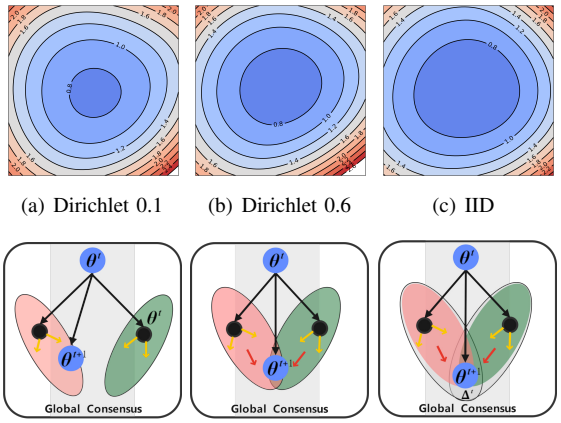
This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2025.3638762

Neighborhood and Global Perturbations Supported Sharpness-aware Minimization in Federated Learning: From Local Tweaks To Global Awareness

Boyuan Li, Zihao Peng, Yafei Li*, Zijian Li*, Shengbo Chen, *Member, IEEE* and Cong Shen, *Senior Member, IEEE*, Tony Q.S. Quek, *Fellow, IEEE*

***Abstract*—Federated Learning (FL) can be coordinated under the orchestration of a central server to build a privacy-preserving model without collaborative data exchange. However, participant data heterogeneity leads to local optima divergence, affecting convergence outcomes. Recent research focused on global sharpness-aware minimization (SAM) and dynamic regularization to enhance consistency between global and local generalization and optimization objectives in FL. Nonetheless, the estimation of global SAM introduces additional computational and memory overhead. At the same time, the local dynamic regularizer cannot capture the global update state due to training isolation. This paper proposes a novel FL algorithm, FedTOGA, designed to consider optimization and generalization objectives while maintaining minimal uplink communication overhead. By linking local perturbations to global updates, we improve global generalization consistency. Additionally, by linking the dynamic regularizer to global updates, FedTOGA improves global gradient perception and strengthens optimization consistency. Crucially, global updates are directly delivered to clients, allowing them to incorporate global knowledge without communication and computational cost. We also propose neighborhood perturbation to enhance local perturbation, analyzing its strengths and working principles. Theoretical analysis shows FedTOGA achieves faster convergence $O(1/T)$ on the non-convex function. Empirical studies demonstrate that FedTOGA outperforms existing algorithms, with a 1% accuracy increase and 30% faster convergence, achieving SOTA.**

***Index Terms*—Deep Learning, Federated Learning, Sharpness-aware Minimization, Data Heterogeneity, Global Perturbation.**



(a) Dirichlet 0.1 (b) Dirichlet 0.6 (c) IID

(d) FedSAM (e) FedSMOO (f) FedTOGA

Fig. 1. Fig.(a)-(c) shows the loss surface [9] under FL IID and the Non-IID setting, and Fig.(d)-(f) shows the FL system, where the gray color represents the global consensus and the colored regions represent the local knowledge. In Fig.(d), no further consensus can be increased in FL only supported by the SAM. In Fig.(e), a dynamic regularizer is introduced in some work to increase global generalization. In Fig.(f), we further introduce Global Update to enhance the generalization.

I. Introduction

The widespread connectivity of mobile terminals has substantially advanced the development of big data–driven industries. However, the massive data throughput has resulted in network congestion and heightened privacy risks. Consequently, to safeguard data privatization and localization, FL [1] has garnered significant attention as a distributed **machine learning** (ML) method that avoids the need for data exchange. Nonetheless, variations in data distribution among participants [2], [3] lead to conflicts in local optimization objectives, potentially causing the global loss function to converge to an undesirable sharp local minimum [4]. As illustrated in Figures 1(a)-1(c), increasing local heterogeneity markedly intensifies the global loss sharpness. Moreover, due to the limitations of uplink bandwidth to the global server [5], FL employs a "**Computation-Then-Aggregation**" (CTA) strategy [6], which utilizes multiple rounds of local training and partial participation to alleviate communication bottlenecks. However, by increasing synchronization intervals and reducing participation rates, the discrepancy between local and global models will be significantly amplified [7], [8].

To address these challenges, most studies tackle global consistency through Empirical Risk Minimization (ERM) [15]. However, with highly heterogeneous datasets, global solutions may become trapped in steep local minima. This not only hampers reliable estimation [13] but may also cause the optimizer to stall. To illustrate this, we conducted a small-scale experiment analyzing the loss landscape of FedAvg under varying degrees of data heterogeneity, as shown in Figures

Boyuan Li and Yafei Li were with the School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou, China, 450001, e-mail: (l202311841010602@gs.zzu.edu.cn;ieyfli@zzu.edu.cn).

Zihao Peng is with the School of Artificial Intelligence, Beijing Normal University, Beijing, China, (e-mail:pzh_cs@mail.bnu.edu.cn).

Zijian Li is with Dalian Maritime University, College of Artificial Intelligence (e-mail:lizj@dlmu.edu.cn).

Shengbo Chen is with Nanchang University, School of Software (e-mail:ccb02kingdom@gmail.com).

Cong Shen is with the University of Virginia (e-mail:cong@virginia.edu).

Tony Q.S. Quek is with the Singapore University of Technology and Design, Singapore 487372, and also with Yonsei Frontier Lab, Yonsei University, Seoul 03722, South Korea (e-mail: tonyquek@sutd.edu.sg).

Corresponding Author: Yafei Li, Zijian Li

TABLE I
ABSTRACT FOR THE SAM-BASED FL ALGORITHMS FOR SOLVING DATA HETEROGENEITY, FOCUSING ON THE BASIC ALGORITHM, SHARPNESS MINIMIZATION OBJECTIVE, PERTURBATION COMPUTATION STRATEGY, ADDITIONAL COMMUNICATION, AND STORAGE OVERHEAD COMPARISON.

Works	Base Algorithm	Minimizing Target	Local Perturbation	Extra Storage	Extra Communication
FedSAM	FedAvg	Local Sharpness	$\rho\frac{g_{i,k}^t}{\|g_{i,k}^t\|}$	$1\times$	$1\times$
MoFedSAM [10]	FedCM	Local Sharpness	$\rho\frac{g_{i,k}^t}{\|g_{i,k}^t\|}$	$1\times$	$1\times$
FedSpeed [11]	FedDyn	Local Sharpness	$\rho\frac{g_{i,k}^t}{\|g_{i,k}^t\|}$	$2\times$	$1\times$
FedGAMMA[12]	SCAFFOLD	Local Sharpness	$\rho\frac{g_{i,k}^t}{\|g_{i,k}^t\|}$	$2\times$	$2\times$
FedSMOO [13]	FedDyn	Local Sharpness With Correction	$\rho\frac{g_{i,k}^t-\mu_i-s}{\|g_{i,k}^t-\mu_i-s\|}$	$3\times$	$2\times$
FedLESAM(S-D)[14]	FedDyn SCAFFOLD	Global Sharpness Estimate	$\rho\frac{\theta_i^{old}-\theta_t}{\|\theta_i^{old}-\theta_t\|}$	$3\times$	$1\times$
FedTOGA(ours)	FedDyn FedCM	Local With Global Sharpness Estimate	$\rho\frac{g_{i,k}^t+[\tilde{g}_{i,k-1}^t]+\kappa\Delta^t}{\|g_{i,k}^t+[\tilde{g}_{i,k-1}^t]+\kappa\Delta^t\|}$	$2\times$	$1\times$

Note: ρ (perturbation radius), g (update gradient), μ, s, Δ (correction vectors), θ (model parameters). Details: Table XVIII.

1(a)–1(c). The loss surface becomes progressively sharper as local data heterogeneity increases. Consequently, recent studies have leveraged SAM [16], which enhances generalization by identifying flatter minima through minimizing the perturbed loss. Nonetheless, methods that minimize only the local sharpness loss [10] cannot adequately capture the flatness of the global loss surface. To this end, FedSMOO [13] introduces a dynamic regularizer to augment local generalization objectives and enforce global consistency, while FedLESAM [14] stores historical models to avoid additional computational overhead. Despite these advances, existing approaches still face several challenges. (1) FedSMOO incurs extra communication and storage costs, which are prohibitive in bandwidth-constrained real-world settings. (2) FedLESAM's global perturbation estimates, which lack real-time local gradient information, may yield insufficient local generalization. (3) When clients remain offline for extended periods, the bias in the global estimate can become significant. Table I summarizes the perturbation strategy of FedTOGA and its distinctions from related methods. Beyond ensuring consistency of the generalization objective, local optimization targets have also been extensively studied, such as dynamic regularization [6], [17], [13]. However, their performance degrades notably as the local interval widens, since the CTA strategy prevents clients from capturing the divergence between global and local update states.

To achieve a reliable, stable, and consistent global model, we propose a novel algorithm named FedTOGA, as illustrated in Fig.1(f). FedTOGA is designed to address three core challenges: (1) **How can we efficiently estimate global perturbations without adding extra computational overhead?** Due to the communication interval, the universal SAM optimizer applied on the global server cannot precisely capture the perturbations occurring during local updates on client devices. Therefore, FedTOGA initially guides the global update gradient to merge with local perturbations, thereby enhancing local generalization consistency without incurring additional cost. (2) **How can we reduce local computational overhead or use local resources to enhance generalization?** We introduce neighborhood gradient perturbation: when the communication interval between local training epochs exceeds one, clients simulate or enhance current perturbations by reusing cached gradients stored in a local gradient register, significantly lowering per-round computation and enhance generalization compared to existing methods. (3) **How can we further align global objectives?** Simultaneously, FedTOGA employs global updates to correct the local dynamic regularizer, originally driven by client drift statistics. Specifically, we incorporate the global update into the local dual term. This correction ensures both local and global stationary conditions, thereby reinforcing consistency with the overall optimization objectives. Moreover, unlike FedCM and MoFedSAM [10], the global update is not treated as a trade-off term with the local perturbation, meaning that their coefficients do not sum to one. As active local clients converge, they ultimately reach a globally stationary state characterized by a smooth loss landscape. This strategy significantly improves performance, even under extreme conditions with highly heterogeneous data or limited client participation.

Theoretically, FedTOGA can achieve a rapid convergence rate of $O(1/T)$ in non-convex settings. Extensive evaluations were conducted on the CIFAR10/100, OfficeHome, DomainNet, and Shakespeare datasets, demonstrating that FedTOGA achieves faster convergence rates and higher generalization accuracy in practice.

- We propose a novel FL algorithm, FedTOGA, the first global perturbation technique that uses a merged global update, and the first local dynamic regularizer that employs the global update. This method effectively reduces uplink communication overhead, ensuring rapid convergence and strong generalization.
- We introduce the concept of neighborhood perturbation to mitigate local computation and enhance generalization for the first time. This approach integrates or substitutes for local perturbation by leveraging gradient registers without incurring additional overhead. We further analyze its benefits and working principles.
- We provide a theoretical convergence analysis, demonstrating that FedTOGA attains $O(1/T)$ convergence rate in non-convex settings. Additionally, we performed extensive evaluations on the CIFAR10/100, OfficeHome, DomainNet and Shakespeare using various neural networks to validate the superior performance of FedTOGA, par-

ticularly in scenarios involving highly heterogeneous and sparse participants, where it significantly outperformed existing methods.

II. Related Works

In this section, we elaborate on the related works in detail and present all the research related to the proposed method.

Optimization-Based Federated Learning Federated learning gained widespread attention upon its introduction due to its data-exchange-free nature. FedAvg [1], as the foundational framework of FL, enables collaborative training without data exchange [18]. However, due to various irresistible factors, the data from cooperative devices show a heterogeneous distribution, which causes the modeling effectiveness to suffer. Therefore, many studies based on empirical loss minimization have been proposed to solve this problem. FedProx [19] employs a simple and intuitive practice, ensuring that the local model is not far from the global model. Specifically, regular terms are introduced during local training to limit the distance between the local and global models. SCAFFOLD [20], Mime [21] uses control variables for local updates. However, they incur greater communication overhead. FedDyn [17] and FedPD [6] consider the inconsistency between the local optimal point and the global optimal point to be a fundamental dilemma, aiming to align the locally optimal solution with the global optimal solution through a dynamic regularizer. FedPA [22] removes bias from client updates by estimating a global posterior. FedDC [23] takes decoupled local and global updates to mitigate heterogeneity. Furthermore, recent research has shown that model bias is similar to catastrophic forgetting in continuous learning [24], [25], [26], [27], [28]. Clients overriding previously important parameters to learn a new task result in the disruption of pre-task performance. Some studies have mitigated global knowledge collapse by task recall [29], [30]. Server momentum-based [31] algorithms also play an important role in FL. [32] investigates the convergence failure of ADAM in certain non-convex settings and develops an adaptive optimizer, YOGI, which aims to improve convergence. [33] integrates it into a FL framework. FedAvgM [34] uses Momentum [35], while FedACG [36] utilizes NAG [37]; FedExp [38] utilizes the extrapolation learning rate on the server. FedCM [39] mitigates local heterogeneity by incorporating a proximal global update gradient into the client-side momentum. FedLADA [40] combines local ADAM with FedCM to dynamically modify local deviations.

Personalized Federated Learning allows each client to maintain a partially local model while sharing global knowledge. Meta-learning strategies such as Per-FedAvg [41] and pFedMe [42] learn initialization parameters that are adaptable to client data. Parameter decomposition methods, including FedRep [43], Ditto [44], and FedAMP [45], separate shared representations from local heads. More recent efforts, pFedHN [46], FedPCL [47], and FedALA [48], employ hypernetworks, contrastive regularization, and meta-adaptive aggregation. These approaches significantly mitigate heterogeneity while improving fairness and stability across diverse clients.

Clustered and Adaptive Aggregation Federated Learning When global consensus is suboptimal, clustered federated learning divides clients into groups with similar distributions. IFCA [49] and CFL [50] established early frameworks; later works, such as FedSoft [51] and FedGroup [52], introduced dynamic assignment and attention-based grouping. Adaptive aggregation methods, including FedMA [53], FedDF [54], and FedDRL [55], weight updates according to divergence or reliability metrics. Graph-based formulations [56] further model inter-client relations, enabling structure-aware aggregation under heterogeneous connectivity.

Federated Learning with Knowledge Distillation Representation level alignment alleviates feature inconsistency caused by non-IID data. MOON [57] introduces contrastive consistency between local and global models, while FedAlign [58] employs embedding matching. Distillation-based methods such as FedMD [59], FedGKD [60], and FedFTG [61] exchange logits instead of weights, facilitating collaboration among heterogeneous architectures. Recent studies incorporate foundation models and self-supervised objectives, FedCLIP [62], FedPrompt [63], and FedDistill [64],to enhance generalization under cross-domain and heterogeneity.

Asynchronous Federated Learning Beyond statistical heterogeneity, device variability and communication delays lead to system heterogeneity. Asynchronous aggregation frameworks (FedAsync [65], FedBuff [66]) allow clients to proceed independently, thereby reducing straggler effects. Resource-aware schemes such as FedSA [67] and FedAdapt [68] dynamically adjust local epochs and participation rates. These system-aware algorithms align practical efficiency with theory.

Sharpness-aware Minimization Many studies [69], [70] have pointed out that flat minima imply superior generalization performance, which is more robust to model perturbations. In order to minimize sharpness[71], [16], [72] proposed sharpness-aware minimization, and many works [73], [74] have been carried out.

SAM in Federated Learning To improve the generalizability of local models in FL, [10], [75] introduced SAM into the FL paradigm, proposing FedSAM. Furthermore, they combined FedSAM with FedCM [39] to develop MoFedSAM. FedGAMMA [12] integrated the variance-reduction technique from SCAFFOLD [20] into FedSAM, achieving improved convergence stability. In addition, FedSpeed [11] leverages SAM to enhance the optimization process of FedDyn [17]. Building upon this, FedSMOO [13] incorporates dynamic regularization into SAM to estimate global perturbations, while FedSOL [25] adopts an orthogonal-learning strategy inspired by continual learning to align local perturbations with the global objective. FedLESAM [14] proposed an efficient algorithm, Local Estimation of Global Perturbations SAM, which optimizes global sharpness while reducing computational costs. FedOMG [76] employs data-free on-server gradient matching to achieve domain generalization, while FedGMT [77] regularizes sharpness through global trajectory tracking.Overall, FedSAM, MoFedSAM, and FedGAMMA optimize sharpness based on locally computed perturbations, which may prevent convergence to the global flat minimum. Several subsequent studies have identified this limitation and proposed remedies: FedSOL restricts perturbation directions to reduce divergence but may cause perturbation deficiencies;

FedSMOO introduces dynamic regularization but incurs additional communication and storage overhead; and FedLESAM estimates global perturbations from historical parameters to avoid excessive computation, yet remains sensitive to network fluctuations. Motivated by these challenges, we propose FedTOGA, which estimates global perturbations directly from the global update to achieve improved generalization under heterogeneous environments.

