac{\beta}{1+\beta} v_t - \frac{1}{(1+\beta)\eta_l sK} \sum_{i \in \mathcal{S}_t} \Delta x_t^i \\
&= \frac{\beta}{1+\beta} v_t - \frac{1}{(1+\beta)\eta_l sK} \sum_{i \in \mathcal{S}_t} \Delta_{t,K}^i \\
&= \frac{\beta}{1+\beta} v_t + \frac{1}{(1+\beta)sK} \sum_{i \in \mathcal{S}_t} \left(\sum_{k=0}^{K-1} \alpha_k \left(\hat{g}_{t,k}^i + \mu x_t \right) + \alpha\beta v_t \right) \\
&= \frac{\beta}{1+\beta} \left(1 + \frac{\alpha}{K} \right) v_t + \frac{1}{(1+\beta)sK} \sum_{i \in \mathcal{S}_t} \sum_{k=0}^{K-1} \alpha_k \left(\hat{g}_{t,k}^i + \mu x_t \right) \\
&= \frac{\gamma\beta}{1+\beta} v_t + \frac{1}{(1+\beta)sK} \sum_{i \in \mathcal{S}_t} G_t^i.
\end{aligned}$$

□

Consider the auxiliary sequence z_t, given by

$$z_t = \frac{1+\beta}{1-\alpha'\beta} x_t - \frac{\gamma\beta}{1-\alpha'\beta} x_{t-1} + \frac{\beta\eta\eta_l}{(1-\alpha'\beta)s} \sum_{i \in \mathcal{S}_{t-1}} G_{t-1}^i, \tag{8}$$

where $\alpha' = \frac{\alpha}{K}$. From the expression of z_t, it can be seen that, as t tends to positive infinity, z_t and x_t both converge to optimum x^*.

Lemma 8. *Consider the sequence* z_t*, we have*

$$z_{t+1} - z_t = -\frac{\eta\eta_l}{(1-\alpha'\beta)s} \sum_{i \in \mathcal{S}_t} G_t^i.$$

Proof.

$$\begin{aligned}
&z_{t+1} - z_t \\
&= \frac{1+\beta}{1-\alpha'\beta}(x_{t+1} - x_t) - \frac{\gamma\beta}{1-\alpha'\beta}(x_t - x_{t-1}) + \frac{\beta\eta\eta_l}{(1-\alpha'\beta)s}\Big(\sum_{i \in \mathcal{S}_t} G_t^i - \sum_{i \in \mathcal{S}_{t-1}} G_{t-1}^i\Big) \\
&= -\frac{(1+\beta)\eta\eta_l}{(1-\alpha'\beta)s} \sum_{i \in \mathcal{S}_t} (G_t^i + \alpha\beta v_t) + \frac{\gamma\beta\eta\eta_l}{(1-\alpha'\beta)s} \sum_{i \in \mathcal{S}_{t-1}} (G_{t-1}^i + \alpha\beta v_{t-1}) \\
&\quad + \frac{\beta\eta\eta_l}{(1-\alpha'\beta)s}\Big(\sum_{i \in \mathcal{S}_t} G_t^i - \sum_{i \in \mathcal{S}_{t-1}} G_{t-1}^i\Big) \\
&= -\frac{\eta\eta_l}{(1-\alpha'\beta)s} \sum_{i \in \mathcal{S}_t} G_t^i.
\end{aligned}$$

the last equality follow from Lemma 7. Besides,

$$\mathbb{E}_t \|z_{t+1} - z_t\|^2 = \mathbb{E}_t \left\| -\frac{\eta\eta_l}{(1-\alpha'\beta)s} \sum_{i \in \mathcal{S}_t} G_t^i \right\|^2$$

$$= \frac{\eta^2\eta_l^2}{(1-\alpha'\beta)^2 s^2}\mathbb{E}_t\left\|\sum_{i\in\mathcal{S}_t}\sum_{k=0}^{K-1}\alpha_k\left(\hat{g}_{t,k}^i+\mu x_t\right)\right\|^2$$
$$\leq \frac{\eta^2\eta_l^2}{(1-\alpha'\beta)^2 s^2}\mathbb{E}_t\left\|\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2+\frac{\alpha\eta^2\eta_l^2}{(1-\alpha'\beta)^2 s}\tilde{\sigma}_l^2.$$

where we denote $\bar{G}_t^i := \sum_{k=0}^{K-1}\alpha_k\left[\nabla F_i(x_{t,k}^i)+\beta_0\left(\nabla F_i(x_{t,k}^i)-\nabla F_i(x_{t,k-1}^i)\right)+\mu x_t\right]$, and $\tilde{\sigma}_l^2 := \left[(1+\beta_0)^2+\beta_0^2\right]\sigma_l^2$. The inequality is based on Assumption 2 and Lemma 4. □

Lemma 9. *Consider the z_t and x_t, then we have*

$$z_t - x_t = -\frac{\alpha(\beta+\beta^2)\eta\eta_l}{(1-\alpha'\beta)}v_t.$$

Proof.

$$\begin{aligned}z_t - x_t &= \frac{\gamma\beta}{1-\alpha'\beta}(x_t - x_{t-1}) + \frac{\beta\eta\eta_l}{(1-\alpha'\beta)s}\sum_{i\in\mathcal{S}_{t-1}}G_{t-1}^i\\ &= -\frac{\gamma\beta\eta\eta_l}{(1-\alpha'\beta)s}(\sum_{i\in\mathcal{S}_{t-1}}G_{t-1}^i+\alpha\beta s v_{t-1}) + \frac{\beta\eta\eta_l}{(1-\alpha'\beta)s}\sum_{i\in\mathcal{S}_{t-1}}G_{t-1}^i\\ &= -\frac{\alpha(\beta+\beta^2)\eta\eta_l}{(1-\alpha'\beta)}v_t.\end{aligned}$$

□

Lemma 10. *Consider the global momentum sequence v_t and summing over t from 0 to $T-1$, then we have*

$$\sum_{k=0}^{T-1}\mathbb{E}_t\|v_t\|^2 \leq \frac{1}{(1-\alpha'\beta)^2}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2+\frac{T\alpha^2}{(1-\alpha'\beta)^2 sK^2}\tilde{\sigma}_l^2.$$

Proof.

