HISTORICAL GRADIENT RECONSTRUCTION FOR ADAPTIVE FEDERATED CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated continual learning (FCL) must balance the acquisition of new knowledge with the retention of earlier knowledge distributed across heterogeneous clients. Gradient-based task weighting offers a way to adapt this balance, but choosing the weights requires historical gradients from data distributed across clients. Private replay buffers provide only a local view of past data, while shared historical gradients can become inaccurate as clients update their models. We propose FEDGRIT (Federated Gradient Reconstruction for Incremental Training), a replay-based framework that reconstructs historical gradients to guide adaptive local updates. By aggregating task-specific gradients and compact curvature information, FEDGRIT enables clients to *reconstruct historical gradients at their own changing models and adapt task weights without cross-client communication at every step.* Periodic refreshes and two-stage curvature compression support repeated sharing. We derive conditional historical-loss retention and weighted stationarity bounds under partial participation, separating the effects of stochasticity, local training, and client heterogeneity. Across four class- and domain-incremental benchmarks, FEDGRIT achieves the highest mean final accuracy among the evaluated methods with ablations examining shared reconstruction and adaptive weighting.

1 INTRODUCTION

Continual learning (CL) studies how models learn sequential tasks with limited access to earlier data. Its central challenge is to balance *stability*, the retention of previously acquired knowledge, and *plasticity*, the acquisition of new knowledge (Lopez-Paz & Ranzato, 2017). When such learning takes place across clients that keep their data local, federated continual learning (FCL) enables collaboration over evolving private data (Yoon et al., 2021). In this paper, we study the shared-model setting, where local updates must support new-task learning while preserving historical knowledge distributed across heterogeneous clients. The appropriate balance can vary across tasks, clients, and training stages, making fixed weights difficult to choose in advance. Gradient-based weighting adapts this balance using the norms and alignment of task gradients, as studied in Pareto-based CL (Wu et al., 2024b) and federated multi-objective optimization (MOO) (Yang et al., 2023; Askin et al., 2025). In FCL, however, these decisions require *historical-gradient estimates that reflect distributed past data at each client's changing local model.*

Estimating these gradients is difficult because each client's private replay buffer captures only a local view of the federation's history. An update that is compatible with a client's past data may still interfere with knowledge retained elsewhere. Sharing replay gradients broadens this view, as in C-FLAG (Keshri et al., 2025), which aggregates a pooled historical gradient at the round's shared model. Adaptive per-task decisions instead require estimates that retain task identity and remain informative as each local model changes. Recomputing these gradients across client after every local step would require repeated communication. This raises our central question:

> *How can distributed historical information support adaptive task balancing throughout local training without requiring cross-client communication at every update?*

We answer this question with FEDGRIT (Federated Gradient Reconstruction for Incremental Training), a replay-based framework that makes shared historical information usable throughout local

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

training. Building on past-gradient reconstruction in CL (Daxberger et al., 2023), participating clients contribute task-specific replay gradients and compact curvature information at a common model. Aggregation incorporates history from other clients, while curvature corrections adjust the estimates as local models move. Each client combines these reconstructed gradients with its current-task gradient to adapt task weights at every local step, without further cross-client queries. Periodic refreshes recenter the summaries, while client compression and server recompression keep each curvature summary compact.

Our main contributions are as follows:

• **Shared reconstruction for adaptive local updates.** We develop FEDGRIT, which combines task-specific historical gradients across clients and adjusts them to each evolving local model, enabling adaptive task weighting without cross-client queries at every local step.

• **Conditional historical-loss retention.** We bound the increase in cumulative historical-task loss under an alignment condition on the adaptive local directions, explicitly separating stochastic noise and client sampling.

• **Weighted stationarity under local adaptation.** We analyze the federation-wide task-gradient combination induced by client-specific local weights, distinguishing optimization and local-update effects from sampling-related and persistent client heterogeneity.

• **Evaluation across incremental settings.** FEDGRIT achieves the highest mean final accuracy among the evaluated methods on four class- and domain-incremental benchmarks. Component and sensitivity studies examine shared reconstruction, weighting, and curvature sharing.

2 PROBLEM SETTING

We consider a server coordinating M clients to train a shared model. Clients follow a shared task sequence with synchronized task transitions (Keshri et al., 2025; Xu & Li, 2026), while their data distributions may differ within each task. At stage K, client m observes its current-task dataset $\mathcal{T}_{K,m}$ and maintains a private replay buffer of fixed total capacity containing samples from earlier tasks. Earlier-task data are accessible only through these buffers, and raw samples are not exchanged between clients or with the server. Replay samples retain their task identities for training.