TABLE II
BASIC NOTATIONS. DETAILED EXPLANATIONS REFER TO TABLE XVIII

i, k, t	Number of the client, local training interval and global epoch.
η_l, ρ	Local learning rate and perturbation learning rate.
D, D_i	Data distributions of global and i -th client.
h, h_i	Global and local dual variables.
Δ^t	Global update gradient in t-th round.
$\theta, \theta^t, \theta^t_{i,k}$	Model weights and weights of global and local models.
δ	Perturbation of θ

III. PRELIMINARIES

Federated Learning The goal of the FL framework is to build a global model that minimizes the average empirical loss of participating clients:

$$\arg\min_{\theta} f(\theta) = \frac{1}{N}\sum_{i\in N} f_i(\theta); f_i(\theta) \triangleq \mathbb{E}_{\xi_i\sim D_i} f_i(\theta, \xi_i). \quad (1)$$

Here $f : \mathbb{R} \to \mathbb{R}^d$ denotes the global objective function, θ is a model parameter, N is the total number of participating clients, and ξ_i is a randomly samples point from the distribution D_i subject to data heterogeneity. f_i is the loss for the i-th client.

Sharpness Aware Minimization Many studies [69], [70] have shown that a flat minimum implies better generalization performance and possesses greater robustness to perturbations. To minimize sharpness, [71], [16] SAM proposes:

$$\arg\min_{\theta} \left\{ f_{sam}(\theta) = \arg\max_{\|\delta\|\le\rho} f(\theta+\delta) \right\}. \quad (2)$$

In practice, the inner maximization is typically approximated by one step of gradient ascent followed by one step of gradient descent to improve generalization and reduce loss. First, calculate the gradient ascent perturbation $\delta = \rho\frac{\nabla f(\theta)}{\|\nabla f(\theta)\|}$. Then, the model gradient is calculated after adding the perturbation, and the model is updated $\theta = \theta - \eta\nabla f(\theta+\delta)$. The basic notation is summarized in Table II.

A. Rethinking FedSAM and Related Works

The limitations of applying the SAM optimizer directly within FL systems [10] have been extensively discussed [13], [14], [25]. The core conflict arises from the fundamental difference between the *centralized training* setting assumed by standard SAM and the *distributed, client-based computing* paradigm of FL. Specifically, the centralized SAM [10] objective requires solving a min-max optimization problem:

$$\begin{aligned}&\min_{\theta}\left[\max_{\|\delta\|\le\rho} \mathbb{E}_{\xi\sim D} f(\theta+\delta,\xi)\right]\\ =&\min_{\theta}\left[\max_{\|\delta\|\le\rho} \mathbb{E}_i\mathbb{E}_{\xi_i\sim D_i} f(\theta+\delta,\xi_i)\right],\end{aligned} \quad (3)$$

where $D = \mathbb{E}_i D_i$ denotes the global data distribution. When directly applying SAM in FL [11], [12], the formulation becomes client-specific:

$$\min_{\theta_i}\left[\max_{\|\delta_i\|\le\rho} \mathbb{E}_{\xi_i\sim D_i} f_i(\theta_i+\delta_i,\xi_i)\right], \quad (4)$$

where $\delta_i = \rho\frac{\nabla f_i(\theta_i)}{\|\nabla f_i(\theta_i)\|}$. Due to the CTA mechanism, this method isolates the global model θ from client-specific perturbations δ_i computed on local parameters θ_i. Consequently, minimizing local sharpness through δ_i fails to produce a global flat minimum. This inconsistency exacerbates as both the local update interval and data heterogeneity increase [14], making global-client alignment progressively more challenging.

We show in Table I how the existing algorithms handle the isolation of perturbations. Some recent studies, FedSAM [75], MoFedSAM [10], FedGAMMA [12], FedSpeed [11] have not resolved the internal perturbation variance (measures the variance between client's and global perturbation, reflecting the inconsistency of local perturbation under data heterogeneity. i.e., $\delta_i \neq \delta$), only handling the local perturbation $\rho\frac{g^t_{i,k}}{\|g^t_{i,k}\|}$. FedSMOO [13] identifies this contradiction for the first time and uses dynamic regularization to correct the discrepancy between local and global perturbations $\rho\frac{g^t_{i,k}-\mu_i-s}{\|g^t_{i,k}-\mu_i-s\|}$. However, FedSMOO introduces additional computation, which increases client-side computational overhead in FL. FedLESAM [14] argues that computing the perturbations requires additional computation and communication; therefore, it allocates additional local storage to approximate the estimated global perturbations $\rho\frac{\theta^{old}_i-\theta_t}{\|\theta^{old}_i-\theta_t\|}$. However, as the set of activated clients S_t decreases sharply, the perturbation estimation by FedLESAM [14] is more affected. For a more detailed description of the limitations, see Appendix C.

B. Rethinking Dynamic Regularizer in FL

Dynamic regularization in FL [78], [17], [13]is designed to mitigate bias in local optimization. A common form uses an augmented Lagrangian across clients:

$$F_{fed} = \frac{1}{N}\sum_{i\in N}\left\{ f_i + \left\langle h^t_i, \theta^t - \theta^t_i\right\rangle + \frac{1}{2\alpha}\|\theta^t - \theta^t_i\|^2 \right\}. \quad (5)$$

Here, each dual variable h^t_i acts as a "correction vector", which guides local parameters θ^t_i toward the global model θ^t [6], [78]. From the first-order condition, $\nabla f_i(\theta^t_i) - h^t_i + \frac{1}{\alpha}(\theta^t_i - \theta^t) = 0$, it follows that local updates converge to stationary points. With the dual update $h^{t+1}_i = h^t_i - \frac{1}{\alpha}(\theta^t_{i,K} - \theta^t_{i,0})$, one obtains $\nabla f_i(\theta^t_i) - \nabla f_i(\theta^{t-1}_i) + \frac{1}{\alpha}(\theta^t_i - \theta^t) = 0$. As $t\to\infty$, $\theta^t_i \to \theta^t$, under ideal conditions [17], [13]. However, data heterogeneity causes each local subproblem to converge to a different stationary point. Global optimality requires $\sum_{i=1}^N \nabla f_i(\theta^*) = 0$, so any individual $-\nabla f_i(\theta^*)$ could be offset by others. Consequently, approaches that rely solely on local correction are insufficient. It is therefore necessary to integrate the global stationary condition into local optimization.

IV. METHODOLOGY

This section introduces FedTOGA, as shown in algorithm 1, and provides an analysis of the three key techniques.

Algorithm 1: FedTOGA Algorithm

1 Initial model parameters θ^0, initial global update Δ^{-1}, local dual variable h_i, global dual variable h, local perturbation gradient $\tilde{g}_{i,-1}$, communication rounds T, penalized coefficient for quadratic term α, Correction coefficient for perturbation and dual term κ, β.
2 **Server execute:**
3 **for** *each round* $t \in [T] \triangleq \{0, 1, 2, \cdots, T-1\}$ **do**
4 Sample the active client set $S_t \subseteq [N]$.
5 **for** $i \in S_t$ *in parallel* **do**
6 $\theta_i^{t+1} \leftarrow$ **Client Update**(θ^t, Δ^t);
7 **communicate** θ_i^t to server ;
8 **end**
9 $h^{t+1} = h^t - \frac{1}{\alpha M} \sum_{i \in S_t} (\theta_i^{t+1} - \theta^t); |S_t| = M$
10 $\Delta^{t+1} = -\frac{1}{MK} \sum_{i \in S_t} (\theta_i^{t+1} - \theta^t)$;
11 $\theta^{t+1} = \frac{1}{M} \sum_{i \in S_t} \theta_i^{t+1} - \alpha h^{t+1}$
12 **end**
13 **Client Update**(θ_t, Δ_t)**:** $\theta_{i,0}^t = \theta^t$
14 **for** *local epoch* $k \in [K] \triangleq \{0, 1, 2, \cdots, K-1\}$ **do**
15 sample a mini-batch data $\xi_{i,k}^t$;
16 gradient estimate: $g_{i,k}^t = \tilde{\nabla} f_i(\theta_{i,k}^t; \xi_{i,k}^t)$
17 Perturbation: $\delta_{i,k}^t = \rho \frac{g_{i,k}^t + [\tilde{g}_{i,k-1}^t] + \kappa \Delta^t}{\|g_{i,k}^t + [\tilde{g}_{i,k-1}^t] + \kappa \Delta^t\|}$
18 extra-step: $\tilde{g}_{i,k}^t = \nabla f_i(\theta_{i,k}^t + \delta_{i,k}^t; \xi_{i,k}^t); \theta_{i,k+1}^t =$
19 $\theta_{i,k}^t - \eta_l \big(\tilde{g}_{i,k}^t - h_i^t + \frac{1}{\alpha} \left(\theta_{i,k}^t - \theta_{i,0}^t \right) + \beta \Delta^t \big)$
20 **end**
21 $h_i^{t+1} = h_i^t - \frac{1}{\alpha} \left(\theta_{i,K}^t - \theta_{i,0}^t \right)$
22 **return** $\theta_i^{t+1} = \theta_{i,K}^t$

A. Estimate Global Perturbation

This subsection describes the computation of the global update and its integration into client-side perturbations, as implemented in Lines 10 and 17 of Algorithm 1. As discussed in Section III-A, our goal is to efficiently estimate the global perturbation for each client while avoiding additional storage or computational overhead. We first recall the formulation of global sharpness-aware minimization in FedLESAM [14],

$$\min_{\theta} \Big\{ f = \frac{1}{N} \sum_{i \in N} \max_{\|\delta_i\| \le \rho} \mathbb{E}_i \mathbb{E}_{\xi_i \sim D_i} f_i(\theta_i + \delta_i, \xi_i) \Big\}. \tag{6}$$

Therefore, at round t and virtual step k, the virtual global perturbation is defined as $\delta_k^t = \rho \frac{\nabla f(\theta^t)}{\|\nabla f(\theta^t)\|} = \rho \frac{\sum_{i \in N} \nabla f_i(\theta_k^t)}{\|\sum_{i \in N} \nabla f_i(\theta_k^t)\|} \approx \rho \frac{\sum_{i \in S} \nabla f_i(\theta_k^t)}{\|\sum_{i \in S} \nabla f_i(\theta_k^t)\|}$, where S denotes the set of selected clients. Here, θ_k^t represents the global model at virtual step k, computed as $\theta_k^t = \frac{1}{M} \sum_{i \in S_t} \theta_{i,k}^t$. However, due to the CTA strategy in FL systems, clients cannot directly access the global model θ_k^t at each time step. Consequently, the global perturbation δ cannot be accurately computed. FedLESAM approximates the global perturbation by storing historical model parameters; however, this estimation neglects local real-time gradient updates, resulting in degraded accuracy when the local training interval increases. Inspired by the FedCM [39] strategy, we estimate the global update $\Delta^t \approx \nabla f(\theta^t)$ by propagating the global update variable to clients and incorporating local real-time updates. Finally, the global perturbation update strategy in FedTOGA is defined as follows: $\delta_k^t = \rho \frac{\nabla f(\theta^t)}{\|\nabla f(\theta^t)\|} \approx \rho \frac{g_{i,k}^t + \kappa \Delta^t}{\|g_{i,k}^t + \kappa \Delta^t\|}; \theta_{i,k}^t = \theta_{i,k-1}^t - \eta_l \nabla f_i(\theta_{i,k}^t + \rho \delta_k^t)$. A comparison of FedTOGA's perturbation strategy with existing methods is provided in Table I.

B. Utilize Neighbourhood Perturbation

The mechanism proposed in this section is implemented in Line 17 of Algorithm 1. Besides, as stated by FedLESAM[14], local perturbations require additional gradient ascent computations, which may consume extra computational overhead. Therefore, how can we estimate the local perturbation without utilizing additional computations? We propose neighborhood perturbation [history gradient] for the first time. Specifically, when the client's local iteration interval exceeds one, the local perturbation gradient $\tilde{g}_{i,k-1}^t$ will be recorded by the cache without opening additional storage space. We can get $g_{i,k} \approx \tilde{g}_{i,k-1}^t$. We can further replace the perturbation term in the local SAM optimization and get: $\delta_{i,k}^t = \rho \frac{\tilde{g}_{i,k-1}^t + \kappa \Delta^t}{\|\tilde{g}_{i,k-1}^t + \kappa \Delta^t\|}$. This operation allows for the approximate estimation of local perturbations in environments with scarce client-side resources. The $[\cdot]$ means that it activates only when $K > 1$.

Perturbation Fusion In the FL paradigm, the client SAM captures only the sharpness of a specific small batch of data, which is mitigated by the global perturbation technique described above to enhance generalization. Let's consider whether neighborhood perturbation may bring additional benefits. Similar to LookAhead[79], it backtracks by perturbing ascent after each gradient descent. Then, perturbation can be rewritten: $\delta_{i,k}^t = \rho \frac{g_{i,k} + \tilde{g}_{i,k-1}^t + \kappa \Delta^t}{\|g_{i,k} + \tilde{g}_{i,k-1}^t + \kappa \Delta^t\|}$. A more in-depth discussion can be found in Appendix D.

C. Global Correction in Dynamic Regularizer

To further enhance the consistency of the optimization objective, FedTOGA incorporates the global update correction term Δ for each local clients' dynamic regularization(Line 19 of Algorithm 1). On the server side, this procedure follows an ADMM-like method, as in FedSMOO [13], to effectively minimize the global objective f [6]. FedTOGA is the first FL framework that explicitly incorporates the **global update** Δ in both perturbation and regularization. We define the global Augmented Lagrangian (AL) function F_{fed}, which introduces a penalty term enforcing the constraint $\theta = \theta_i$, as follows:

$$F_{\text{fed}} = \frac{1}{N} \sum_{i \in N} \left\{ f_i + \langle h_i^t, \theta^t - \theta_i^t \rangle + \frac{1}{2\alpha} \|\theta^t - \theta_i^t\|^2 \right\}. \tag{7}$$

In general, the global objective is decomposed across clients, and each client minimizes its local AL subproblem:

$$\theta_{i,K}^t = \min_{\theta_i} \left\{ f_i - \langle h_i^t, \theta_i^t \rangle + \frac{1}{2\alpha} \|\theta^t - \theta_i^t\|^2 \right\}. \tag{8}$$

The global dual variable h is updated at each communication, while each client maintains a local dual variable h_i. As discussed in Section III-B, storing h_i helps mitigate local target drift but ignores the global gradient trend, which previous studies [17], [13] have not addressed. To resolve this issue, FedTOGA estimates a global update variable Δ^t by aggregating client-side gradients. Unlike the dual variable h_i,

which drives the global aggregation step, Δ^t serves as a **global direction estimator**, providing additional gradient information to align local optimization steps with the global objective. Specifically, the corrected local dual variable is expressed as $h_i^t - \beta\Delta^t$, leading to the modified local subproblem:

$$\theta_{i,K}^t = \min_{\theta_i} \left\{ f_i - \langle h_i^t - \beta\Delta^t, \theta_i^t \rangle + \frac{1}{2\alpha}\|\theta^t - \theta_i^t\|^2 \right\}. \quad (9)$$

Here, β controls the strength of global update injection. After solving the local problem, each client updates its dual variable as $h_i^{t+1} = h_i^t - \frac{1}{\alpha}\left(\theta_{i,K}^t - \theta_{i,0}^t\right)$. The global model θ^{t+1} is then updated using the aggregated dual variables according to Eq. 7.

Specifically, integrating Δ into perturbation calculations improves the accuracy of planar global minimum estimation, thereby enhancing generalization consistency. Meanwhile, adding Δ to the dual regularizer not only takes into account the local stability points but also further considers the global objective, enhancing the consistency of the optimization. The combined effect of these two mechanisms enables FedTOGA to converge rapidly to a consistent and flat global optimum.

D. Overview of FedTOGA

Algorithm 1 shows the detailed flow of FedTOGA. First, initialize the server-side global model θ. In the global synchronization round t, a set S_t containing M clients is randomly selected from all clients N, and the global model θ_t is sent to the set of authorized clients S_t with the global update Δ^t of the $t-1$ round. The client first computes the original gradient $g_{i,k}^t$ according to Line 16 and subsequently computes the SAM gradient $\delta_{i,k}^t$ corrected by Δ^t in Line 17, with the neighborhood perturbation variable $\tilde{g}_i$ being optional. In Line 19, we use the formula 9 for local dual variable correction to update the local model θ_i and update the local dual variable h_i via Line 21. After local training, FedTOGA sends only θ_i^t to the server for aggregation. In lines 9-11 of the algorithm, the server updates the global model from θ^t to θ^{t+1} by minimizing the function 7. This process is repeated until $T-1$.