$$\begin{aligned}&\mathbb{E}_t\|v_t\|^2\\ &= \mathbb{E}_t\left\|\frac{\gamma\beta}{1+\beta}v_{t-1}+\frac{1}{(1+\beta)sK}\sum_{i\in\mathcal{S}_{t-1}}G_{t-1}^i\right\|^2\\ &= \mathbb{E}_t\left\|\frac{1}{(1+\beta)sK}\sum_{\tau=0}^{t-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\sum_{i\in\mathcal{S}_\tau}G_\tau^i\right\|^2\\ &\leq \mathbb{E}_t\left\|\frac{1}{(1+\beta)sK}\sum_{\tau=0}^{t-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2+\frac{\alpha^2[1-(\frac{\gamma\beta}{1+\beta})^t]^2}{(1-\alpha'\beta)^2 sK^2}\tilde{\sigma}_l^2\\ &\overset{(a)}{=} p_t^2\mathbb{E}_t\left\|\frac{1}{(1+\beta)sK}\sum_{\tau=0}^{t-1}p_t^{-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2+\frac{\alpha^2[1-(\frac{\gamma\beta}{1+\beta})^t]^2}{(1-\alpha'\beta)^2 sK^2}\tilde{\sigma}_l^2\\ &\leq \frac{p_t^2}{(1+\beta)^2}\sum_{\tau=0}^{t-1}p_t^{-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2+\frac{\alpha^2[1-(\frac{\gamma\beta}{1+\beta})^t]^2}{(1-\alpha'\beta)^2 sK^2}\tilde{\sigma}_l^2 \qquad (9)\\ &= \frac{1-(\frac{\gamma\beta}{1+\beta})^t}{(1+\beta)(1-\alpha'\beta)}\sum_{\tau=0}^{t-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2+\frac{\alpha^2[1-(\frac{\gamma\beta}{1+\beta})^t]^2}{(1-\alpha'\beta)^2 sK^2}\tilde{\sigma}_l^2\end{aligned}$$

$$\leq \frac{1}{(1+\beta)(1-\alpha'\beta)}\sum_{\tau=0}^{t-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2+\frac{\alpha^2}{(1-\alpha'\beta)^2sK^2}\tilde{\sigma}_l^2.$$

The first inequality is based on Assumption 2 and Lemma 4. In (a), $p_t=\sum_{\tau=0}^{t-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}$. Now we sum (9) over t from 0 to $T-1$:

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}_t\|v_t\|^2\\
&\leq \frac{1}{(1+\beta)(1-\alpha'\beta)}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}(\frac{\gamma\beta}{1+\beta})^{t-1-\tau}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2+\frac{T\alpha^2}{(1-\alpha'\beta)^2sK^2}\tilde{\sigma}_l^2\\
&= \frac{1}{(1+\beta)(1-\alpha'\beta)}\sum_{\tau=0}^{T-2}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_\tau}\bar{G}_\tau^i\right\|^2\sum_{s=\tau+1}^{T-1}(\frac{\gamma\beta}{1+\beta})^{s-1-\tau}+\frac{T\alpha^2}{(1-\alpha'\beta)^2sK^2}\tilde{\sigma}_l^2\\
&\leq \frac{1}{(1-\alpha'\beta)^2}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2+\frac{T\alpha^2}{(1-\alpha'\beta)^2sK^2}\tilde{\sigma}_l^2.
\end{aligned}$$

□

Lemma 11. *Suppose the local learning rate satisfies $\eta_l\leq\frac{1}{KL}$, we can bound the client drift by:*

$$\frac{1}{m}\sum_{i\in[M]}\mathbb{E}_t\|x_{t,k}^i-x_t\|^2\leq 42\eta_l^2K^2\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+42\eta_l^2K^2\sigma_g^2+3\eta_l^2K\tilde{\sigma}_l^2+24\beta^2\eta_l^2K^2\mathbb{E}_t\|v_t\|^2.$$

Proof. For simplicity, we denote

$$\mathcal{E}_{t,k}^i:=\mathbb{E}_t\|x_{t,k}^i-x_t\|^2,\qquad \bar{\mathcal{E}}_{t,k}^i:=\max_{0\leq r\leq k}\mathcal{E}_{t,r}^i.$$

Recall that the corrected stochastic gradient is denoted by

$$\hat{g}_{t,k}^i=g_{t,k}^i+\beta_0(g_{t,k}^i-g_{t,k-1}^i),$$

where

$$g_{t,k}^i=\nabla F_i(x_{t,k}^i,\xi_{t,k}^i).$$

For notational convenience, we set $g_{t,-1}^i=g_{t,0}^i$.

According to the local update rule, we have

$$\begin{aligned}
&\mathbb{E}_t\|x_{t,k}^i-x_t\|^2\\
&=\mathbb{E}_t\left\|x_{t,k-1}^i-x_t-\eta_l\left(\beta v_t+\hat{g}_{t,k-1}^i+(\lambda+\mu)x_{t,k-1}^i-\lambda x_t\right)\right\|^2\\
&=\mathbb{E}_t\left\|\left(1-\eta_l(\lambda+\mu)\right)\left(x_{t,k-1}^i-x_t\right)-\eta_l\left(\hat{g}_{t,k-1}^i+\mu x_t+\beta v_t\right)\right\|^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_t\|x_{t,k}^i-x_t\|^2\\
&\leq\mathbb{E}_t\Big\|\left(1-\eta_l(\lambda+\mu)\right)\left(x_{t,k-1}^i-x_t\right)-\eta_l\Big[\nabla F_i(x_{t,k-1}^i)+\beta_0\left(\nabla F_i(x_{t,k-1}^i)-\nabla F_i(x_{t,k-2}^i)\right)\\
&\qquad+\mu x_t+\beta v_t\Big]\Big\|^2+\eta_l^2\tilde{\sigma}_l^2\\
&\leq(1+a)\left(1-\eta_l(\lambda+\mu)\right)^2\mathbb{E}_t\|x_{t,k-1}^i-x_t\|^2+\eta_l^2\tilde{\sigma}_l^2\\
&\quad+\left(1+\frac{1}{a}\right)\eta_l^2\mathbb{E}_t\left\|\nabla F_i(x_{t,k-1}^i)+\beta_0\left(\nabla F_i(x_{t,k-1}^i)-\nabla F_i(x_{t,k-2}^i)\right)+\mu x_t+\beta v_t\right\|^2,
\end{aligned}$$