V. Theoretical Analysis

Assumption 1. *The loss function f_i is L-Smooth, i.e., $f_i(y) - f_i(x) \leq \langle \nabla f_i(x), y-x\rangle + \frac{L}{2}\|y-x\|^2$.*

Assumption 2. *Unbiased and bounded variance of stochastic gradient. The stochastic gradient $\tilde{\nabla} f_i(x) = \nabla f_i(x, \xi_i)$ computed by the i-th client using mini-batch ξ is an unbiased estimator of $\nabla f_i(x)$, i.e.*

$$\mathbb{E}[\tilde{\nabla} f_i(x)] = \nabla f_i(x), \mathbb{E}\|\tilde{\nabla} f_i(x) - \nabla f_i(x)\|^2 \leq \sigma_l^2.$$

Assumption 3. *Bounded Heterogeneity, for all $x \in \mathbb{R}^d$, we establish the following inequality: $\mathbb{E}\|\nabla f_i(x) - \nabla f(x)\| \leq \sigma_g$ Besides, the variance of the unit gradient is bounded: $\mathbb{E}\|\frac{\nabla f_i(x)}{\|\nabla f_i(x)\|} - \frac{\nabla f(x)}{\|\nabla f(x)\|}\| \leq \sigma_g'$ [13], [75].*

Theorem 1. *Under Assumption 1-3, for any training interval t on the i-th client, model divergence satisfies:*

$$\|\theta_{i,k}^t - v_k^t\|^2 \leq H_i(k) \quad (10)$$

where $H_i(\tau) \leq \frac{L^2\rho^2\sigma_g'^2 + \sigma_g^2}{2L^2}((1+2\eta_l^2L^2)^\tau - 1)$ and $\{v^t\}$ are virtual sequences representing the global model. More details are in Appendix A.

Remark 1. *The discrepancy between local and global models tends to grow geometrically as the local training interval increases, primarily due to the accumulation of internal perturbations and update variance. Therefore, it is crucial to improve the consistency between the optimization and generalization objectives (see Section III-A III-B IV).*

Theorem 2. *Under Assumption 1-3. When $\eta_l \leq \min\{\frac{1}{\sqrt{2128L^2K}}, \alpha\}$, and the perturbation learning rate satisfies $\rho = O(1/\sqrt{T})$, and the local interval $K > \frac{\alpha}{\eta_l}$, let $\omega = \frac{1}{2} + \beta - 2128\eta_l^2L^2K - \beta^2 - L\alpha\beta^2$ be a positive constant while selecting the suitable η_l. the auxiliary sequence $\{z^t\}$ generated by executing Algorithm 1 satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(z^t)\|^2 \leq \frac{f(z^0) - f^*}{T\alpha\omega} + \Upsilon + \frac{16\alpha^3L^2}{T\omega}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^0\right\|^2 + \frac{112L^2\eta_l^2K}{TN\omega}\sum_{i\in N}\mathbb{E}\|h_i^0\|^2$$

where the f^* is the optimum of the non-convex function f, and the term Υ is:

$$\Upsilon = \frac{1}{\omega}\left(56\eta_l^2L^2K(3\sigma_l^2 + 16\sigma_g^2 + 5L^2\rho^2) + L^2\rho^2\right)$$

More details are in Appendix B. Here, T denotes the number of global communication rounds, while K denotes the number of local iterations per round; together, they determine the algorithm's convergence behavior.

Remark 2. *When we set the local learning rate η_l to satisfy $\eta_l = O(1/K)$ and the perturbation learning rate to be $O(1/T)$, FedTOGA can achieve a fast convergence rate of $O(1/T)$ when the local interval K satisfies $K = O(T)$.*

Remark 3. *As mentioned in Section III-B, the proof of [78], [17], [13] relies on the strict assumption that the client must approach a stationary point in each round of training, which cannot be strictly fulfilled in the FL system. Therefore, we do not consider the strict assumption of the first-order condition $\tilde{g}_i^t - h_i^t + \frac{1}{\alpha}(\theta_i^t - \theta^t) + \beta\Delta^t = 0$ of the equation (9) in each round t. Inspired by [11], we relax this assumption by enlarging the local intervals, which also achieves $O(1/T)$ convergence speed [13]. Appendix G provides a further extended analysis.*

Remark 4. *Inspired by [11], FedTOGA can also speed up convergence by increasing the local interval K setting, which is helpful for bandwidth-constrained FL systems. However, the local perturbation learning rate ρ in FedSpeed is specified by the upper bound $\frac{1}{\sqrt{6}\alpha L}$, and our proof relaxes the limitation so that the ρ only needs to satisfy $O(1/T)$. We can also tighten the boundary of $\frac{1}{\omega}$ by adjusting β appropriately.*

A. Connection with other works

We show how to generalize FedTOGA to **FedDyn**, **FedPD**, and **FedProx** without considering local perturbations, recalling the AL function defined in Eqn.(7). By setting h_i to zero (i.e., ignoring the dual update) and neglecting the global update (i.e., $\Delta^t \equiv 0$) , the local objective of FedTOGA reduces to the form of FedProx. Additionally, setting $\Delta^t \equiv 0$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2025.3638762

TABLE III

DIRICHLET COEFFICIENTS u SELECTED FROM $\{0.1, 0.6\}$, AND c IS THE PATHOLOGICAL COEFFICIENT, I.E., THE NUMBER OF ACTIVE CATEGORIES IN EACH CLIENT. THE TWO DATASETS HAVE 100 CLIENTS IN THE UPPER PART WITH 10% ACTIVE IN EACH ROUND, 200 CLIENTS IN THE LOWER PART WITH 5% ACTIVE IN EACH ROUND.(LENET)

Method	CIFAR10				CIFAR100			
Partition	Dirichlet		Pathological		Dirichlet		Pathological	
Coefficient	$u=0.6$	$u=0.1$	$c=6$	$c=3$	$u=0.6$	$u=0.1$	$c=20$	$c=10$
FedAvg [1]	$80.28^{\pm0.14}$	$74.68^{\pm0.19}$	$80.59^{\pm0.18}$	$78.10^{\pm0.23}$	$47.35^{\pm0.16}$	$45.56^{\pm0.20}$	$46.46^{\pm0.20}$	$43.43^{\pm0.27}$
FedAdam	$80.39^{\pm0.17}$	$71.52^{\pm0.29}$	$81.02^{\pm0.20}$	$77.88^{\pm0.23}$	$48.94^{\pm0.21}$	$43.62^{\pm0.25}$	$44.86^{\pm0.25}$	$41.58^{\pm0.27}$
FedYogi [33]	$80.11^{\pm0.19}$	$73.58^{\pm0.25}$	$81.08^{\pm0.21}$	$78.10^{\pm0.20}$	$48.41^{\pm0.21}$	$45.44^{\pm0.22}$	$46.18^{\pm0.22}$	$42.07^{\pm0.25}$
SCAFFOLD[20]	$82.87^{\pm0.12}$	$78.00^{\pm0.16}$	$83.31^{\pm0.10}$	$80.29^{\pm0.15}$	$53.68^{\pm0.21}$	$50.33^{\pm0.24}$	$51.30^{\pm0.22}$	$47.71^{\pm0.22}$
FedACG[36]	$82.87^{\pm0.14}$	$77.51^{\pm0.16}$	$82.86^{\pm0.12}$	$80.84^{\pm0.17}$	$52.88^{\pm0.20}$	$48.72^{\pm0.23}$	$50.24^{\pm0.21}$	$46.08^{\pm0.24}$
FedCM[39]	$77.04^{\pm0.30}$	$62.75^{\pm0.31}$	$66.58^{\pm0.29}$	$71.20^{\pm0.33}$	$43.08^{\pm0.19}$	$34.69^{\pm0.26}$	$36.27^{\pm0.18}$	$28.48^{\pm0.30}$
FedDyn[17]	$82.31^{\pm0.13}$	$78.05^{\pm0.19}$	$83.13^{\pm0.18}$	$79.96^{\pm0.19}$	$49.97^{\pm0.19}$	$45.85^{\pm0.29}$	$47.41^{\pm0.21}$	$43.29^{\pm0.19}$
FedDC[23]	$83.58^{\pm0.14}$	$78.50^{\pm0.19}$	$84.00^{\pm0.16}$	$81.72^{\pm0.17}$	$51.99^{\pm0.15}$	$48.75^{\pm0.21}$	$49.53^{\pm0.19}$	$44.82^{\pm0.23}$
FedRCL [80]	$77.62^{\pm0.11}$	$68.79^{\pm0.16}$	$78.28^{\pm0.15}$	$76.04^{\pm0.19}$	$46.34^{\pm0.24}$	$42.28^{\pm0.17}$	$44.06^{\pm0.19}$	$39.64^{\pm0.21}$
FedSAM	$81.58^{\pm0.15}$	$77.67^{\pm0.15}$	$82.15^{\pm0.17}$	$79.23^{\pm0.23}$	$48.08^{\pm0.21}$	$46.86^{\pm0.26}$	$46.71^{\pm0.25}$	$43.41^{\pm0.22}$
MoFedSAM [10]	$77.17^{\pm0.12}$	$66.24^{\pm0.15}$	$77.44^{\pm0.15}$	$72.15^{\pm0.19}$	$43.30^{\pm0.18}$	$34.43^{\pm0.21}$	$36.50^{\pm0.19}$	$29.92^{\pm0.24}$
FedGAMMA[12]	$83.88^{\pm0.13}$	$78.61^{\pm0.15}$	$83.79^{\pm0.14}$	$79.68^{\pm0.15}$	$53.94^{\pm0.20}$	$49.95^{\pm0.24}$	$51.20^{\pm0.22}$	$48.11^{\pm0.29}$
FedSMOO [13]	$\mathbf{84.82}^{\pm0.15}$	$\mathbf{80.06}^{\pm0.16}$	$\mathbf{85.07}^{\pm0.17}$	$81.26^{\pm0.19}$	$\mathbf{56.57}^{\pm0.18}$	$52.17^{\pm0.17}$	$53.42^{\pm0.21}$	$48.12^{\pm0.19}$
FedSpeed [11]	$84.14^{\pm0.15}$	$80.16^{\pm0.16}$	$84.74^{\pm0.14}$	$\mathbf{82.20}^{\pm0.19}$	$53.96^{\pm0.19}$	$\mathbf{52.29}^{\pm0.21}$	$53.78^{\pm0.18}$	$\mathbf{48.33}^{\pm0.20}$
FedLESAM	$80.94^{\pm0.18}$	$77.02^{\pm0.15}$	$81.79^{\pm0.18}$	$78.85^{\pm0.15}$	$48.13^{\pm0.18}$	$46.55^{\pm0.21}$	$46.08^{\pm0.23}$	$43.57^{\pm0.17}$
FedLESAM-D	$83.28^{\pm0.15}$	$79.12^{\pm0.18}$	$84.20^{\pm0.19}$	$80.91^{\pm0.16}$	$54.88^{\pm0.18}$	$52.08^{\pm0.22}$	$\mathbf{54.14}^{\pm0.19}$	$48.28^{\pm0.22}$
FedLESAM-S[14]	$83.39^{\pm0.12}$	$78.23^{\pm0.17}$	$83.99^{\pm0.19}$	$81.20^{\pm0.15}$	$53.29^{\pm0.15}$	$50.12^{\pm0.21}$	$52.20^{\pm0.20}$	$47.29^{\pm0.17}$
FedOMG[76]	$85.31^{\pm0.14}$	$80.56^{\pm0.18}$	$85.41^{\pm0.17}$	$81.88^{\pm0.19}$	$56.82^{\pm0.17}$	$52.46^{\pm0.16}$	$53.70^{\pm0.20}$	$48.45^{\pm0.20}$
FedGMT[77]	$85.67^{\pm0.13}$	$81.04^{\pm0.14}$	$85.55^{\pm0.15}$	$82.30^{\pm0.16}$	$56.98^{\pm0.15}$	$52.80^{\pm0.16}$	$54.00^{\pm0.15}$	$48.73^{\pm0.18}$
FedTOGA(ours)	$\mathbf{86.01}^{\pm0.12}$	$\mathbf{82.05}^{\pm0.11}$	$\mathbf{85.71}^{\pm0.13}$	$\mathbf{84.00}^{\pm0.12}$	$\mathbf{57.25}^{\pm0.13}$	$\mathbf{53.45}^{\pm0.13}$	$\mathbf{55.49}^{\pm0.13}$	$\mathbf{51.27}^{\pm0.18}$
FedAvg [1]	$77.53^{\pm0.17}$	$74.60^{\pm0.23}$	$79.21^{\pm0.25}$	$76.20^{\pm0.23}$	$43.86^{\pm0.21}$	$42.70^{\pm0.24}$	$42.94^{\pm0.25}$	$42.28^{\pm0.29}$
FedAdam	$79.39^{\pm0.19}$	$74.49^{\pm0.31}$	$79.53^{\pm0.23}$	$76.09^{\pm0.25}$	$45.34^{\pm0.25}$	$42.79^{\pm0.23}$	$43.57^{\pm0.25}$	$40.66^{\pm0.29}$
FedYogi [33]	$79.95^{\pm0.21}$	$75.29^{\pm0.25}$	$79.73^{\pm0.22}$	$77.64^{\pm0.23}$	$46.67^{\pm0.25}$	$43.02^{\pm0.24}$	$44.70^{\pm0.27}$	$41.33^{\pm0.30}$
SCAFFOLD[20]	$81.18^{\pm0.15}$	$76.11^{\pm0.19}$	$82.44^{\pm0.17}$	$78.52^{\pm0.17}$	$51.45^{\pm0.25}$	$47.19^{\pm0.27}$	$48.26^{\pm0.28}$	$46.82^{\pm0.26}$
FedACG[36]	$82.57^{\pm0.17}$	$78.47^{\pm0.20}$	$82.09^{\pm0.16}$	$80.50^{\pm0.19}$	$51.96^{\pm0.24}$	$49.34^{\pm0.26}$	$50.01^{\pm0.27}$	$46.82^{\pm0.25}$
FedCM[39]	$76.08^{\pm0.30}$	$64.33^{\pm0.31}$	$76.64^{\pm0.29}$	$68.61^{\pm0.33}$	$40.32^{\pm0.19}$	$33.05^{\pm0.26}$	$34.19^{\pm0.18}$	$27.88^{\pm0.30}$
FedDyn[17]	$80.60^{\pm0.17}$	$77.53^{\pm0.21}$	$81.54^{\pm0.22}$	$79.39^{\pm0.24}$	$48.40^{\pm0.20}$	$45.04^{\pm0.31}$	$46.87^{\pm0.24}$	$43.04^{\pm0.29}$
FedDC[23]	$81.83^{\pm0.17}$	$78.87^{\pm0.21}$	$82.44^{\pm0.17}$	$80.93^{\pm0.19}$	$48.74^{\pm0.19}$	$45.11^{\pm0.26}$	$45.94^{\pm0.22}$	$43.94^{\pm0.27}$
FedRCL [80]	$76.06^{\pm0.15}$	$66.88^{\pm0.19}$	$76.51^{\pm0.19}$	$72.28^{\pm0.23}$	$42.05^{\pm0.27}$	$38.60^{\pm0.20}$	$40.56^{\pm0.24}$	$37.28^{\pm0.26}$
FedSAM	$79.74^{\pm0.18}$	$74.69^{\pm0.19}$	$79.87^{\pm0.18}$	$76.90^{\pm0.23}$	$44.78^{\pm0.25}$	$43.50^{\pm0.24}$	$44.14^{\pm0.29}$	$43.36^{\pm0.25}$
MoFedSAM [10]	$76.36^{\pm0.15}$	$65.74^{\pm0.19}$	$76.74^{\pm0.17}$	$70.74^{\pm0.21}$	$41.07^{\pm0.19}$	$34.11^{\pm0.23}$	$35.91^{\pm0.17}$	$28.55^{\pm0.27}$
FedGAMMA[12]	$80.89^{\pm0.17}$	$75.34^{\pm0.19}$	$81.73^{\pm0.16}$	$78.74^{\pm0.19}$	$49.78^{\pm0.25}$	$46.31^{\pm0.27}$	$47.91^{\pm0.26}$	$45.26^{\pm0.33}$
FedSMOO [13]	$\mathbf{84.17}^{\pm0.19}$	$\mathbf{80.92}^{\pm0.17}$	$\mathbf{84.78}^{\pm0.19}$	$\mathbf{82.79}^{\pm0.21}$	$\mathbf{52.31}^{\pm0.24}$	$\mathbf{49.42}^{\pm0.20}$	$50.59^{\pm0.21}$	$46.08^{\pm0.25}$
FedSpeed [11]	$82.76^{\pm0.19}$	$79.95^{\pm0.19}$	$83.36^{\pm0.18}$	$80.72^{\pm0.22}$	$49.93^{\pm0.23}$	$49.04^{\pm0.24}$	$\mathbf{50.61}^{\pm0.23}$	$\mathbf{46.85}^{\pm0.25}$
FedLESAM	$80.11^{\pm0.23}$	$74.35^{\pm0.22}$	$78.35^{\pm0.21}$	$71.23^{\pm0.25}$	$44.35^{\pm0.19}$	$43.75^{\pm0.21}$	$43.97^{\pm0.23}$	$43.21^{\pm0.22}$
FedLESAM-D	$83.26^{\pm0.19}$	$79.89^{\pm0.20}$	$83.99^{\pm0.23}$	$81.89^{\pm0.21}$	$49.77^{\pm0.20}$	$45.35^{\pm0.22}$	$50.58^{\pm0.19}$	$46.55^{\pm0.21}$
FedLESAM-S[14]	$83.76^{\pm0.17}$	$79.02^{\pm0.18}$	$83.12^{\pm0.20}$	$81.57^{\pm0.21}$	$49.52^{\pm0.19}$	$47.83^{\pm0.22}$	$48.21^{\pm0.23}$	$45.75^{\pm0.24}$
FedOMG[76]	$84.56^{\pm0.18}$	$81.24^{\pm0.17}$	$84.93^{\pm0.20}$	$83.05^{\pm0.20}$	$52.63^{\pm0.23}$	$49.70^{\pm0.19}$	$50.87^{\pm0.20}$	$46.32^{\pm0.23}$
FedGMT[77]	$84.82^{\pm0.17}$	$81.56^{\pm0.18}$	$85.10^{\pm0.21}$	$83.26^{\pm0.20}$	$52.84^{\pm0.22}$	$49.92^{\pm0.19}$	$51.09^{\pm0.20}$	$46.51^{\pm0.22}$
FedTOGA(ours)	$\mathbf{84.91}^{\pm0.15}$	$\mathbf{81.78}^{\pm0.17}$	$\mathbf{84.90}^{\pm0.19}$	$\mathbf{83.49}^{\pm0.14}$	$\mathbf{54.90}^{\pm0.16}$	$\mathbf{51.00}^{\pm0.15}$	$\mathbf{53.25}^{\pm0.17}$	$\mathbf{49.90}^{\pm0.21}$

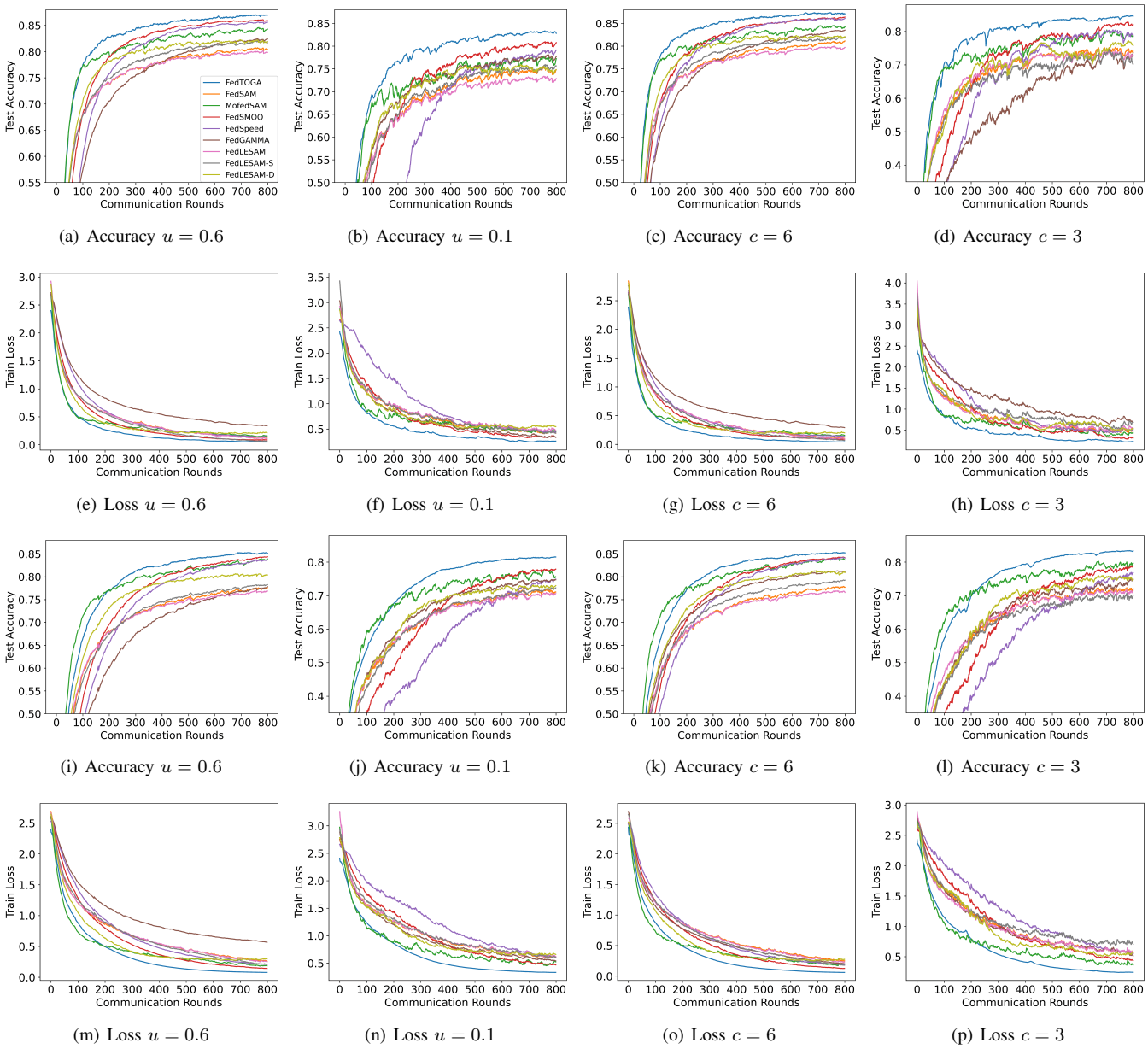
Note: All the experiments involved were set up in accordance with the fairness of the previous studies.

TABLE IV

DIRICHLET COEFFICIENTS u SELECTED FROM $\{0.1, 0.6\}$, AND c IS THE PATHOLOGICAL COEFFICIENT, I.E., THE NUMBER OF ACTIVE CATEGORIES IN EACH CLIENT. THE CIFAR10 HAS 100 CLIENTS IN THE LEFT PART WITH 10% ACTIVE IN EACH ROUND AND 200 CLIENTS IN THE RIGHT PART WITH 5% ACTIVE IN EACH ROUND.(RESNET18) **NOTE**: THE EXTENDED TABLE SEES TAB. X IN APPENDIX.

Method	CIFAR10							
Partition	Dirichlet		Pathological		Dirichlet		Pathological	
Coefficient	$u=0.6$	$u=0.1$	$c=6$	$c=3$	$u=0.6$	$u=0.1$	$c=6$	$c=3$
FedAvg [1]	$79.52^{\pm0.13}$	$76.00^{\pm0.18}$	$79.91^{\pm0.17}$	$74.08^{\pm0.22}$	$75.90^{\pm0.21}$	$72.93^{\pm0.19}$	$77.47^{\pm0.34}$	$71.68^{\pm0.34}$
FedAdam[33]	$77.08^{\pm0.31}$	$73.41^{\pm0.33}$	$77.05^{\pm0.26}$	$72.44^{\pm0.20}$	$75.55^{\pm0.38}$	$69.70^{\pm0.32}$	$75.74^{\pm0.22}$	$70.49^{\pm0.26}$
SCAFFOLD[20]	$81.81^{\pm0.17}$	$78.57^{\pm0.14}$	$83.07^{\pm0.10}$	$77.02^{\pm0.18}$	$79.00^{\pm0.26}$	$76.15^{\pm0.15}$	$80.69^{\pm0.21}$	$74.05^{\pm0.31}$
FedCM[39]	$82.97^{\pm0.21}$	$77.82^{\pm0.16}$	$83.44^{\pm0.17}$	$77.82^{\pm0.19}$	$80.52^{\pm0.29}$	$77.28^{\pm0.22}$	$81.76^{\pm0.24}$	$76.72^{\pm0.25}$
FedDyn[17]	$83.22^{\pm0.18}$	$78.08^{\pm0.19}$	$83.18^{\pm0.17}$	$77.63^{\pm0.14}$	$80.69^{\pm0.23}$	$76.82^{\pm0.17}$	$82.21^{\pm0.18}$	$74.93^{\pm0.22}$
FedSAM	$81.46^{\pm0.12}$	$77.03^{\pm0.17}$	$81.13^{\pm0.23}$	$78.30^{\pm0.24}$	$78.32^{\pm0.16}$	$74.00^{\pm0.14}$	$78.75^{\pm0.27}$	$75.12^{\pm0.29}$
MoFedSAM [10]	$85.29^{\pm0.13}$	$80.25^{\pm0.17}$	$84.74^{\pm0.16}$	$83.09^{\pm0.24}$	$84.76^{\pm0.20}$	$\mathbf{80.10}^{\pm0.14}$	$85.00^{\pm0.27}$	$\mathbf{82.13}^{\pm0.23}$
FedGAMMA[12]	$82.82^{\pm0.16}$	$79.91^{\pm0.15}$	$83.51^{\pm0.18}$	$77.11^{\pm0.14}$	$80.72^{\pm0.19}$	$76.70^{\pm0.14}$	$81.81^{\pm0.27}$	$77.44^{\pm0.29}$
FedSMOO [13]	$\mathbf{86.08}^{\pm0.14}$	$\mathbf{81.80}^{\pm0.18}$	$\mathbf{86.38}^{\pm0.15}$	$\mathbf{82.79}^{\pm0.16}$	$\mathbf{84.96}^{\pm0.19}$	$79.76^{\pm0.19}$	$84.82^{\pm0.18}$	$81.01^{\pm0.19}$
FedSpeed [11]	$86.01^{\pm0.16}$	$81.02^{\pm0.16}$	$86.09^{\pm0.19}$	$82.50^{\pm0.16}$	$84.12^{\pm0.18}$	$76.74^{\pm0.14}$	$84.78^{\pm0.27}$	$79.09^{\pm0.29}$
FedLESAM	$81.04^{\pm0.19}$	$76.92^{\pm0.16}$	$81.37^{\pm0.17}$	$78.21^{\pm0.21}$	$77.80^{\pm0.18}$	$73.73^{\pm0.22}$	$78.44^{\pm0.20}$	$74.53^{\pm0.19}$
FedLESAM-D	$84.27^{\pm0.14}$	$80.08^{\pm0.19}$	$85.62^{\pm0.18}$	$83.00^{\pm0.22}$	$82.53^{\pm0.19}$	$79.56^{\pm0.27}$	$\mathbf{85.04}^{\pm0.21}$	$81.10^{\pm0.19}$
FedLESAM-S[14]	$84.94^{\pm0.12}$	$79.52^{\pm0.17}$	$85.88^{\pm0.19}$	$82.18^{\pm0.15}$	$83.22^{\pm0.22}$	$78.69^{\pm0.17}$	$85.02^{\pm0.24}$	$80.57^{\pm0.17}$
FedTOGA(ours)	$\mathbf{86.99}^{\pm0.13}$	$\mathbf{83.16}^{\pm0.17}$	$\mathbf{87.21}^{\pm0.18}$	$\mathbf{84.55}^{\pm0.15}$	$\mathbf{85.21}^{\pm0.17}$	$\mathbf{81.60}^{\pm0.16}$	$\mathbf{85.24}^{\pm0.19}$	$\mathbf{83.25}^{\pm0.20}$

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2025.3638762



(a) Accuracy $u = 0.6$ (b) Accuracy $u = 0.1$ (c) Accuracy $c = 6$ (d) Accuracy $c = 3$

(e) Loss $u = 0.6$ (f) Loss $u = 0.1$ (g) Loss $c = 6$ (h) Loss $c = 3$

(i) Accuracy $u = 0.6$ (j) Accuracy $u = 0.1$ (k) Accuracy $c = 6$ (l) Accuracy $c = 3$

(m) Loss $u = 0.6$ (n) Loss $u = 0.1$ (o) Loss $c = 6$ (p) Loss $c = 3$

Fig. 2. Accyracy/ Loss on the CIFAR-10 dataset under 10% /5% participation of total 100/200 clients

recovers the local training problems of FedPD and FedDyn. When the value of $1/\alpha$ is set to zero, the local training problem of FedAvg is restored. These terms revisit the core challenge of heterogeneous FL: local consensus inconsistency. In addition to the quadratic proximal term introduced by FedProx, FedDyn and FedPD employ dual variables, which are beneficial in guiding local model updates, as discussed in Sec.III-B. However, focusing solely on local stationary points is insufficient due to the inability of clients in real FL to guaranty convergence to local stationary points after each training. Therefore, we further introduce global stationary conditions to enhance local consensus. The advantage of this approach is that clients are not burdened with additional storage or computational overhead, while the extra uplink overhead is reduced, alleviating the communication bottleneck.

VI. Experiments

A. Experimental Setups

Baselines We compare FedTOGA with the vanilla baseline FedAvg[1] and existing SAM-based FL methods, including FedSAM, MoFedSAM[10], FedGAMMA[12], FedSMOO[13], and FedSpeed[11], alongside the recent study FedLESAM[14]. Also, we compare with the momentum-based FL algorithms; for example, FedAdam, FedYogi[33], FedACG[36], and FedCM[39]. In addition, methods based on local consistency are also considered, including FedDyn[17], SCAFFOLD[20], FedDC[23], and FedRCL [80].

Dataset and Splits We use the benchmark datasets CIFAR-10 and CIFAR-100 in our experiments. Following the settings described in [12], [81], [14], client data are partitioned using Dirichlet and pathological splits under non-IID settings. To

emulate real-world conditions involving a large number of clients, we consider two experimental configurations. First, data are distributed to 100 clients, with 10% participating in each training round. Second, data are distributed to 200 clients, with 5% participation per round. Additionally, FedTOGA exhibits strong generalization performance on the OfficeHome, DomainNet, TinyImageNet, and Shakespeare datasets. Additional experimental results are provided in Appendix G.

Experimental Details To ensure a fair comparison, we adopt the experimental setup used in [13], [14]. SGD is used as the optimizer, with the client learning rate $\eta_l = 0.1$ and the global learning rate set to 1. The weight decay is fixed at 10^{-3}. To further assess the generalization performance of our method, we conduct experiments using two models: LeNet and ResNet-18 [82]. For LeNet, the learning rate decays with a multiplicative factor of 0.997 per epoch, while for ResNet18, it decays with a factor of 0.998. For CIFAR-10, the batch size is 50 and the number of local epochs is 5, whereas for CIFAR-100, the batch size is 20 and the number of local epochs is 2. In FedTOGA, the local perturbation correction coefficient is $\kappa = 1$, the dual-variable correction coefficient is $\beta = 0.8$, and the penalty coefficient is $\alpha = 0.1$. Following previous works, the perturbation magnitude ρ is set to 0.1, except for FedSAM and FedLESAM, where it is 0.01. Detailed information on the experimental setup is provided in Appendix E.

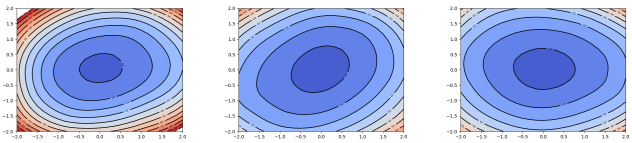
Evaluation metrics Test accuracy serves as the primary metric for fair comparison of FL algorithms across IID/non-IID data, client participation, and system configurations. The standard deviation captures performance variance from stochastic factors, supplemented by training loss and convergence curves.

B. Performance Evaluation

Performance compared with benchmarks As shown in Tables III and IV, the proposed FedTOGA algorithm performs excellently on various heterogeneous datasets concerning convergence speed and final achieved accuracy. Table III, which details the test accuracy of the LeNet model, demonstrates that FedTOGA significantly outperforms other algorithms under different heterogeneous data conditions. Specifically, under the Dirichlet-0.1 setting on the CIFAR10 dataset, FedTOGA attains an accuracy of 82.05%, marking a significant improvement of over 7.37% compared to vanilla FedAvg and a 1.99% increase over the second-highest baseline accuracy. Similar results are observed in Table IV for the ResNet18 model; FedTOGA outperforms all current baselines.