where the last inequality follows from Lemma 2. We further bound the third term as

$$\mathbb{E}_t\left\|\nabla F_i(x_{t,k-1}^i)+\mu x_t+\beta v_t+\beta_0\left(\nabla F_i(x_{t,k-1}^i)-\nabla F_i(x_{t,k-2}^i)\right)\right\|^2$$

$$
\begin{aligned}
&\leq (1+\beta_0)\mathbb{E}_t\|\nabla F_i(x^i_{t,k-1}) + \mu x_t + \beta v_t\|^2 + \beta_0(1+\beta_0)\mathbb{E}_t\|\nabla F_i(x^i_{t,k-1}) - \nabla F_i(x^i_{t,k-2})\|^2 \\
&\leq (1+\beta_0)\mathbb{E}_t\left\|\nabla F_i(x^i_{t,k-1}) - \nabla F_i(x_t) + \nabla F_i(x_t) + \mu x_t + \beta v_t\right\|^2 \\
&\quad + (\beta_0+\beta_0^2)L^2\mathbb{E}_t\|x^i_{t,k-1} - x^i_{t,k-2}\|^2 \\
&\leq (1+\beta_0)\mathbb{E}_t\Big(3\|\nabla F_i(x^i_{t,k-1}) - \nabla F_i(x_t)\|^2 + 3\|\nabla F_i(x_t) + \mu x_t\|^2 + 3\beta^2\|v_t\|^2\Big) \\
&\quad + 2(\beta_0+\beta_0^2)L^2\mathbb{E}_t\|x^i_{t,k-1} - x_t\|^2 + 2(\beta_0+\beta_0^2)L^2\mathbb{E}_t\|x^i_{t,k-2} - x_t\|^2 \\
&\leq \Big(3(1+\beta_0) + 2(\beta_0+\beta_0^2)\Big)L^2\mathcal{E}^i_{t,k-1} + 2(\beta_0+\beta_0^2)L^2\mathcal{E}^i_{t,k-2} \\
&\quad + 6(1+\beta_0)\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2 + 6(1+\beta_0)\sigma_g^2 + 3\beta^2(1+\beta_0)\mathbb{E}_t\|v_t\|^2.
\end{aligned}
$$

Since $0 < \beta_0 < 0.1$, we have

$$
3(1+\beta_0) + 4(\beta_0+\beta_0^2) \leq 4, \qquad 6(1+\beta_0) \leq 7, \qquad 3(1+\beta_0) \leq 4.
$$

Therefore,

$$
\begin{aligned}
\mathcal{E}^i_{t,k} \leq &\left[(1-\eta_l(\lambda+\mu))^2(1+a) + 4\left(1+\frac{1}{a}\right)\eta_l^2L^2\right]\bar{\mathcal{E}}^i_{t,k-1} + \left(1+\frac{1}{a}\right)7\eta_l^2\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2 \\
&+ \left(1+\frac{1}{a}\right)7\eta_l^2\sigma_g^2 + \eta_l^2\tilde{\sigma}_l^2 + \left(1+\frac{1}{a}\right)4\beta^2\eta_l^2\mathbb{E}_t\|v_t\|^2.
\end{aligned}
$$

For $K = 1$, take $a = 1$ and the lemma holds. Suppose that $K \geq 2$ thereafter and take $a = \frac{1}{2K-1}$. It follows from $\eta_l \leq \frac{1}{15LK}$ and $0 \leq \eta_l(\lambda+\mu) \leq 1$ that

$$
\bar{\mathcal{E}}^i_{t,k} \leq \left(1+\frac{1}{K-1}\right)\bar{\mathcal{E}}^i_{t,k-1} + 14\eta_l^2K\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2 + 14\eta_l^2K\sigma_g^2 + 8\beta^2\eta_l^2K\mathbb{E}_t\|v_t\|^2 + \eta_l^2\tilde{\sigma}_l^2.
$$

Unrolling the recursion, noting that $\bar{\mathcal{E}}^i_{t,0} = 0$ and

$$
\sum_{r=0}^{K-1}\left(1+\frac{1}{K-1}\right)^r \leq 3K
$$

for $K \geq 2$, we obtain

$$
\bar{\mathcal{E}}^i_{t,k} \leq 42\eta_l^2K^2\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2 + 42\eta_l^2K^2\sigma_g^2 + 24\beta^2\eta_l^2K^2\mathbb{E}_t\|v_t\|^2 + 3\eta_l^2K\tilde{\sigma}_l^2.
$$

Since $\mathcal{E}^i_{t,k} \leq bar\mathcal{E}^i_{t,k}$, averaging over all clients gives

$$
\begin{aligned}
\frac{1}{m}\sum_{i\in[M]}\mathbb{E}_t\|x^i_{t,k} - x_t\|^2 \leq &\, 42\eta_l^2K^2\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2 + 42\eta_l^2K^2\sigma_g^2 \\
&+ 3\eta_l^2K\tilde{\sigma}_l^2 + 24\beta^2\eta_l^2K^2\mathbb{E}_t\|v_t\|^2.
\end{aligned}
$$

□

Lemma 12. *Suppose that the local learning rate satisfies $\eta_l \leq \frac{1}{15KL}$. We can bound the sum of gradients by:*

$$
\begin{aligned}
\frac{\kappa_1}{mK}\sum_{t=0}^{T-1}\sum_{i\in[M]}\mathbb{E}_t\left\|\sum_{k=0}^{K-1}\alpha_k\nabla\bar{F}_i(x^i_{t,k})\right\|^2 \leq &\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K + \frac{4(1+\beta_0)\alpha^2}{K}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2 \\
+ \left(84C_{\beta_0}\alpha^2\eta_l^2L^2K + \frac{4(1+\beta_0)\alpha^2}{K}\right)&T\sigma_g^2 + 6C_{\beta_0}\alpha^2\eta_l^2L^2T\tilde{\sigma}_l^2 + T\frac{48C_{\beta_0}\alpha^4\beta^2\eta_l^2L^2}{(1-\alpha'\beta)^2sK}\tilde{\sigma}_l^2.
\end{aligned}
$$

Proof. For notational simplicity, we denote

$$
\nabla\bar{F}_i(x^i_{t,k}) := \nabla F_i(x^i_{t,k}) + \beta_0\left(\nabla F_i(x^i_{t,k}) - \nabla F_i(x^i_{t,k-1})\right) + \mu x_t.
$$

Then we have

$$\frac{1}{mK}\sum_{i\in[M]}\mathbb{E}_t\left\|\sum_{k=0}^{K-1}\alpha_k\nabla\bar{F}_i(x_{t,k}^i)\right\|^2=\frac{\alpha^2}{mK}\sum_{i\in[M]}\mathbb{E}_t\left\|\sum_{k=0}^{K-1}\frac{\alpha_k}{\alpha}\nabla\bar{F}_i(x_{t,k}^i)\right\|^2$$
$$\leq\frac{\alpha}{mK}\sum_{i\in[M]}\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\left\|\nabla\bar{F}_i(x_{t,k}^i)\right\|^2.$$

For all $i\in[M]$, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|\nabla\bar{F}_i(x_{t,k}^i)\|^2\\
&=\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\left\|\nabla F_i(x_{t,k}^i)+\beta_0\left(\nabla F_i(x_{t,k}^i)-\nabla F_i(x_{t,k-1}^i)\right)+\mu x_t\right\|^2\\
&\leq(1+\beta_0)\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|\nabla F_i(x_{t,k}^i)+\mu x_t\|^2+\beta_0(1+\beta_0)\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|\nabla F_i(x_{t,k}^i)-\nabla F_i(x_{t,k-1}^i)\|^2\\
&\leq(1+\beta_0)\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\left\|\nabla F_i(x_{t,k}^i)-\nabla F_i(x_t)+\nabla F_i(x_t)+\mu x_t\right\|^2\\
&\quad+\beta_0(1+\beta_0)L^2\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|x_{t,k}^i-x_{t,k-1}^i\|^2\\
&\leq 2(1+\beta_0)L^2\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|x_{t,k}^i-x_t\|^2+4(1+\beta_0)\alpha\sigma_g^2+4(1+\beta_0)\alpha\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\\
&\quad+2\beta_0(1+\beta_0)L^2\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|x_{t,k}^i-x_t\|^2+2\beta_0(1+\beta_0)L^2\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|x_{t,k-1}^i-x_t\|^2.
\end{aligned}$$

Define

$$C_{\beta_0}:=(1+\beta_0)(1+2\beta_0).$$

Using the client drift bound, we have

$$\begin{aligned}
&\sum_{k=0}^{K-1}\alpha_k\mathbb{E}_t\|\nabla\bar{F}_i(x_{t,k}^i)\|^2\\
&\leq 2C_{\beta_0}\alpha L^2\Big(42\eta_l^2K^2\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+42\eta_l^2K^2\sigma_g^2+3\eta_l^2K\tilde{\sigma}_l^2+24\beta^2\eta_l^2K^2\mathbb{E}_t\|v_t\|^2\Big)\\
&\quad+4(1+\beta_0)\alpha\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+4(1+\beta_0)\alpha\sigma_g^2\\
&\leq\left(84C_{\beta_0}\alpha\eta_l^2L^2K^2+4(1+\beta_0)\alpha\right)\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+\left(84C_{\beta_0}\alpha\eta_l^2L^2K^2+4(1+\beta_0)\alpha\right)\sigma_g^2\\
&\quad+6C_{\beta_0}\alpha\eta_l^2L^2K\tilde{\sigma}_l^2+48C_{\beta_0}\alpha\beta^2\eta_l^2L^2K^2\mathbb{E}_t\|v_t\|^2.
\end{aligned}$$

Substituting it back into the above inequality, we get

$$\begin{aligned}
&\frac{1}{mK}\sum_{i\in[M]}\mathbb{E}_t\left\|\sum_{k=0}^{K-1}\alpha_k\nabla\bar{F}_i(x_{t,k}^i)\right\|^2\\
&\leq\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\\
&\quad+\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)\sigma_g^2+6C_{\beta_0}\alpha^2\eta_l^2L^2\tilde{\sigma}_l^2+48C_{\beta_0}\alpha^2\beta^2\eta_l^2L^2K\mathbb{E}_t\|v_t\|^2.
\end{aligned}$$

We sum the above inequality by using weight 1, and get

$$
\begin{aligned}
&\frac{1}{mK}\sum_{t=0}^{T-1}\sum_{i\in[M]}\mathbb{E}_t\left\|\sum_{k=0}^{K-1}\alpha_k\nabla\bar{F}_i(x_{t,k}^i)\right\|^2\\
&\leq\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\\
&\quad+\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)T\sigma_g^2+6C_{\beta_0}\alpha^2\eta_l^2L^2T\tilde{\sigma}_l^2\\
&\quad+48C_{\beta_0}\alpha^2\beta^2\eta_l^2L^2K\sum_{t=0}^{T-1}\mathbb{E}_t\|v_t\|^2\\
&\leq\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\\
&\quad+\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)T\sigma_g^2+6C_{\beta_0}\alpha^2\eta_l^2L^2T\tilde{\sigma}_l^2\\
&\quad+\frac{48C_{\beta_0}\alpha^2\beta^2\eta_l^2L^2K}{(1-\alpha'\beta)^2}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2+T\frac{48C_{\beta_0}\alpha^4\beta^2\eta_l^2L^2}{(1-\alpha'\beta)^2sK}\tilde{\sigma}_l^2.
\end{aligned}
$$

We define a positive constant $\kappa_1<1$ such that $0<\kappa_1\leq 1-\frac{48C_{\beta_0}\alpha^2\beta^2\eta_l^2L^2}{(1-\alpha'\beta)^2}$, Therefore, we get

$$
\begin{aligned}
&\frac{\kappa_1}{mK}\sum_{t=0}^{T-1}\sum_{i\in[M]}\mathbb{E}_t\left\|\sum_{k=0}^{K-1}\alpha_k\nabla\bar{F}_i(x_{t,k}^i)\right\|^2\\
&\leq\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\\
&\quad+\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{4(1+\beta_0)\alpha^2}{K}\right)T\sigma_g^2+6C_{\beta_0}\alpha^2\eta_l^2L^2T\tilde{\sigma}_l^2\\
&\quad+T\frac{48C_{\beta_0}\alpha^4\beta^2\eta_l^2L^2}{(1-\alpha'\beta)^2sK}\tilde{\sigma}_l^2.
\end{aligned}
$$