As seen in Table V, FedTOGA exhibits a significant advantage in convergence speed. When reaching 80% accuracy, FedTOGA converges 3.6× faster than FedSAM and 1.2× faster than the second-best. Similarly, when reaching 82% accuracy, FedTOGA converges 4.7× faster compared to FedSAM and 1.5× faster than the second-best. This indicates that FedTOGA achieves the target accuracy with significantly reduced computation and communication overhead compared to other methods. The exceptional global consensus capability of FedTOGA enables it to more effectively mitigate the impact of data heterogeneity and facilitate faster convergence.

Loss Surface As shown in Fig.3, we conduct a visualized loss landscape on CIFAR10 with the Dirichlet $u = 0.6$ setup. We compared FedSAM and FedSMOO, and the results show that the flatness of FedTOGA far exceeds that of FedSAM and is equal to that of FedSMOO, which means that it still performs well without increasing computation.



(a) FedSAM (b) FedSMOO (c) FedTOGA

Fig. 3. Visualization of global loss surface on CIFAR10 of FedSAM, FedSMOO, FedTOGA.

Impact of heterogeneity We use the Dirichlet and Pathological methods for data partitioning. For the Dirichlet distribution, we adopt variance coefficients u of 0.1 and 0.6. We use coefficients c of 3 and 6 for the Pathological distribution. As shown in Tables III and IV, increased data heterogeneity leads to decreased accuracy across all algorithms. However, FedTOGA exhibits the smallest accuracy drop. Specifically, for the Resnet18 model, as u changes from 0.6 to 0.1 under the Dirichlet distribution on the CIFAR10 dataset, FedSAM's accuracy decreases from 81.46% to 77.03%, a 4.43% reduction. At the same time, the second-best algorithm, FedSMOO, shows a drop from 86.08% to 81.80%, a 4.28% reduction. In contrast, FedTOGA's accuracy declines from 86.99% to 83.16%, a 3.83% drop. Similar trends are observed under the Pathological split, underscoring FedTOGA's superior stability and accuracy across varying levels of data heterogeneity.

Impact of partial participation We fix all hyperparameters except the client participation rate to assess its effect on accuracy. As illustrated in Table IV, a reduction in the client participation rate from 10% to 5% results in a modest decline in accuracy across all algorithms. For instance, on the CIFAR10 dataset, under the challenging pathological distribution with $c = 3$, FedTOGA's accuracy decreases marginally from 84.55% to 83.35%, a reduction of just 1.40%, while FedSMOO experiences a sharper decline from 82.79% to 81.01%, a reduction of 1.78%. Similarly, under the Dirichlet distribution with $u = 0.1$, FedTOGA's accuracy decreases from 86.99% to 85.21%, a decrease of 1.78%, whereas FedSMOO's accuracy drops from 86.08% to 84.96%, a reduction of 1.12%. Despite these reductions, FedTOGA consistently outperforms the accuracy of other algorithms', highlighting its generalization capability and stability.

As shown in Fig.2, FedTOGA significantly outperforms other algorithms in scenarios with significant heterogeneity (e.g., Dirichlet-0.1 and Pathological-3). Our algorithm still shows stability even when the number of clients decreases(e.g., 200 clients with 5% participation). These results are in line with our expectations. We aim to design an algorithm that enhances global consistency while efficiently finding a global flat minimum to improve generalization and reduce edge node computation and storage requirements.

Hyperparameter Sensitivity We study the sensitivity of the hyperparameters: learning rate decay, penalty coefficient α, correction coefficients β and κ, and perturbation coefficient ρ. As shown in Fig.4, our experiments demonstrate FedTOGA's

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2025.3638762

TABLE V
NUMBER OF COMMUNICATION ROUNDS TO ACHIEVE A TARGET ACCURACY. WE RECORDED THE FIRST ROUND OF COMMUNICATION TO REACH A TARGET ACCURACY. WE IMPROVED THE NUMBER OF TRAINING ROUNDS COMPARED TO THE OTHER ALGORITHMS IN THE DIRICHLET-0.1/0.6 AND PATHOLOGICAL-6.0/3.0 SETTINGS. WE MAINLY COMPARED THE SAM-BASED FL ALGORITHMS.

Partition	Dirichlet								Pathological							
Coefficient	$u = 0.6$				$u = 0.1$				$c = 6$				$c = 3$			
Acc/Rounds	80%	cost	82%	cost	76%	cost	78%	cost	80%	cost	82%	cost	76%	cost	78%	cost
FedSAM	481	3.6×	800+	4.7×	587	3.2×	800+	3.5×	443	3.3×	790	4.8×	465	2.9×	691	3.5×
MoFedSAM [10]	167	1.2×	270	1.6×	303	1.6×	425	2.9×	135	1.0×	253	1.5×	167	1.1×	265	1.3×
FedGAMMA[12]	458	3.4×	630	3.7×	369	2.0×	591	2.6×	407	3.0×	550	3.3×	701	4.4×	800+	4.0×
FedSMOO [13]	190	1.4×	253	1.5×	302	1.6×	402	1.8×	205	1.5×	263	1.6×	262	1.7×	322	1.6×
FedSpeed [11]	262	1.9×	318	1.9×	445	2.4×	530	2.3×	233	1.7×	292	1.8×	349	2.2×	438	2.2×
FedLESAM	588	4.4×	800+	4.7×	800+	4.3×	800+	3.5×	620	4.6×	800+	4.8×	497	3.1×	778	3.9×
FedLESAM-D	248	1.8×	418	2.5×	369	2.0×	663	2.9×	224	1.7×	376	2.3×	393	2.5×	452	2.3×
FedLESAM-S[14]	390	2.9×	643	3.8×	529	2.8×	800+	3.5×	348	2.6×	602	3.6×	497	3.1×	800+	4.0×
FedTOGA(ours)	135	1.0×	170	1.0×	184	1.0×	226	1.0×	134	1.0×	166	1.0×	158	1.0×	200	1.0×

Note: The SGD method is not considered.

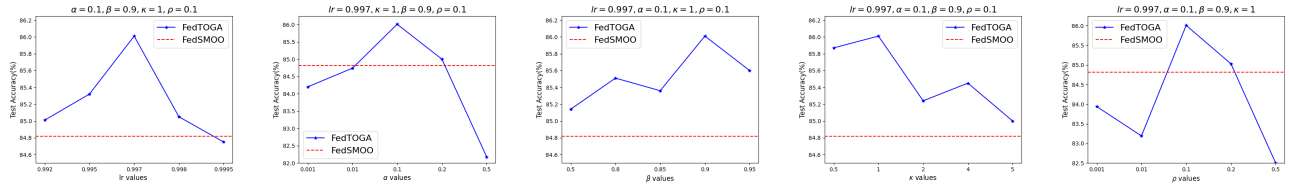


Fig. 4. Hyperparameters sensitivity studies of lr decay, penalized coefficient α, Correction coeficient β, κ and perturbations coefficient ρ on CIFAR-10.

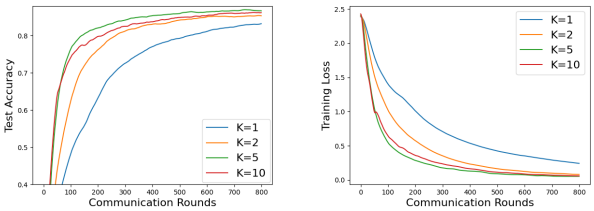


Fig. 5. Test accuracy and training loss of FedTOGA with different intervals K on CIFAR10. K is set as $1, 2, 5, 10$, and other parameters are the same as mentioned above.

resilience to variations in these hyperparameters. By systematically adjusting each parameter while holding the others constant, FedTOGA remains remarkably stable under changes in Lr decay and the correction coefficients β and κ. Additionally, the penalty coefficient α and the perturbation coefficient ρ effectively maintain robustness when appropriately selected.

Local Interval K measures the communication interval, which refers to the number of local training steps. In Theorem 2, we observe that increasing K can help the global model achieve a higher convergence rate when T is large enough. However, although increasing K improves the convergence speed, it also amplifies the negative effects of local heterogeneity. Figure 5 shows the impact of different values of K. Some previous studies suggested making K large enough to approach the suboptimal value of the objective function. However, in most practical FL setups, K represents a trade-off between training convergence speed and local overfitting. In our experiments, when $K = 2$, the training convergence rate and generalization of FedTOGA improved compared to $K = 1$. When K increased to 5, the convergence rate was about 1.5 times faster than $K = 2$, achieving the best accuracy, which aligns with our theoretical analysis. As K increases, when $K = 10$, the acceleration effect remains but becomes less significant, while generalization performance starts to decline. We believe that a larger K means more local updates, which forces local clients to move toward their local optima, interfering with generalization.

As the communication intervals increase, the model accuracy does not significantly decline, which demonstrates FedTOGA's robustness in long-interval communication scenarios and highlights the importance of enhancing global generalization consistency.

C. Training Speed

According to the above table V in the main test, we can see that FedTOGA performs far better than the other algorithms. It has the fastest convergence rate while maintaining high accuracy. The SAM optimizer usually slows down the whole training process due to the need to compute additional perturbations to the ascent process, which will be improved by enhancing consistency. MoFedSAM enforces consistency by employing global momentum on each local client and weighting it by a factor α (usually 0.1), which means that local knowledge will be forcibly overwritten by the global gradient while speeding up convergence in the early stages. However, it may not be able to draw further adequate learning progress from the locals in the later stage. FedTOGA corrects local perturbations and dynamic regularizers by guiding global updates, greatly enhancing the consistency of generalization and optimization. Therefore, our method can effectively accelerate the modeling speed and improve the modeling accuracy, especially in the case of large-scale heterogeneity. Table VI shows that FedTOGA has a similar local computation time to the SAM-based FL algorithm.

TABLE VI
WALL CLOCK TIME(TRAINING, LOADING, EVALUATION) ON CIFAR10 RESNET18 $u = 0.6, 0.1$ AND 100 CLIENTS.

	FedSAM	MoFedSAM	FedGAMMA	FedSpeed	FedSMOO	FedLESAM-D	**FedTOGA**
time	25.71s	28.73s	29.88s	28.98s	29.67s	25.70s	29.12s

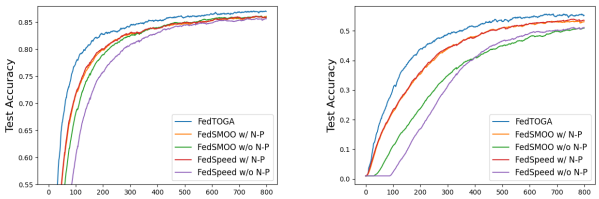
TABLE VII
ABLATION STUDIES OF DIFFERENT MODULES.

SAM	Dynamic Regularization	Dual Correction (β)	SAM Correction (κ)	CIFAR10 Acc	CIFAR100 Acc
✓	-	-	-	81.39%	48.08%
✓	✓	-	-	84.14%	53.79%
✓	✓	✓	-	85.54%	56.85%
✓	✓	✓	✓	86.01%	57.25%

D. Ablation Studies

We evaluated the contribution of individual components, namely *SAM*, *Dynamic Regularization*, *Dual Variable Correction*, and *SAM Perturbation Correction*, using the LeNet network on CIFAR10/100 datasets partitioned with a Dirichlet 0.6 distribution. FedSAM is used as the baseline for comparison. With the sequential addition of these components, the accuracy on CIFAR-10 increased by 2.75%, 4.15%, and 4.62%, while that on CIFAR-100 improved by 5.71%, 8.77%, and 9.17%, relative to FedSAM.

E. Neighbourhood Perturbation Analysis



(a) CIFAR10 (b) CIFAR100

Fig. 6. The Impact of FedSpeed and FedSMOO with neighborhood perturbation (N-P) Modules on CIFAR10/100, Dirichlet 0.6, 100 Client

As shown in Fig. 6, we evaluated the performance of SAM-based federated learning algorithms with the proposed **Neighborhood Perturbation (NP)** mechanism enabled. Enabling NP was found to effectively enhance the performance of FedSpeed [11] and FedSMOO [13]. This result validates our hypothesis, as discussed in Section IV-B.

VII. CONCLUSION

In this paper, we propose a novel federated learning algorithm, FedTOGA. For the first time, it estimates global perturbations by aggregating global training gradients and enhances the local dynamic regularization mechanism. This design ensures that local clients can align their updates with the global optimization and generalization objectives. FedTOGA enables the efficient discovery of globally consistent flat minima and accelerates convergence while avoiding additional local storage and uplink communication overhead. In summary, FedTOGA introduces global and neighborhood perturbation estimation to reduce communication costs and improve generalization performance. Our theoretical analysis establishes an $O(1/T)$ convergence rate, and experiments across multiple benchmark datasets demonstrate superior accuracy and efficiency under heterogeneous settings.

ACKNOWLEDGMENT

This work was supported in part by the Major Science and Technology Projects of Longmen Laboratory (Grant Nos. 231100220400, 231100220300), in part by the Liaoning Natural Science Foundation under Grant No. 2025-BS-0212, and in part by the Dalian Science and Technology Innovation Fund under Grant No. 2024JB11GX001.

REFERENCES

[1] B. McMahan, E. Moore, and et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.
[2] Z. Fan, Y. Wang, and et al., "Fedskip: Combatting statistical heterogeneity with federated skip aggregation," in *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2022, pp. 131–140.
[3] Z. Fan, J. Yao, R. Zhang, L. Lyu, Y. Zhang, and Y. Wang, "Federated learning under partially class-disjoint data via manifold reshaping," 2024. [Online]. Available: https://arxiv.org/abs/2405.18983
[4] B. Woodworth, K. K. Patel, S. Stich, and et al., "Is local sgd better than minibatch sgd?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 334–10 343.
[5] Speedtest, "speedtest.net," https://www.speedtest.net/global-index, 2024.
[6] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021.
[7] J. Wang, Q. Liu, and et al., "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.
[8] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," 2019.
[9] H. Li, Z. Xu, and et al., "Visualizing the loss landscape of neural nets," 2018. [Online]. Available: https://arxiv.org/abs/1712.09913
[10] Z. Qu, X. Li, R. Duan, Y. Liu, B. Tang, and Z. Lu, "Generalized federated learning via sharpness aware minimization," in *International conference on machine learning*. PMLR, 2022, pp. 18 250–18 280.
[11] Y. Sun and et al., "Fedspeed: Larger local interval, less communication round, and higher generalization accuracy," in *ICLR*, 2023.
[12] R. Dai, X. Yang, and et al., "Fedgamma: Federated learning with global sharpness-aware minimization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[13] Y. Sun, L. Shen, S. Chen, L. Ding, and D. Tao, "Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape," in *International Conference on Machine Learning*. PMLR, 2023, pp. 32 991–33 013.
[14] Z. Fan, S. Hu, J. Yao, G. Niu, Y. Zhang, M. Sugiyama, and Y. Wang, "Locally estimated global perturbations are better than local perturbations for federated sharpness-aware minimization," in *International Conference on Machine Learning*, 2024.
[15] G. Malinovskiy and et al., "From local sgd to local fixed-point methods for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6692–6701.

[16] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," 2020.

[17] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *International Conference on Learning Representations*, 2021.

[18] S. U. Stich, "Local sgd converges fast and communicates little," *arXiv preprint arXiv:1805.09767*, 2018.

[19] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[20] S. P. Karimireddy and et al., "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.

[21] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. Reddi, S. U. Stich, and A. T. Suresh, "Breaking the centralized barrier for cross-device federated learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28 663–28 676, 2021.

[22] M. Al-Shedivat, J. Gillenwater, E. P. Xing, and A. Rostamizadeh, "Federated learning via posterior averaging: A new perspective and practical algorithms," in *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[23] L. Gao, H. Fu, L. Li, Y. Chen, M. Xu, and C.-Z. Xu, "Feddc: Federated learning with non-iid data via local drift decoupling and correction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 112–10 121.

[24] G. Lee, M. Jeong, and et al., "Preservation of the global knowledge by not-true distillation in federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 461–38 474, 2022.

[25] G. Lee, M. Jeong, S. Kim, J. Oh, and S.-Y. Yun, "Fedsol: Stabilized orthogonal learning with proximal restrictions in federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 512–12 522.

[26] N. Shoham, T. Avidor, A. Keren, N. Israel, D. Benditkis, L. Mor-Yosef, and I. Zeitak, "Overcoming forgetting in federated learning on non-iid data," 2019. [Online]. Available: https://arxiv.org/abs/1910.07796

[27] Z. Wang, E. Yang, L. Shen, and H. Huang, "A comprehensive survey of forgetting in deep learning beyond continual learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[28] W. Huang, M. Ye, Z. Shi, and B. Du, "Generalizable heterogeneous federated cross-correlation and instance similarity learning," *IEEE TPAMI*, vol. 46, no. 2, pp. 712–728, 2023.

[29] S.-A. Rebuffi and et al., "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[30] J. Dong, L. Wang, and et al., "Federated class-incremental learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 164–10 173.

[31] J. Sun, X. Wu, H. Huang, and A. Zhang, "On the role of server momentum in federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 13, 2024, pp. 15 164–15 172.

[32] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar, "Adaptive methods for nonconvex optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[33] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," 2021.

[34] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[35] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural networks*, vol. 12, no. 1, pp. 145–151, 1999.

[36] G. Kim, J. Kim, and B. Han, "Communication-efficient federated learning with accelerated client gradient," in *Proceedings of the IEEE/CVF CVPR*, 2024, pp. 12 385–12 394.

[37] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence o (1/k2)," in *Dokl. Akad. Nauk. SSSR*, vol. 269, no. 3, 1983, p. 543.

[38] D. Jhunjhunwala, S. Wang, and G. Joshi, "Fedexp: Speeding up federated averaging via extrapolation," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[39] J. Xu, S. Wang, L. Wang, and A. C.-C. Yao, "Fedcm: Federated learning with client-level momentum," *arXiv preprint arXiv:2106.10874*, 2021.

[40] Y. Sun, L. Shen, H. Sun, L. Ding, and D. Tao, "Efficient federated learning via local adaptive amended optimizer with linear speedup," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 453–14 464, 2023.

[41] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning: A meta-learning approach," *arXiv:2002.07948*, 2020.

[42] C. T Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," *Advances in neural information processing systems*, vol. 33, pp. 21 394–21 405, 2020.

[43] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International conference on machine learning*. PMLR, 2021, pp. 2089–2099.

[44] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *International conference on machine learning*. PMLR, 2021, pp. 6357–6368.

[45] Y. Huang, L. Chu, and et al., "Personalized cross-silo federated learning on non-iid data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 9, 2021, pp. 7865–7873.

[46] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," in *International conference on machine learning*. PMLR, 2021, pp. 9489–9502.

[47] Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, and J. Jiang, "Federated learning from pre-trained models: A contrastive learning approach," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[48] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, and H. Guan, "Fedala: Adaptive local aggregation for personalized federated learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 237–11 244.

[49] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," *Advances in neural information processing systems*, vol. 33, pp. 19 586–19 597, 2020.

[50] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 8, pp. 3710–3722, 2020.

[51] Y. Ruan and C. Joe-Wong, "Fedsoft: Soft clustered federated learning with proximal local updating," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 7, 2022, pp. 8124–8131.

[52] M. Duan, D. Liu, X. Ji, R. Liu, L. Liang, X. Chen, and Y. Tan, "Fedgroup: Efficient clustered federated learning via decomposed data-driven measure," *arXiv preprint arXiv:2010.06870*, 2020.

[53] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," *arXiv preprint arXiv:2002.06440*, 2020.

[54] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in neural information processing systems*, vol. 33, pp. 2351–2363, 2020.

[55] L. Chen, W. Zhang, and et al., "Feddrl: Trustworthy federated learning model fusion method based on staged reinforcement learning," *Computing and Informatics*, 2024.

[56] X. Zhu, G. Li, and W. Hu, "Heterogeneous federated knowledge graph embedding learning and unlearning," in *Proceedings of the ACM web conference 2023*, 2023, pp. 2444–2454.

[57] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 713–10 722.

[58] S. Gupta, V. Sutar, and et al., "Fedalign: Federated domain generalization with cross-client feature alignment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1801–1810.

[59] D. Li and J. Wang, "Fedmd: Heterogenous federated learning via model distillation." *arXiv preprint arXiv:1910.03581*, 2019.

[60] D. Yao, W. Pan, and et al., "Fedgkd: Toward heterogeneous federated learning via global knowledge distillation," *IEEE Transactions on Computers*, vol. 73, no. 1, pp. 3–17, 2023.

[61] L. Zhang, L. Shen, L. Ding, D. Tao, and L.-Y. Duan, "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 174–10 183.

[62] W. Lu, X. Hu, J. Wang, and X. Xie, "Fedclip: Fast generalization and personalization for clip in federated learning," *arXiv preprint arXiv:2302.13485*, 2023.

[63] H. Zhao, W. Du, F. Li, P. Li, and G. Liu, "Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[64] C. Song, D. Saxena, J. Cao, and Y. Zhao, "Feddistill: Global model distillation for local model de-biasing in non-iid federated learning," *arXiv preprint arXiv:2404.09210*, 2024.

[65] C. Xie, S. Koyejo, and I. Gupta, "Asynchronous federated optimization," *arXiv preprint arXiv:1903.03934*, 2019.

[66] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, and D. Huba, "Federated learning with buffered asynchronous aggregation," in *International conference on artificial intelligence and statistics*. PMLR, 2022, pp. 3581–3607.

[67] Q. Ma, Y. Xu, H. Xu, Z. Jiang, L. Huang, and H. Huang, "Fedsa: A semi-asynchronous federated learning mechanism in heterogeneous edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3654–3672, 2021.

[68] D. Wu, R. Ullah, and et al., "Fedadapt: Adaptive offloading for iot devices in federated learning," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 20 889–20 901, 2022.

[69] S. Hochreiter and J. Schmidhuber, "Simplifying neural nets by discovering flat minima," *Advances in neural information processing systems*, vol. 7, 1994.

[70] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp minima can generalize for deep nets," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1019–1028.

[71] N. S. Keskar, D. Mudigere, and et al., "On large-batch training for deep learning: Generalization gap and sharp minima," 2016.

[72] M. Becker, F. Altrock, and B. Risse, "Momentum-sam: Sharpness aware minimization without computational overhead," 2024.

[73] B. Li and G. Giannakis, "Enhancing sharpness-aware optimization through variance suppression," *Advances in Neural Information Processing Systems*, vol. 36, pp. 70 861–70 879, 2023.

[74] M. Mueller, T. Vlaar, D. Rolnick, and M. Hein, "Normalization layers are all that sharpness-aware minimization needs," *Advances in Neural Information Processing Systems*, vol. 36, pp. 69 228–69 252, 2023.

[75] D. Caldarola, B. Caputo, and M. Ciccone, "Improving generalization in federated learning by seeking flat minima," in *European Conference on Computer Vision*. Springer, 2022, pp. 654–672.

[76] T.-B. Nguyen, M.-D. Nguyen, J. Park, Q.-V. Pham, and W. J. Hwang, "Federated domain generalization with data-free on-server gradient matching," in *The Thirteenth International Conference on Learning Representations*, May 2025. [Online]. Available: https://openreview.net/forum?id=8TERgu1Lb2

[77] Y. Li, T. Liu, and et al., "One arrow, two hawks: Sharpness-aware minimization for federated learning via global model trajectory," in *International Conference on Machine Learning (ICML)*, 2025.

[78] H. Wang, S. Marella, and J. Anderson, "Fedadmm: A federated primal-dual algorithm allowing partial participation," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 287–294.

[79] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, "Lookahead optimizer: k steps forward, 1 step back," *Advances in neural information processing systems*, vol. 32, 2019.

[80] S. Seo, J. Kim, G. Kim, and B. Han, "Relaxed contrastive learning for federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 279–12 288.

[81] Y. Sun, L. Shen, and D. Tao, "Understanding how consistency works in federated learning via stage-wise relaxed initialization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[83] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

APPENDIX

A. Theorem 1's Proof

Proof. Calculated by SAM rules, $\tilde{\theta}^t_{i,k-1} = \theta^t_{i,k-1} + \rho\frac{\nabla f_i(\theta^t_{i,k-1})}{\|\nabla f_i(\theta^t_{i,k-1})\|}$. For the induction, we assume that $\|\theta^t_{i,k-1} - v^t_{k-1}\|^2 \le H_i(k-1)$, then

$$
\begin{aligned}
&\|\theta^t_{i,k} - v^t_k\|^2 \\
=&\|\theta^t_{i,k-1} - \eta_l\nabla f_i(\tilde{\theta}^t_{i,k-1}) - (v^t_{k-1} - \eta_l\nabla f_i(\tilde{v}^t_{k-1}))\|^2 \\
=&\left\|\theta^t_{i,k-1} - \eta_l\nabla f_i\left(\theta^t_{i,k-1} + \rho\frac{\nabla f_i(\theta^t_{i,k-1})}{\|\nabla f_i(\theta^t_{i,k-1})\|}\right)\right. \\
&\left. - \left(v^t_{i,k-1} - \eta_l\nabla f\left(v^t_{k-1} + \rho\frac{\nabla f(v^t_{k-1})}{\|\nabla f(v^t_{k-1})\|}\right)\right)\right\|^2 \\
=&\left\|\theta^t_{i,k-1} - \eta_l\nabla f_i\left(\theta^t_{i,k-1} + \rho\frac{\nabla f_i(\theta^t_{i,k-1})}{\|\nabla f_i(\theta^t_{i,k-1})\|}\right) - v^t_{i,k-1}\right. \\
&\left. +\eta_l\nabla f\left(v^t_{k-1} + \rho\frac{\nabla f(v^t_{k-1})}{\|\nabla f(v^t_{k-1})\|}\right)\right\|^2 \\
\overset{(1)}{\le}&\|\theta^t_{i,k-1} - v^t_{i,k-1}\|^2 + \eta_l^2L^2\left\|\theta^t_{i,k-1} - v^t_{i,k-1}\right. \\
&\left. +\rho\frac{\nabla f_i(\theta^t_{i,k-1})}{\|\nabla f_i(\theta^t_{i,k-1})\|} - \rho\frac{\nabla f(v^t_{k-1})}{\|\nabla f(v^t_{k-1})\|}\right\|^2 + \eta_l^2\sigma_g^2 \\
\overset{(2)}{\le}&(1+2\eta_l^2L^2)\|\theta^t_{i,k-1} - v^t_{i,k-1}\|^2 + \eta_l^2(L^2\rho^2\sigma_g^{'2} + \sigma_g^2) \\
\le&\frac{L^2\rho^2\sigma_g^{'2} + \sigma_g^2}{2L^2}(1+2\eta_l^2L^2)^k - (1+2\eta_l^2L^2)\frac{L^2\rho^2\sigma_g^{'2} + \sigma_g^2}{2L^2} \\
&+ \eta_l^2(L^2\rho^2\sigma_g^{'2} + \sigma_g^2) \\
=&\frac{L^2\rho^2\sigma_g^{'2} + \sigma_g^2}{2L^2}(1+2\eta_l^2L^2)^k - \frac{L^2\rho^2\sigma_g^{'2} + \sigma_g^2}{2L^2} \\
=&\frac{L^2\rho^2\sigma_g^{'2} + \sigma_g^2}{2L^2}((1+2\eta_l^2L^2)^k - 1).
\end{aligned}
$$

Equation (1-2) holds due to the assumption 1. Recursively, it follows that the theorem 1 holds. □

B. Theorem 2's Proof

Recalling the Algorithm 1, based on the FL paradigm, we propose an Augmented Lagrangian(AL) function:

$$
\begin{aligned}
f(\theta, h) &\triangleq \frac{1}{N}\sum_{i\in N} f_i(\theta, \theta_i, h_i); F_i(\theta, \theta_i, h_i) \\
&\triangleq f_i(\theta_i) + \langle h_i, \theta - \theta_i\rangle + \frac{1}{\alpha}\|\theta_i - \theta\|^2
\end{aligned} \tag{11}
$$

By fixing θ, the AL function can be separated into local pairs $\{\theta_i, h_i\}$. However, the local optimization cannot perceive the global gradient trend due to the CTA policy. Therefore, we direct the global update Δ to be merged into the local dual variables. Thus, the local AL function is rewritten as:

$$
F_i(\theta, \theta_i, h_i) \triangleq f_i(\theta_i) + \langle h_i - \beta\Delta, \theta - \theta_i\rangle + \frac{1}{\alpha}\|\theta_i - \theta\|^2 \tag{12}
$$

Unlike [78], [17], [13], inspired by [11], we relax the strict assumption that $\tilde{g}^t_i - h^t_i + \frac{1}{\alpha}(\theta^t_i - \theta^t) + \beta\Delta^t = 0$ as a strict assumption and extend the local interval to K rounds, so we obtain the workflow in Algorithm 1. First, we give all the lemmas needed for proof analysis.

Lemma A.1. *For $\forall\theta^t_{i,k} \in \mathbb{R}^d$ and i in S_t, we have $\psi^t_{i,k} = \theta^t_{i,k} - \theta^t_{i,k-1}$ with the fact $\psi^t_{i,0} = 0$, and $\Psi^t_{i,K} = \sum_{k=0}^{K-1}\psi^t_{i,k} = \theta^t_{i,K} - \theta^t_{i,0}$, under the workflow in Algorithm 1, we have:*

$$
\Psi^t_{i,K} = -\alpha\gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}^t_{i,k} + \gamma\alpha(h^t_i - \beta\Delta^t) \tag{13}
$$

where $\gamma = \sum_{k=0}^{K-1}\gamma_k = \sum_{k=0}^{K-1}\frac{\eta_l}{\alpha}\left(1 - \frac{\eta_l}{\alpha}\right)^{K-1-k} = 1 - \left(1 - \frac{\eta_l}{\alpha}\right)^K$.

Proof. According to Line 19 in Algorithm 1, we have

$$
\begin{aligned}
\psi^t_{i,k} &= \Psi^t_{i,k} - \Psi^t_{i,k-1} = \theta^t_{i,k} - \theta^t_{i,k-1} \\
&= -\eta_l\left(\tilde{g}^t_{i,k} - h^t_i + \frac{1}{\alpha}\left(\theta^t_{i,k} - \theta^t_{i,0}\right) + \beta\Delta^t\right) \\
&= -\eta_l\left(\tilde{g}^t_{i,k} - h^t_i + \frac{1}{\alpha}\Psi^t_{i,k-1} + \beta\Delta^t\right).
\end{aligned}
$$

Then we can build the $\Psi_{i,k}^t$ as:

$$\begin{aligned}\Psi_{i,k}^t &= \Psi_{i,k-1}^t - \eta_l\big(\tilde{g}_{i,k}^t - h_i^t + \frac{1}{\alpha}\Psi_{i,k-1}^t + \beta\Delta^t\big)\\ &= \Big(1-\frac{\eta_l}{\alpha}\Big)\Psi_{i,k-1}^t - \eta_l(\tilde{g}_{i,k}^t - h_i^t + \beta\Delta^t).\end{aligned}$$

Taking the iteration on k,

$$\begin{aligned}\theta_{i,K}^t - \theta_{i,0}^t &= \Psi_{i,K}^t = \Big(1-\frac{\eta_l}{\alpha}\Big)^K \Psi_{i,0}^t\\ &- \eta_l\sum_{k=0}^{K-1}\Big(1-\frac{\eta_l}{\alpha}\Big)^{K-1-k}(\tilde{g}_{i,k}^t - h_i^t + \beta\Delta^t)\\ &\overset{(1)}{=} -\eta_l\sum_{k=0}^{K-1}\Big(1-\frac{\eta_l}{\alpha}\Big)^{K-1-k}(\tilde{g}_{i,k}^t - h_i^t + \beta\Delta^t)\\ &= -\alpha\sum_{k=0}^{K-1}\frac{\eta_l}{\alpha}\Big(1-\frac{\eta_l}{\alpha}\Big)^{K-1-k}\tilde{g}_{i,k}^t\\ &+ \Big(1-(1-\frac{\eta_l}{\alpha})^K\alpha(h_i^t - \beta\Delta^t)\Big)\\ &= -\alpha\gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t + \gamma\alpha(h_i^t - \beta\Delta^t).\end{aligned}$$

(1) applies $\Psi_{i,k}^t = 0$. □

Lemma A.2. *Under the workflow in Algorithm 1, we have:*

$$h_i^{t+1} = (1-\gamma)h_i^t + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t) \tag{14}$$

Proof. According to the update rule of Line 21 in Algorithm 1, we have

$$\begin{aligned}h_i^{t+1} &= h_i^t - \frac{1}{\alpha}(\theta_{i,K}^t - \theta_{i,0}^t)\\ &\overset{(1)}{=} h_i^t - \frac{1}{\alpha}(-\alpha\gamma\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t + \gamma\alpha(h_i^t - \beta\Delta^t))\\ &= h_i^t + \frac{\eta_l}{\alpha}\sum_{k=0}^{K-1}\Big(1-\frac{\eta_l}{\alpha}\Big)^{K-1-k}(\tilde{g}_{i,k}^t - h_i^t + \beta\Delta^t)\\ &= (1-\gamma)h_i^t + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t).\end{aligned}$$

(1) holds due to Lemma A.1. □

Lemma A.3. *We define the $u^{t+1} = \frac{1}{N}\sum_{i\in N}\theta_{i,K}^t$ as the averaged model among the last iteration of clients at t, the auxiliary sequence $\{z^t = u^t + \frac{1-\gamma}{\gamma}(u^t - u^{t-1})\}_{t>0}$ satisfies the rule as:*

$$z^{t+1} = z^t - \alpha\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t - \alpha\beta\Delta^t \tag{15}$$