□

Lemma 13. *Under Assumption 1-2 and the loss function is non-convex, we define* $D_0:=\mathbb{E}(\hat{f}(z_0)-\hat{f}(z_T))$. *When the learning rates satisfy* $\eta_l\leq\frac{1}{16KL}$ *and*

$$
\eta\eta_l\leq\frac{m(s-1)}{s(m-1)}\min\left\{\frac{(1-\alpha'\beta)^2\sqrt{K}}{4\alpha\beta L\sqrt{3\alpha}},\frac{(1-\alpha'\beta)\eta}{6\beta L\sqrt{2\alpha}},\frac{1-\alpha'\beta}{2\alpha L}\right\}.
$$

Then the auxiliary sequence z_t *in equation* (8) *generated by executing FedSEMA satisfies:*

$$
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\leq\mathcal{O}\bigg(\frac{LD_0}{\sqrt{sKT}}+\frac{\sigma_l^2}{\sqrt{sKT}}+\frac{1}{T}\bigg(\frac{\sigma_l^2}{K^2}+K\sigma_l^2+\frac{\sigma_l^2}{sK}+\frac{\sigma_g^2}{K}\bigg)+\Psi_l+\Psi_g\bigg),
$$

where

$$
\begin{aligned}
\Psi_l&=\left(\frac{1}{T}+\frac{K}{sT}\right)\frac{(m-s)}{m-1}\left(\frac{1}{T}+\frac{1}{sK^2T}+\frac{1}{\sqrt{sKT}}\right)\sigma_l^2,\\
\Psi_g&=\left(\frac{K}{T}+K\right)\frac{(m-s)}{m-1}\left(\frac{1}{T}+\frac{1}{sK^2T}+\frac{1}{\sqrt{sKT}}\right)\sigma_g^2.
\end{aligned}
$$

Proof. For the general non-convex case, according to the Assumptions and the smoothness of $\hat{f}$, we take the conditional expectation at round $t+1$ and expand the $\hat{f}(z_{t+1})$ as

$$\mathbb{E}_t[\hat{f}(z_{t+1})] \leq \hat{f}(z_t) + \langle \nabla \hat{f}(z_t), \mathbb{E}_t[z_{t+1} - z_t] \rangle + \frac{L}{2}\mathbb{E}_t\|z_{t+1} - z_t\|^2.$$

Note that the second term can be split into the following:

$$\begin{aligned}
&\langle \nabla \hat{f}(z_t), \mathbb{E}_t[z_{t+1} - z_t] \rangle \\
&= \langle \nabla \hat{f}(z_t) - \nabla \hat{f}(x_t) + \nabla \hat{f}(x_t), \mathbb{E}_t[z_{t+1} - z_t] \rangle \\
&= \underbrace{\langle \nabla \hat{f}(z_t) - \nabla \hat{f}(x_t), \mathbb{E}_t[z_{t+1} - z_t] \rangle}_{A_1} + \underbrace{\langle \nabla \hat{f}(x_t), \mathbb{E}_t[z_{t+1} - z_t] \rangle}_{A_2}.
\end{aligned}$$

We first bound the A_1:

$$\begin{aligned}
A_1 &= \left\langle \nabla \hat{f}(z_t) - \nabla \hat{f}(x_t), -\frac{\eta\eta_l}{1-\alpha'\beta}\mathbb{E}_t\left[\frac{1}{m}\sum_{i\in[M]} G_t^i\right] \right\rangle \\
&= \left\langle \nabla \hat{f}(z_t) - \nabla \hat{f}(x_t), -\frac{\eta\eta_l}{1-\alpha'\beta}\frac{1}{m}\sum_{i\in[M]} \bar{G}_t^i \right\rangle \\
&= \left\langle \frac{\sqrt{2\eta\eta_l K}}{\sqrt{1-\alpha'\beta}}\left(\nabla \hat{f}(z_t) - \nabla \hat{f}(x_t)\right), -\frac{\sqrt{\eta\eta_l}}{\sqrt{2(1-\alpha'\beta)K}}\frac{1}{m}\sum_{i\in[M]} \bar{G}_t^i \right\rangle \\
&\leq \frac{\eta\eta_l K}{1-\alpha'\beta}\|\nabla \hat{f}(z_t) - \nabla \hat{f}(x_t)\|^2 + \frac{\eta\eta_l}{4(1-\alpha'\beta)m^2 K}\mathbb{E}_t\left\|\sum_{i\in[M]} \bar{G}_t^i\right\|^2 \\
&\leq \frac{\eta\eta_l L^2 K}{1-\alpha'\beta}\|z_t - x_t\|^2 + \frac{\eta\eta_l}{4(1-\alpha'\beta)m^2 K}\mathbb{E}_t\left\|\sum_{i\in[M]} \bar{G}_t^i\right\|^2.
\end{aligned}$$

Then the A_2 can be bounded as:

$$\begin{aligned}
A_2 &= \left\langle \nabla \hat{f}(x_t), -\frac{\eta\eta_l}{1-\alpha'\beta}\frac{1}{m}\sum_{i\in[M]} \bar{G}_t^i \right\rangle \\
&= \left\langle \nabla \hat{f}(x_t), -\frac{\eta\eta_l}{1-\alpha'\beta}\left(\frac{1}{m}\sum_{i\in[M]} \bar{G}_t^i - \alpha\nabla \hat{f}(x_t) + \alpha\nabla \hat{f}(x_t)\right) \right\rangle \\
&= -\frac{\eta\eta_l\alpha}{1-\alpha'\beta}\|\nabla \hat{f}(x_t)\|^2 - \left\langle \frac{\sqrt{\eta\eta_l\alpha}}{\sqrt{1-\alpha'\beta}}\nabla \hat{f}(x_t), \frac{\sqrt{\eta\eta_l}}{\sqrt{(1-\alpha'\beta)\alpha}}\left(\frac{1}{m}\sum_{i\in[M]} \bar{G}_t^i - \alpha\nabla \hat{f}(x_t)\right) \right\rangle \\
&\leq -\frac{\eta\eta_l\alpha}{2(1-\alpha'\beta)}\|\nabla \hat{f}(x_t)\|^2 + \frac{\eta\eta_l}{2(1-\alpha'\beta)\alpha}\mathbb{E}_t\left\|\frac{1}{m}\sum_{i\in[M]} \bar{G}_t^i - \alpha\nabla \hat{f}(x_t)\right\|^2 \\
&\quad - \frac{\eta\eta_l}{2(1-\alpha'\beta)\alpha m^2}\mathbb{E}_t\left\|\sum_{i\in[M]} \bar{G}_t^i\right\|^2.
\end{aligned}$$

By the definition of the corrected deterministic accumulated gradient, we have

$$\sum_{i\in[M]} \bar{G}_t^i - m\alpha\nabla \hat{f}(x_t) = \sum_{i\in[M]}\sum_{k=0}^{K-1} \alpha_k\left[\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)\right]$$

$$+ \beta_0 \sum_{i\in[M]} \sum_{k=0}^{K-1} \alpha_k \left[\nabla F_i(x_{t,k}^i) - \nabla F_i(x_{t,k-1}^i)\right].$$

Using Lemma 2 and L-smoothness, we obtain

$$\begin{aligned}
&\mathbb{E}_t \left\| \sum_{i\in[M]} \bar{G}_t^i - m\alpha \nabla \hat{f}(x_t) \right\|^2 \\
&\leq (1+\beta_0)\mathbb{E}_t \left\| \sum_{i\in[M]} \sum_{k=0}^{K-1} \alpha_k \left[\nabla F_i(x_{t,k}^i) - \nabla F_i(x_t)\right] \right\|^2 \\
&\quad + \beta_0(1+\beta_0)\mathbb{E}_t \left\| \sum_{i\in[M]} \sum_{k=0}^{K-1} \alpha_k \left[\nabla F_i(x_{t,k}^i) - \nabla F_i(x_{t,k-1}^i)\right] \right\|^2 \\
&\leq m\alpha(1+\beta_0)L^2 \sum_{i\in[M]} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}_t \|x_{t,k}^i - x_t\|^2 \\
&\quad + 2m\alpha\beta_0(1+\beta_0)L^2 \sum_{i\in[M]} \sum_{k=0}^{K-1} \alpha_k \left(\mathbb{E}_t \|x_{t,k}^i - x_t\|^2 + \mathbb{E}_t \|x_{t,k-1}^i - x_t\|^2\right).
\end{aligned}$$

Since $0 < \beta_0 < 0.1$, we have

$$(1+\beta_0)(1+4\beta_0) \leq 2.$$

Therefore, the above term can be bounded by

$$\mathbb{E}_t \left\| \sum_{i\in[M]} \bar{G}_t^i - m\alpha \nabla \hat{f}(x_t) \right\|^2 \leq 2m\alpha L^2 \sum_{i\in[M]} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}_t \|x_{t,k}^i - x_t\|^2.$$

Then, by the client drift bound, we have

$$\begin{aligned}
A_2 &\leq \frac{\eta\eta_l L^2}{1-\alpha'\beta}\Big(42\eta_l^2 K^2 \mathbb{E}_t \|\nabla \hat{f}(x_t)\|^2 + 42\eta_l^2 K^2 \sigma_g^2 + 3\eta_l^2 K \tilde{\sigma}_l^2 + 24\beta^2 \eta_l^2 K^2 \mathbb{E}_t \|v_t\|^2\Big) \\
&\quad - \frac{\eta\eta_l \alpha}{2(1-\alpha'\beta)} \|\nabla \hat{f}(x_t)\|^2 - \frac{\eta\eta_l}{2(1-\alpha'\beta)\alpha m^2} \mathbb{E}_t \left\| \sum_{i\in[M]} \bar{G}_t^i \right\|^2.
\end{aligned}$$

Now we have

$$\begin{aligned}
&\mathbb{E}_t[\hat{f}(z_{t+1})] \\
&\leq \hat{f}(z_t) + \langle \nabla \hat{f}(z_t), \mathbb{E}_t[z_{t+1} - z_t] \rangle + \frac{L}{2} \mathbb{E}_t \|z_{t+1} - z_t\|^2 \\
&\leq \hat{f}(z_t) + \frac{\eta\eta_l L^2 K}{1-\alpha'\beta} \|z_t - x_t\|^2 + \frac{\eta\eta_l}{4(1-\alpha'\beta)m^2 K} \mathbb{E}_t \left\| \sum_{i\in[M]} \bar{G}_t^i \right\|^2 - \frac{\eta\eta_l \alpha}{2(1-\alpha'\beta)} \|\nabla \hat{f}(x_t)\|^2 \\
&\quad + \frac{\eta\eta_l L^2}{1-\alpha'\beta}\Big(42\eta_l^2 K^2 \mathbb{E}_t \|\nabla \hat{f}(x_t)\|^2 + 42\eta_l^2 K^2 \sigma_g^2 + 3\eta_l^2 K \tilde{\sigma}_l^2 + 24\beta^2 \eta_l^2 K^2 \mathbb{E}_t \|v_t\|^2\Big) \\
&\quad - \frac{\eta\eta_l}{2(1-\alpha'\beta)\alpha m^2} \mathbb{E}_t \left\| \sum_{i\in[M]} \bar{G}_t^i \right\|^2 + \frac{L}{2} \mathbb{E}_t \|z_{t+1} - z_t\|^2 \\
&\leq \hat{f}(z_t) + \left(\frac{\alpha^2\beta^2(1+\beta)^2\eta^3\eta_l^3 L^2 K}{(1-\alpha'\beta)^3} + \frac{24\beta^2\eta\eta_l^3 L^2 K^2}{1-\alpha'\beta}\right) \mathbb{E}_t \|v_t\|^2 - \frac{\eta\eta_l \alpha}{2(1-\alpha'\beta)} \mathbb{E}_t \|\nabla \hat{f}(x_t)\|^2
\end{aligned}$$

$$+\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+\left(\frac{3\eta\eta_l^3L^2K}{1-\alpha'\beta}+\frac{\alpha\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s}\right)\tilde{\sigma}_l^2+\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\sigma_g^2$$
$$+\frac{\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s^2}\mathbb{E}_t\left\|\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2-\frac{\eta\eta_l}{4(1-\alpha'\beta)\alpha m^2}\mathbb{E}_t\left\|\sum_{i\in[M]}\bar{G}_t^i\right\|^2.$$