Proof. Firstly, recalling the lemma A.1 and $\theta_{i,0}^t = \theta^t = \frac{1}{N}\sum_{i\in N}(\theta_{i,K}^{t-1} - \alpha h_i^t)$ in Algorithm 1, we have:

$$\begin{aligned}u^{t+1} - u^t &= \frac{1}{N}\sum_{i\in N}(\theta_{i,K}^t - \theta_{i,K}^{t-1})\\ &= \frac{1}{N}\sum_{i\in N}(-\alpha\gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t + \gamma\alpha(h_i^t - \beta\Delta^t) - \alpha h_i^t)\\ &= -\alpha\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\gamma(\tilde{g}_{i,k}^t + \beta\Delta^t) + (1-\gamma)h_i^t).\end{aligned}$$

Here, we define a virtual observation sequence $\{u^t\}$, and its update rule is:

$$\begin{aligned}u_{i,k+1}^t &= u_{i,k}^t - \alpha\frac{\gamma_k}{\gamma}(\gamma(\tilde{g}_{i,k}^t + \beta\Delta^t) + (1-\gamma)h_i^t);\\ u_{i,0}^{t+1} &= u^{t+1} = \frac{1}{N}\sum_{i\in N}u_{i,K}^t.\end{aligned}$$

Recalling the lemma A.2 and update rule $u_{i,K}^t - u_{i,0}^t = -\alpha(1-\gamma)h_i^t - \alpha\gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t)$, we can get:

$$\begin{aligned}h_i^{t+1} &= (1-\gamma)h_i^t + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t)\\ &= -\frac{1}{\alpha}(u_{i,K}^t - u_{i,0}^t).\end{aligned}$$

Then, we can expand the auxiliary sequence z^t as:

$$\begin{aligned}z^{t+1} - z^t &= \frac{1}{\gamma}(u^{t+1} - u^t) - \frac{1-\gamma}{\gamma}(u^t - u^{t-1})\\ &= \Delta - \frac{1-\gamma}{\gamma}\frac{1}{N}\sum_{i\in N}(\theta_{i,K}^{t-1} - \theta_{i,K}^{t-2} + \alpha h_i^t)\\ &\overset{(1)}{=} \Delta - \frac{1-\gamma}{\gamma}\frac{1}{N}\sum_{i\in N}(\theta_{i,K}^{t-1} - \theta_{i,0}^{t-1} + \alpha h_i^t - \alpha h_i^{t-1})\\ &= -\alpha\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t - \alpha\beta\Delta^t,\end{aligned}$$

where $\Delta = -\alpha\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\Big(\tilde{g}_{i,k}^t + \beta\Delta^t\Big)$ and (1) hold due to Line 21 in Algorithm 1. □

Lemma A.4. *(Bounded global dual update)The global dual variable $\frac{1}{N}\sum_{i\in N}h_i^{t+1}$ holds an upper bound of:*

$$\begin{aligned}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2 &\le \frac{1}{\gamma}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2 - \mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)\\ &+ 2\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2 + 2\beta^2\mathbb{E}_t\|\Delta^t\|^2.\end{aligned} \tag{16}$$

Proof. According to lemma A.2, we have:

$$\frac{1}{N}\sum_{i\in N}h_i^{t+1} = (1-\gamma)\frac{1}{N}\sum_{i\in N}h_i^t + \gamma\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t).$$

Take $L2$-norm, we have $\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2$:

$$\begin{aligned}&= \left\|(1-\gamma)\frac{1}{N}\sum_{i\in N}h_i^t + \gamma\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t)\right\|^2\\ &\le (1-\gamma)\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2 + 2\gamma\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2 + 2\beta^2\gamma\|\Delta^t\|^2.\end{aligned}$$

Take expectations. Thus, we have the following recursion:

$$\begin{aligned}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2 &\le \frac{1}{\gamma}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2 - \mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)\\ &+ 2\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2 + 2\beta^2\mathbb{E}_t\|\Delta^t\|^2.\end{aligned} \tag{17}$$

□

Lemma A.5. *(Bounded local dual update)The local dual variable h_i^{t+1} holds an upper bound of $\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|h_i^t\|^2$:*

$$\begin{aligned}
&\leq \frac{C}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2 - \mathbb{E}_t\|h_i^{t+1}\|^2) + 4CL^2\rho^2 \\
&+ \frac{24CL^2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 \\
&+ (12+2\beta^2)C\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2 + 2C(6\sigma_g^2+\sigma_l^2). \qquad (18)
\end{aligned}$$

where $\frac{1}{C} = 1 - \frac{24\alpha^2L^2(1-2\gamma)^2}{\gamma^2}$ is the constant.

Proof. Recalling lemma A.2,

$$h_i^{t+1} = (1-\gamma)h_i^t + \gamma\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}(\tilde{g}_{i,k}^t + \beta\Delta^t).$$

same as lemma A.4's proof, we have:

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|h_i^t\|^2 \leq \frac{1}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2 - \mathbb{E}_t\|h_i^{t+1}\|^2) \\
&+ \frac{2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\tilde{g}_{i,k}^t\|^2 + 2\beta^2\mathbb{E}_t\|\Delta^t\|^2.
\end{aligned}$$

We provide an upper bound for the quasi-stochastic gradient:

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\tilde{g}_{i,k}^t\|^2 \\
&\leq 2L^2\rho^2 + \frac{2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\nabla f_i(\theta_{i,k}^t) - \nabla f_i(z^t)\right. \\
&\left.+\nabla f_i(z^t) - \nabla f(z^t) + \nabla f(z^t)\right\|^2 + \sigma_l^2 \\
&\leq 2L^2\rho^2 + \frac{12L^2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 \\
&+ 12L^2\left\|\theta^t - u^t + u^t - z^t\right\|^2 + 6\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2 + (6\sigma_g^2+\sigma_l^2) \\
&\overset{(1)}{\leq} 2L^2\rho^2 + \frac{12L^2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 + (6\sigma_g^2+\sigma_l^2) \\
&+ 12L^2\frac{1}{N}\sum_{i\in N}\left\|-\alpha h_i^t + \frac{1-\gamma}{\gamma}\alpha h_i^t\right\|^2 + 6\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2 \\
&\leq 2L^2\rho^2 + \frac{12L^2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 + (6\sigma_g^2+\sigma_l^2) \\
&+ \frac{12\alpha^2L^2(1-2\gamma)^2}{\gamma^2 N}\sum_{i\in N}\mathbb{E}_t\left\|h_i^t\right\|^2 + 6\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2. \qquad (19)
\end{aligned}$$

Inequality (1) holds because $u^t - z^t = -\frac{1-\gamma}{\gamma}(u^t - u^{t-1}); \theta^t - u^t = -\alpha\frac{1}{N}\sum_{i\in N}h_i^t$. Let $\frac{1}{C} = 1 - \frac{24\alpha^2L^2(1-2\gamma)^2}{\gamma^2}$ be the constant. Combining the above inequalities, we have $\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|h_i^t\|^2$:

$$\begin{aligned}
&\leq \frac{C}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2 - \mathbb{E}_t\|h_i^{t+1}\|^2) + 4CL^2\rho^2 \\
&+ \frac{24CL^2}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 \\
&+ (12+2\beta^2)C\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2 + 2C(6\sigma_g^2+\sigma_l^2).
\end{aligned}$$

We set $\Delta^t \approx \nabla f(z^t)$ like [39], [10], [14]. □

Now we have completed all the preparations for the proof of Theorem 2. For the non-convex case, based on assumptions 1-3, we take the conditional expectation at round $t+1$ and expand the $f(z^{t+1})$ as $\mathbb{E}_t f(z^{t+1})$:

$$\begin{aligned}
&\leq \mathbb{E}_t f(z^t) + \mathbb{E}_t\langle\nabla f(z^t), z^{t+1} - z^t\rangle + \frac{L}{2}\mathbb{E}_t\|z^{t+1} - z^t\|^2 \\
&= \mathbb{E}_t f(z^t) + \frac{L}{2}\mathbb{E}_t\|z^{t+1} - z^t\|^2 \\
&\quad + \mathbb{E}_t\left\langle\nabla f(z^t), -\alpha\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\left(\tilde{g}_{i,k}^t + \beta\nabla f(z^t)\right)\right\rangle \\
&= \mathbb{E}_t f(z^t) - \alpha(1+\beta)\|\nabla f(z^t)\|^2 + \frac{L}{2}\underbrace{\mathbb{E}_t\|z^{t+1} - z^t\|^2}_{\textbf{A.2}} \\
&\quad \underbrace{-\alpha\mathbb{E}_t\left\langle\nabla f(z^t), \frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t - \nabla f(z^t)\right\rangle}_{\textbf{A.1}}. \qquad (20)
\end{aligned}$$

Firstly, the term **A.1** can be bounded:

$$\begin{aligned}
&-\alpha\mathbb{E}_t\left\langle\nabla f(z^t), \frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t - \nabla f(z^t)\right\rangle \\
&\overset{(1)}{=} -\alpha\mathbb{E}_t\left\langle\nabla f(z^t), \frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t - \frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\nabla f_i(z^t)\right\rangle \\
&\overset{(2)}{=} \frac{\alpha}{2}\|\nabla f(z^t)\|^2 + \frac{\alpha}{2}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\left(\mathbb{E}\tilde{g}_{i,k}^t - \nabla f_i(z^t)\right)\right\|^2 \\
&\quad - \frac{\alpha}{2N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{g}_{i,k}^t\right\|^2 \\
&\overset{(3)}{\leq} \frac{\alpha}{2}\|\nabla f(z^t)\|^2 + \frac{\alpha}{2}\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\mathbb{E}\tilde{g}_{i,k}^t - \nabla f_i(z^t)\right\|^2 \\
&\quad - \frac{\alpha}{2N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{g}_{i,k}^t\right\|^2.
\end{aligned}$$

(1) holds due to the fact $\frac{1}{N}\sum_{i\in N}\nabla f_i(z^t) = \nabla f(z^t)$. (b) applies $-\langle x,y\rangle = \frac{1}{2}(\|x\|^2 + \|y\|^2 - \|x+y\|^2)$ (c) holds due to Jensen's inequality. And, according to the SAM update rule, we have $\mathbb{E}\tilde{g}_{i,k}^t = \nabla f(\theta_{i,k}^t + \delta_{i,k}^t)$. Then, we can bound the term $\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\mathbb{E}\tilde{g}_{i,k}^t - \nabla f_i(z^t)\right\|^2$ as follows:

$$\begin{aligned}
&\leq \frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\nabla f(\theta_{i,k}^t + \delta_{i,k}^t) - \nabla f_i(z^t)\right\|^2 \\
&= \frac{2L^2}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t + \theta^t - u^t + u^t - z^t\right\|^2 + 2L^2\rho^2 \\
&\leq \frac{4L^2}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 \\
&\quad + 4L^2\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta^t - u^t + u^t - z^t\right\|^2 + 2L^2\rho^2 \\
&\overset{(1)}{\leq} \frac{4L^2}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\theta_{i,k}^t - \theta^t\right\|^2 \\
&\quad + \frac{4\alpha^2L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2 - \mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)
\end{aligned}$$

$$+\frac{8\alpha^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2$$
$$+\frac{8\alpha^2L^2(1-2\gamma)^2}{\gamma^2}\beta^2\mathbb{E}_t\|\nabla f(z^t)\|^2+2L^2\rho^2. \tag{21}$$

(1) applied the lemma A.4.

Then, we assume $\epsilon^t=\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}_t\|\theta_{i,k}^t-\theta^t\|^2$ term as the local offset after k iterations. We first bounded $\epsilon_k^t=\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|\theta_{i,k}^t-\theta^t\|^2$ as:

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|\theta_{i,k}^t-\theta^t\|^2=\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|\theta_{i,k}^t-\theta_{i,k-1}^t+\theta_{i,k-1}^t-\theta_{i,0}^t\|^2\\
&\overset{(1)}{=}\frac{1}{N}\sum_{i\in N}\left\|-\eta_l(\tilde{g}_{i,k-1}^t-h_i^t)+\left(1-\frac{\eta_l}{\alpha}\right)\left(\theta_{i,k-1}^t-\theta_{i,0}^t\right)-\eta_l\beta\Delta^t\right\|^2\\
&\overset{(2)}{\leq}(1+b)\left(1-\frac{\eta_l}{\alpha}\right)^2\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|\theta_{i,k-1}^t-\theta_{i,0}^t\|^2\\
&+(1+\frac{1}{b})\frac{\eta_l^2}{N}\sum_{i\in N}\mathbb{E}_t\|\tilde{g}_{i,k-1}^t-h_i^t+\beta\Delta^t\|^2\\
&\leq(1+\frac{1}{b})\frac{3\eta_l^2}{N}\sum_{i\in N}\left(\mathbb{E}_t\|\nabla f_i(\theta_{i,k-1}^t+\delta_{i,k-1}^t)\|^2+\mathbb{E}_t\|h_i^t\|^2\right.\\
&\left.+\mathbb{E}_t\|\beta\Delta^t\|^2\right)+(1+\frac{1}{b})\eta_l^2\sigma_l^2+(1+b)\left(1-\frac{\eta_l}{\alpha}\right)^2\epsilon_{k-1}^t\\
&\leq(1+\frac{1}{b})\frac{6\eta_l^2}{N}\sum_{i\in N}\mathbb{E}_t\|\nabla f_i(\theta_{i,k-1}^t)-\nabla f_i(\theta^t)+\nabla f_i(\theta^t)\\
&-\nabla f_i(z^t)+\nabla f_i(z^t)-\nabla f(z^t)+\nabla f(z^t)\|^2\\
&+(1+\frac{1}{b})\frac{3\eta_l^2}{N}\sum_{i\in N}\mathbb{E}_t\|h_i^t\|^2+(1+\frac{1}{b})3\eta_l^2\beta^2\mathbb{E}_t\|\nabla f(z^t)\|^2\\
&+(1+\frac{1}{b})\eta_l^2(\sigma_l^2+6L^2\rho^2)+(1+b)\left(1-\frac{\eta_l}{\alpha}\right)^2\epsilon_{k-1}^t\\
&\overset{(3)}{\leq}\left((1+b)\left(1-\frac{\eta_l}{\alpha}\right)^2+(1+\frac{1}{b})24\eta_l^2L^2\right)\epsilon_{k-1}^t\\
&+(1+\frac{1}{b})\eta_l^2\left(\frac{24L^2\alpha^2(1-2\gamma)^2}{\gamma^2}+3\right)\frac{1}{N}\sum_{i\in N}\mathbb{E}_t\|h_i^t\|^2\\
&+(1+\frac{1}{b})3\eta_l^2(8+\beta^2)\mathbb{E}_t\|\nabla f(z^t)\|^2\\
&+(1+\frac{1}{b})\eta_l^2(\sigma_l^2+6L^2\rho^2+24\sigma_g^2)\\
&\overset{(4)}{\leq}\left((1+b)\left(1-\frac{\eta_l}{\alpha}\right)^2+(1+\frac{1}{b})24\eta_l^2L^2\right)\epsilon_{k-1}^t\\
&+(1+\frac{1}{b})\eta_l^2(\sigma_l^2+6L^2\rho^2+24\sigma_g^2)\\
&+(1+\frac{1}{b})\frac{7\eta_l^2}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2)\\
&+14(1+\frac{1}{b})\eta_l^2(\sigma_l^2+2L^2\rho^2+6\sigma_g^2)\\
&+168(1+\frac{1}{b})\eta_l^2L^2\epsilon^t+(1+\frac{1}{b})7\eta_l^2(12+2\beta^2)\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2\\
&+(1+\frac{1}{b})3\eta_l^2(8+\beta^2)\mathbb{E}_t\|\nabla f(z^t)\|^2.
\end{aligned}$$

(1) holds due to Line.19 in Algorithm 1, (2) uses the fact $\|x+y\|^2\leq(1+b)\|x\|^2+(1+\frac{1}{b})\|y\|^2$, (3) applies lemma A.5, (4) applies C satisfies $C\leq 2$, which means $\left(\frac{24L^2\alpha^2(1-2\gamma)^2}{\gamma^2}+3\right)=\frac{4C-1}{C}$, $\frac{1}{C}=1-\frac{24\alpha^2L^2(1-2\gamma)^2}{\gamma^2}\geq\frac{1}{2}$.

We let the weight satisfy that [11]:

$$\begin{aligned}
&(1+b)\left(1-\frac{\eta_l}{\alpha}\right)^2+(1+\frac{1}{b})24\eta_l^2L^2\\
&\leq\frac{\gamma_{K-2}}{\gamma_{K-1}}=\frac{\gamma_{K-3}}{\gamma_{K-2}}=\cdots=\frac{\gamma_1}{\gamma_0}=1-\frac{\eta_l}{\alpha}
\end{aligned}\tag{22}$$

let $\eta_l\leq\alpha$, we have $\epsilon^t=\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\epsilon_k^t$:

$$\begin{aligned}
&\leq 7(1+\frac{1}{b})\eta_l^2\sum_{k=0}^{K-1}\left(\sum_{\check{k}=0}^{K-1}\frac{\gamma_k}{\gamma}\right)\left(3\sigma_l^2+5L^2\rho^2+16\sigma_g^2+16\epsilon^t\right.\\
&\left.+(16+3\beta^2)\mathbb{E}_t\|\nabla f(z^t)\|^2+\frac{1}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2)\right)\\
&=7(1+\frac{1}{b})\eta_l^2K\left(3\sigma_l^2+5L^2\rho^2+16\sigma_g^2+(16+3\beta^2)\mathbb{E}_t\|\nabla f(z^t)\|^2\right.\\
&\left.+\frac{1}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2)\right)+168(1+\frac{1}{b})\eta_l^2KL^2\epsilon^t.
\end{aligned}\tag{23}$$

Let η_l satisfy the bound of $\eta_l\leq\frac{1}{\sqrt{336(1+1/b)KL}}$ for convenience, we can bound the ϵ^t as:

$$\begin{aligned}
\epsilon^t\leq 14(1+\frac{1}{b})\eta_l^2K&\left(3\sigma_l^2+5L^2\rho^2+16\sigma_g^2\right.\\
&\left.+(16+3\beta^2)\mathbb{E}_t\|\nabla f(z^t)\|^2+\frac{1}{\gamma N}\sum_{i\in N}(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2)\right).
\end{aligned}\tag{24}$$

Let $b=1$ for convenience, we can get:

$$\begin{aligned}
&\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\mathbb{E}\tilde{g}_{i,k}^t-\nabla f_i(z^t)\right\|^2\\
&\leq\frac{4\alpha^2L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2-\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)\\
&+\frac{8\alpha^2L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2+2L^2\rho^2\\
&+\frac{8\alpha^2L^2(1-2\gamma)^2}{\gamma^2}\beta^2\mathbb{E}_t\|\nabla f(z^t)\|^2\\
&+\frac{112L^2\eta_l^2K}{\gamma N}\sum_{i\in N}\left(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2\right)\\
&+112\eta_l^2L^2K(3\sigma_l^2+5L^2\rho^2+16\sigma_g^2)\\
&+112\eta_l^2L^2K(16+3\beta^2)\|\nabla f(z^t)\|^2.
\end{aligned}\tag{25}$$

Thus we can bound the **A.1** as follows:

$$\begin{aligned}
&\leq\frac{\alpha}{2}\mathbb{E}_t\|\nabla f(z^t)\|^2+\frac{\alpha}{2}\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\mathbb{E}_t\left\|\mathbb{E}\tilde{g}_{i,k}^t-\nabla f_i(z^t)\right\|^2\\
&-\frac{\alpha}{2N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{g}_{i,k}^t\right\|^2\\
&\leq\left(\frac{\alpha}{2}+896\alpha\eta_l^2L^2K+168\alpha\eta_l^2L^2K\beta^2+\alpha\beta^2\right)\mathbb{E}_t\|\nabla f(z^t)\|^2\\
&+\frac{56\alpha L^2\eta_l^2K}{\gamma N}\sum_{i\in N}\left(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2\right)+\frac{2\alpha^3L^2}{\gamma^3}.
\end{aligned}$$

$$
\begin{aligned}
&(1-2\gamma)^2\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2-\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)\\
&+\alpha L^2\rho^2-\frac{\alpha}{2N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{g}_{i,k}^t\right\|^2\\
&+\frac{4\alpha^3L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2\\
&+56\alpha\eta_l^2L^2K(3\sigma_l^2+16\sigma_g^2+5L^2\rho^2). \qquad (26)
\end{aligned}
$$

We notice that **A.1** contains the term $\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2$ with a negative weight, thus we can set a suitable α to eliminate this term. Besides, the upper bound of **A.2** can be easy to get:

$$
\begin{aligned}
&=\mathbb{E}_t\|z^{t+1}-z^t\|^2\\
&=\mathbb{E}_t\left\|\alpha\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t+\alpha\beta\Delta^t\right\|^2\\
&\le\frac{2\alpha^2}{N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2+2\alpha^2\beta^2\mathbb{E}_t\left\|\Delta^t\right\|^2. \qquad (27)
\end{aligned}
$$

As we have bounded the term **A.1** and **A.2**, we combine the inequalities above and get $\mathbb{E}_t f(z^{t+1})$:

$$
\begin{aligned}
&\le\mathbb{E}_t f(z^t)-\alpha(1+\beta)\|\nabla f(z^t)\|^2+\textbf{A.1}+\frac{L}{2}\textbf{A.2}\\
&\le\mathbb{E}_t f(z^t)-\alpha(1+\beta)\|\nabla f(z^t)\|^2+\frac{L\alpha^2}{N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2\\
&+L\alpha^2\beta^2\mathbb{E}_t\left\|\nabla f(z^t)\right\|^2+\frac{56\alpha L^2\eta_l^2K}{\gamma N}\sum_{i\in N}\left(\mathbb{E}_t\|h_i^t\|^2\right.\\
&\left.-\mathbb{E}_t\|h_i^{t+1}\|^2\right)-\left(\frac{\alpha}{2}+\alpha\beta-896\alpha\eta_l^2L^2K-168\alpha\eta_l^2L^2K\beta^2\right.\\
&\left.-\alpha\beta^2\right)\mathbb{E}_t\|\nabla f(z^t)\|^2+\frac{2\alpha^3L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2\right.\\
&\left.-\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)+\alpha L^2\rho^2\\
&-\frac{\alpha}{2N^2}\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\mathbb{E}\tilde{g}_{i,k}^t\right\|^2\\
&+\frac{4\alpha^3L^2(1-2\gamma)^2}{\gamma^2}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}\sum_{k=0}^{K-1}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2\\
&+56\alpha\eta_l^2L^2K(3\sigma_l^2+16\sigma_g^2+5L^2\rho^2)\\
&\overset{(1)}{=}\mathbb{E}_t f(z^t)-\left(\frac{\alpha}{2}+\alpha\beta-1064\alpha\eta_l^2L^2K-\alpha\beta^2\right.\\
&\left.-L\alpha^2\beta^2\right)\mathbb{E}_t\|\nabla f(z^t)\|^2+56\alpha\eta_l^2L^2K(3\sigma_l^2+16\sigma_g^2+5L^2\rho^2)\\
&+\left(\frac{4\alpha^3L^2(1-2\gamma)^2}{N^2\gamma^2}+\frac{L\alpha^2}{N^2}-\frac{\alpha}{2N^2}\right)\mathbb{E}_t\left\|\sum_{i\in N}\sum_{k=0}^{K}\frac{\gamma_k}{\gamma}\tilde{g}_{i,k}^t\right\|^2\\
&+\frac{2\alpha^3L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2-\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)\\
&+\frac{56\alpha L^2\eta_l^2K}{\gamma N}\sum_{i\in N}\left(\mathbb{E}_t\|h_i^t\|^2-\mathbb{E}_t\|h_i^{t+1}\|^2\right)+\alpha L^2\rho^2. \qquad (28)
\end{aligned}
$$

(2) holds due to the fact $\beta\in(0,1)$. We set α to satisfy $\frac{4\alpha^3L^2(1-2\gamma)^2}{N^2\gamma^2}+\frac{L\alpha^2}{N^2}-\frac{\alpha}{2N^2}\le 0$, and we make $\alpha\omega=\frac{\alpha}{2}+\alpha\beta-896\alpha\eta_l^2L^2K-168\alpha\eta_l^2L^2K\beta^2-\alpha\beta^2-L\alpha^2\beta^2$; ω can be regarded as a constant.

proof for ω can be regarded as a constant.. First, Let $\beta=0$ means no dual variable correction exists. There exist a constant $c\in(0,1/2)$, we let $\omega=\frac{1}{2}-1064\eta_l^2L^2K\ge\frac{1}{2}-c>0$. Thus, $\omega=\frac{1}{2}-1064\eta_l^2L^2K\ge\frac{1}{2}-c$ when the $\eta_l\le\frac{\sqrt{c}}{\sqrt{1064KL}}<\frac{1}{\sqrt{2128KL}}$. For the final convergence, $\frac{1}{\omega}\le\frac{2c}{1-2c}$ is a constant upper bound. When β satisfy $\beta\le\frac{1}{1+L\alpha}$, the upper bound on $\frac{1}{w}$ does not changed. □

We take the full expectation on the bounded global gradient as $\alpha\omega\mathbb{E}\|\nabla f(z^t)\|^2$:

$$
\begin{aligned}
&\le\left(\mathbb{E}f(z^t)-\mathbb{E}f(z^{t+1})\right)+\frac{56\alpha L^2\eta_l^2K}{\gamma N}\sum_{i\in N}\left(\mathbb{E}\|h_i^t\|^2-\mathbb{E}\|h_i^{t+1}\|^2\right)\\
&+\frac{2\alpha^3L^2(1-2\gamma)^2}{\gamma^3}\left(\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^t\right\|^2-\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^{t+1}\right\|^2\right)\\
&+56\alpha\eta_l^2L^2K(3\sigma_l^2+16\sigma_g^2+5L^2\rho^2)+\alpha L^2\rho^2. \qquad (29)
\end{aligned}
$$

Take the full expectation and telescope the sum on the above inequality $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(z^t)\|^2$:

$$
\begin{aligned}
&\le\frac{1}{T\alpha\omega}\left(f(z^0)-\mathbb{E}_t f(z^T)\right)+\frac{56\alpha L^2\eta_l^2K}{T\gamma N\omega}\sum_{i\in N}\mathbb{E}\|h_i^0\|^2\\
&+\frac{2\alpha^2L^2(1-2\gamma)^2}{T\gamma^3\omega}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^0\right\|^2\\
&+\frac{1}{\omega}\left(56\eta_l^2L^2K(3\sigma_l^2+16\sigma_g^2+5L^2\rho^2)+L^2\rho^2\right). \qquad (30)
\end{aligned}
$$

Here, we summarize the conditions and some constraints in the conclusion mentioned above. Like [11], we note that $(1-(1-\eta_l/\alpha)^K)<1$ when $\eta_l\le\alpha$. We have $1/\gamma>1$. When $K>\alpha/\eta_l$, $(1-\frac{\eta_l}{\alpha})^K\le e^{-\eta_lK/\alpha}\le e^{-1}$, then $\gamma>1-e^{-1}$ and $\frac{1}{\gamma}<\frac{e}{e-1}<2$. And apply the fact $f^*\le f(x),\quad\forall x\in\mathbb{R}^d$:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(z^t)\|^2\le\frac{1}{T\alpha\omega}\left(f(z^0)-f^*\right)\\
&+\frac{112L^2\eta_l^2K}{TN\omega}\sum_{i\in N}\mathbb{E}\|h_i^0\|^2\\
&+\frac{16\alpha^2L^2}{T\omega}\mathbb{E}_t\left\|\frac{1}{N}\sum_{i\in N}h_i^0\right\|^2\\
&+\frac{1}{\omega}\left(56\eta_l^2L^2K(3\sigma_l^2+16\sigma_g^2+5L^2\rho^2)+L^2\rho^2\right). \qquad (31)
\end{aligned}
$$

This completes our proof of Theorem 2.

This article has been accepted for publication in IEEE Transactions on Network Science and Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2025.3638762

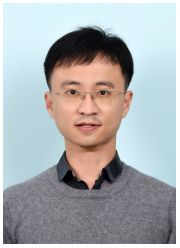
Boyuan Li received the master's degree from Henan University, China. He is now pursuing the Ph.D. degree at Zhengzhou University. His research interests include distributed computing, network science, and federated learning.

Zihao Peng received the master's degree from Nanchang University, China. He is now pursuing the Ph.D. degree at Beijing Normal University. His research interests include federated learning and large models.

Yafei Li received the PhD degree in computer science from Hong Kong Baptist University in 2015. He is currently a professor with the School of Computer and Artificial Intelligence, Zhengzhou University, China.

His research interests include crowd intelligence, mobile and spatial data management, location-based services, and urban computing. He has authored more than 30 journal and conference papers in these areas, including IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Mobile Computing, IEEE Transactions on Services Computing, IEEE Transactions on Neural Networks and Learning Systems, PVLDB, IEEE ICDE, WWW, etc.

Zijian Li (Member, IEEE) received the B.S. and Ph.D. degrees in computer science from Jilin University, Changchun, China, in 2017 and 2023, respectively. During his doctoral studies, he also worked as a visiting Ph.D. student with the Wireless Networks and Decision Systems (WNDS) Group of Singapore University of Technology and Design, Singapore. He was a postdoctoral research fellow with Dalian Maritime University, Dalian, China, from 2023 to 2025. Currently, he is a lecturer with the College of Artificial Intelligence, Dalian Maritime University. His research interests include edge intelligence, federated learning, distributed computing, and crowdsensing.



Shengbo Chen (Member, IEEE) received the B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2006 and 2008, respectively, and the Ph.D. degree from The Ohio State University, Columbus, OH, USA, in 2013. From 2013 to 2019, he was a Senior Research Engineer with the Qualcomm Research Center, San Diego, CA, USA. He is currently a Professor with the School of Software, Nanchang University, China. He holds more than 50 U.S. patents in the 5G and AI areas. He received the Best Student Paper Award from WiOpt 2013.

Cong Shen (Senior Member, IEEE) received his B.E. and M.E. degrees from the Department of Electronic Engineering, Tsinghua University, China. He received the Ph.D. degree from the Electrical Engineering Department, University of California Los Angeles (UCLA). He is an Associate Professor of the Electrical and Computer Engineering Department at the University of Virginia (UVa). Prior to joining UVa, he was a professor in the School of Information Science and Technology at the University of Science and Technology of China (USTC). He also has extensive industry experience, having worked for Qualcomm Research, SpiderCloud Wireless, Silvus Technologies, and Xsense.ai, in various full-time and consulting roles. His general research interests are in the area of machine learning, signal processing, communication systems, and networking. In particular, his current research focuses on generative models, in-context learning, reinforcement learning, federated learning, and their engineering applications.

He received the NSF CAREER award in 2022. He was the recipient of the Best Paper Award in the 2021 IEEE International Conference on Communications (ICC), and the Excellent Paper Award in the 9th International Conference on Ubiquitous and Future Networks (ICUFN 2017). Currently, he serves as an associate editor for the IEEE Transactions on Communications, an editor for the IEEE Transactions on Wireless Communications, an editor for the IEEE Transactions on Green Communications and Networking, and an associate editor for IEEE Transactions on Machine Learning in Communications and Networking. He was the TPC co-chair of the Wireless Communications Symposium of IEEE Globecom 2021, and actively serves as (senior) program committee members/reviewers for the Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), International Conference on Learning Representations (ICLR), International Conference on Artificial Intelligence and Statistics (AISTATS), International Joint Conference on Artificial Intelligence (IJCAI), and the AAAI Conference on Artificial Intelligence (AAAI). He is a member of SpectrumX, an NSF Spectrum Innovation Center.

Tony Q.S. Quek (S'98-M'08-SM'12-F'18) received the B.E. and M.E. degrees in electrical and electronics engineering from the Tokyo Institute of Technology in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology in 2008. Currently, he is the Cheng Tsang Man Chair Professor with Singapore University of Technology and Design (SUTD) and ST Engineering Distinguished Professor. He also serves as the Director of the Future Communications R&D Programme, the Head of ISTD Pillar, and the AI on RAN Working Group Chair in AI-RAN Alliance. His current research topics include wireless communications and networking, network intelligence, non-terrestrial networks, open radio access network, and 6G.

Dr. Quek has been actively involved in organizing and chairing sessions and has served as a member of the Technical Program Committee as well as symposium chairs in a number of international conferences. He is currently serving as an Area Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Dr. Quek was honored with the 2008 Philip Yeo Prize for Outstanding Achievement in Research, the 2012 IEEE William R. Bennett Prize, the 2015 SUTD Outstanding Education Awards – Excellence in Research, the 2016 IEEE Signal Processing Society Young Author Best Paper Award, the 2017 CTTC Early Achievement Award, the 2017 IEEE ComSoc AP Outstanding Paper Award, the 2020 IEEE Communications Society Young Author Best Paper Award, the 2020 IEEE Stephen O. Rice Prize, the 2020 Nokia Visiting Professor, and the 2022 IEEE Signal Processing Society Best Paper Award. He is an IEEE Fellow, a WWRF Fellow, and a Fellow of the Academy of Engineering Singapore.