Summing from $t=0,\cdots,T-1$, we get

$$\begin{aligned}
&\mathbb{E}[\hat{f}(z_T)]-\hat{f}(z_0)\\
\leq&-\left(\frac{\eta\eta_l\alpha}{2(1-\alpha'\beta)}-\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+T\left(\frac{3\eta\eta_l^3L^2K}{1-\alpha'\beta}+\frac{\alpha\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s}\right)\tilde{\sigma}_l^2\\
&+\left(\frac{\alpha^2\beta^2(1+\beta)^2\eta^3\eta_l^3L^2K}{(1-\alpha'\beta)^3}+\frac{24\beta^2\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|v_t\|^2+T\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\sigma_g^2\\
&+\frac{\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s^2}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2-\frac{\eta\eta_l}{4(1-\alpha'\beta)\alpha m^2}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\sum_{i\in[M]}\bar{G}_t^i\right\|^2\\
\leq&-\left(\frac{\eta\eta_l\alpha}{2(1-\alpha'\beta)}-\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+T\left(\frac{3\eta\eta_l^3L^2K}{1-\alpha'\beta}+\frac{\alpha\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s}\right)\tilde{\sigma}_l^2\\
&+T\left(\frac{4\alpha^4\beta^2\eta^3\eta_l^3L^2}{(1-\alpha'\beta)^5sK}+\frac{24\alpha^2\beta^2\eta\eta_l^3L^2}{(1-\alpha'\beta)^3s}\right)\tilde{\sigma}_l^2+T\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\sigma_g^2\\
&+\left(\frac{4\alpha^2\beta^2\eta^3\eta_l^3L^2K}{(1-\alpha'\beta)^5}+\frac{24\beta^2\eta\eta_l^3L^2K^2}{(1-\alpha'\beta)^3}+\frac{\eta^2\eta_l^2LK^2}{2(1-\alpha'\beta)^2}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\frac{1}{sK}\sum_{i\in\mathcal{S}_t}\bar{G}_t^i\right\|^2\\
&-\frac{\eta\eta_l}{4(1-\alpha'\beta)\alpha m^2}\sum_{t=0}^{T-1}\mathbb{E}_t\left\|\sum_{i\in[M]}\bar{G}_t^i\right\|^2.
\end{aligned}$$

Then we use Lemma 5 and let $\eta\eta_l$ satisfy $\eta\eta_l\leq\frac{m(s-1)}{s(m-1)}\min\left\{\frac{(1-\alpha'\beta)^2\sqrt{K}}{4\alpha\beta L\sqrt{3\alpha}},\frac{(1-\alpha'\beta)\eta}{6\beta L\sqrt{2\alpha}},\frac{1-\alpha'\beta}{2\alpha L}\right\}$, we have

$$\begin{aligned}
&\mathbb{E}[\hat{f}(z_T)]-\hat{f}(z_0)\\
\leq&-\left(\frac{\eta\eta_l\alpha}{2(1-\alpha'\beta)}-\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\right)\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2+T\left(\frac{3\eta\eta_l^3L^2K}{1-\alpha'\beta}+\frac{\alpha\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s}\right)\tilde{\sigma}_l^2\\
&+T\left(\frac{4\alpha^4\beta^2\eta^3\eta_l^3L^2}{(1-\alpha'\beta)^5sK}+\frac{24\alpha^2\beta^2\eta\eta_l^3L^2}{(1-\alpha'\beta)^3s}\right)\tilde{\sigma}_l^2+T\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\sigma_g^2\\
&+\frac{(m-s)}{ms(m-1)}\left(\frac{4\alpha^2\beta^2\eta^3\eta_l^3L^2}{(1-\alpha'\beta)^5K}+\frac{24\beta^2\eta\eta_l^3L^2}{(1-\alpha'\beta)^3}+\frac{\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2}\right)\sum_{t=0}^{T-1}\sum_{i\in[M]}\mathbb{E}_t\left\|\bar{G}_t^i\right\|^2\\
\leq&T\bigg(\frac{3\eta\eta_l^3L^2K}{1-\alpha'\beta}+\frac{\alpha\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2s}+\frac{4\alpha^4\beta^2\eta^3\eta_l^3L^2}{(1-\alpha'\beta)^5sK}+\frac{24\alpha^2\beta^2\eta\eta_l^3L^2}{(1-\alpha'\beta)^3s}\\
&+\left(6C_{\beta_0}\alpha^2\eta_l^2L^2K+\frac{48C_{\beta_0}\alpha^4\beta^2\eta_l^2L^2}{(1-\alpha'\beta)^2s}\right)\frac{(m-s)}{\kappa_1s(m-1)}\left(\frac{4\alpha^2\beta^2\eta^3\eta_l^3L^2}{(1-\alpha'\beta)^5K}+\frac{24\beta^2\eta\eta_l^3L^2}{(1-\alpha'\beta)^3}\right.\\
&\left.+\frac{\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2}\right)\bigg)\tilde{\sigma}_l^2+T\bigg(\frac{\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K^2+4(1+\beta_0)\alpha^2\right)(m-s)}{\kappa_1s(m-1)}\\
&\times\left(\frac{4\alpha^2\beta^2\eta^3\eta_l^3L^2}{(1-\alpha'\beta)^5K}+\frac{24\beta^2\eta\eta_l^3L^2}{(1-\alpha'\beta)^3}+\frac{\eta^2\eta_l^2L}{2(1-\alpha'\beta)^2}\right)+\frac{42\eta\eta_l^3L^2K^2}{1-\alpha'\beta}\bigg)\sigma_g^2\\
&-\sum_{t=0}^{T-1}\mathbb{E}_t\|\nabla\hat{f}(x_t)\|^2\bigg(\frac{\eta\eta_l\alpha}{2(1-\alpha'\beta)}-\frac{\left(84C_{\beta_0}\alpha^2\eta_l^2L^2K^2+4(1+\beta_0)\alpha^2\right)(m-s)}{\kappa_1s(m-1)}
\end{aligned}$$

$$\times \left(\frac{4\alpha^2\beta^2\eta^3\eta_l^3 L^2}{(1-\alpha'\beta)^5 K} + \frac{24\beta^2\eta\eta_l^3 L^2}{(1-\alpha'\beta)^3} + \frac{\eta^2\eta_l^2 L}{2(1-\alpha'\beta)^2} \right) - \frac{42\eta\eta_l^3 L^2 K^2}{1-\alpha'\beta} \Bigg).$$

We define a positive constant c such that $0 < c \leq \frac{1}{2} - \frac{42\eta_l^2 L^2 K^2}{\alpha} - \frac{\left(84C_{\beta_0}\alpha\eta_l^2 L^2 K^2 + 4(1+\beta_0)\alpha\right)(m-s)}{\kappa_1 s(m-1)} \left(\frac{4\alpha^2\beta^2\eta^2\eta_l^2 L^2}{(1-\alpha'\beta)^4 K} + \frac{24\beta^2\eta_l^2 L^2}{(1-\alpha'\beta)^2} + \frac{\eta\eta_l L}{2(1-\alpha'\beta)} \right)$. Rearranging the above inequality, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_t \|\nabla \hat{f}(x_t)\|^2 \\
&\leq \Bigg(\frac{3\eta_l^2 L^2 K}{\alpha} + \frac{\eta\eta_l L}{2(1-\alpha'\beta)s} + \frac{4\alpha^3\beta^2\eta^2\eta_l^2 L^2}{(1-\alpha'\beta)^4 sK} + \frac{24\alpha\beta^2\eta_l^2 L^2}{(1-\alpha'\beta)^2 s} \\
&\quad + \left(6C_{\beta_0}\alpha\eta_l^2 L^2 K + \frac{48C_{\beta_0}\alpha^3\beta^2\eta_l^2 L^2}{(1-\alpha'\beta)^2 s} \right) \frac{(m-s)}{\kappa_1 s(m-1)} \\
&\quad \times \left(\frac{4\alpha^2\beta^2\eta^2\eta_l^2 L^2}{(1-\alpha'\beta)^4 K} + \frac{24\beta^2\eta_l^2 L^2}{(1-\alpha'\beta)^2} + \frac{\eta\eta_l L}{2(1-\alpha'\beta)} \right) \Bigg) \frac{\tilde{\sigma}_l^2}{c} \\
&\quad + \Bigg(\frac{\left(84C_{\beta_0}\alpha\eta_l^2 L^2 K^2 + 4(1+\beta_0)\alpha\right)(m-s)}{\kappa_1 s(m-1)} \\
&\quad \times \left(\frac{4\alpha^2\beta^2\eta^2\eta_l^2 L^2}{(1-\alpha'\beta)^4 K} + \frac{24\beta^2\eta_l^2 L^2}{(1-\alpha'\beta)^2} + \frac{\eta\eta_l L}{2(1-\alpha'\beta)} \right) + \frac{42\eta_l^2 L^2 K^2}{\alpha} \Bigg) \frac{\sigma_g^2}{c} \\
&\quad + \frac{(1-\alpha'\beta)\left(\hat{f}(z_0) - \mathbb{E}[\hat{f}(z_T)]\right)}{c\alpha\eta\eta_l T}.
\end{aligned}$$

Then we set $\eta_l = \mathcal{O}\left(\frac{1}{LK\sqrt{T}}\right)$, $\eta = \mathcal{O}(\sqrt{sK})$. Since $\alpha = \mathcal{O}(K)$ and $0 < \beta_0 < 0.1$, we get:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_t \|\nabla \hat{f}(x_t)\|^2 \\
&\leq \frac{(1-\beta)L\left(\hat{f}(z_0) - \mathbb{E}[\hat{f}(z_T)]\right)}{c\sqrt{sKT}} + \Bigg(\frac{3}{K^2 T} + \frac{1}{2(1-\beta)\sqrt{sKT}} + \frac{4\beta^2 K}{(1-\beta)^4 T} + \frac{24\beta^2}{(1-\beta)^2 sKT} \\
&\quad + \left(\frac{6}{T} + \frac{48\beta^2 K}{(1-\beta)^2 sT} \right) \frac{(m-s)}{\kappa_1(m-1)} \left(\frac{4\beta^2}{(1-\beta)^4 T} + \frac{24\beta^2}{(1-\beta)^2 sK^2 T} + \frac{1}{2\sqrt{sKT}} \right) \Bigg) \frac{\sigma_l^2}{c} \\
&\quad + \left(\left(\frac{84K}{T} + 4K \right) \frac{(m-s)}{\kappa_1(m-1)} \left(\frac{4\beta^2}{(1-\beta)^4 T} + \frac{24\beta^2}{(1-\beta)^2 sK^2 T} + \frac{1}{2\sqrt{sKT}} \right) + \frac{42}{KT} \right) \frac{\sigma_g^2}{c}.
\end{aligned}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the main claims made in the abstract and introduction include our primary contributions and the assumptions underlying our scenarios.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our theoretical analysis relies on standard assumptions such as L-smooth, unbiased local gradient estimate, bounded client-variance, and Lipschitz continuity.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions used in our theoretical analysis are stated in the Preliminaries subsection. Complete proofs for all theoretical results are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the proposed method and provide pseudocode in Algorithm 1. We also report the experimental implementation details in Appendix A and provide code availability information in Section 5.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide open access to our code and detailed instructions on how to use the open-source data in our experiments in Appendix.A.3.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the training and test details in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the variance of the best validation accuracy across multiple runs. The corresponding results are provided in Appendix A.10.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computer resources in Appendix.A.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms to the NeurIPS Code of Ethics. The experiments are conducted on standard benchmark datasets and do not involve human subjects, personally identifiable information, or real-world deployment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is theoretical and does not target a specific application domain or real-world deployment. Thus, concrete societal impacts are not directly applicable.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve the release of data or models with a high risk of misuse. Therefore, specific safeguards for responsible release are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The datasets, code, and models used in this paper are properly credited, and their licenses and terms of use are explicitly mentioned and respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects or crowdsourced data collection. Therefore, participant risk disclosure and IRB approval are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for writing and language editing purposes, which do not affect the core methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.