For model parameters $\theta \in \mathbb{R}^d$ and task i, let $f_{i,m}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{T}_{i,m}}[f_{i,m}(\theta; x, y)]$ denote the loss of client m, where (x, y) is an input-label pair. The federation-wide task loss is $f_i(\theta) = M^{-1}\sum_{m=1}^{M} f_{i,m}(\theta)$. The cumulative learning objective is

$$\min_{\theta\in\mathbb{R}^d} f_{1:K}(\theta) := \sum_{i=1}^{K} f_i(\theta) = \frac{1}{M}\sum_{i=1}^{K}\sum_{m=1}^{M} f_{i,m}(\theta). \tag{1}$$

This objective accounts for both learning the current task and retaining knowledge from earlier tasks. During new-task training, however, the full earlier-task datasets are unavailable, and each client's replay buffer contains only a limited sample of its own history. *An update that preserves this local history may therefore still degrade performance on historical data held by other clients.*

At communication round t, the server samples a set $\mathcal{S}_K^t$ of N clients uniformly without replacement. We use T for the number of rounds per task and E for the number of local gradient steps performed by each participating client per round. The shared model at the start of round t is θ_K^t, and $\theta_{K,m}^{t,e}$ denotes client m's model after e local steps, with $\theta_{K,m}^{t,0} = \theta_K^t$. For $K > 1$. Training on task K starts from $\theta_K^0 = \theta_{K-1}^T$. The local and server learning rates are η_L and η_G, respectively.

3 FEDGRIT: ADAPTIVE UPDATES FROM RECONSTRUCTED HISTORY

Gradient-based task balancing coordinates current- and previous-task gradients to support new learning while limiting interference with earlier knowledge (Wu et al., 2024b). In FCL, each client learns from its current-task data, while the historical knowledge to be retained is distributed across clients. Its local update should therefore account for how learning from its current data affects earlier tasks

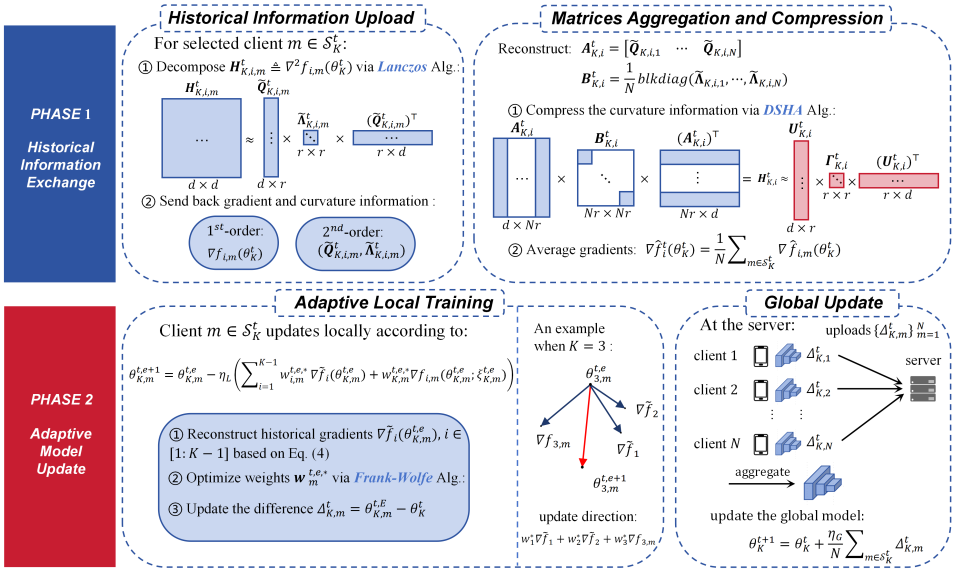


Figure 1: **FEDGRIT in one communication round.** Clients share task-specific replay gradients and low-rank curvature factors, which the server aggregates, recompresses, and broadcasts. During E local steps, each client reconstructs historical gradients at its current model and adapts task weights without further communication. The server then aggregates the model updates

across the federation. For client m training on task K, we consider the following idealized direction at local step e of communication round t:

$$\mathbf{d}_{K,m}^{t,e} = \sum_{i=1}^{K-1} w_{i,m}^{t,e} \nabla f_i(\theta_{K,m}^{t,e}) + w_{K,m}^{t,e} \nabla f_{K,m}(\theta_{K,m}^{t,e}; \xi_{K,m}^{t,e}). \tag{2}$$

Here, $\nabla f_i(\theta_{K,m}^{t,e})$ is the federation-wide gradient of historical task i, evaluated at client m's current local model. The current-task gradient $\nabla f_{K,m}(\theta_{K,m}^{t,e}; \xi_{K,m}^{t,e})$ is computed from a minibatch $\xi_{K,m}^{t,e}$ drawn from client m's current-task data. The weights $\boldsymbol{w}_m^{t,e} = (w_{1,m}^{t,e}, \ldots, w_{K,m}^{t,e})^\top$ determine each task's contribution, with $w_{i,m}^{t,e} \geq 0$ and $\sum_{i=1}^K w_{i,m}^{t,e} = 1$.

Constructing this direction requires historical gradients at the changing local model and weights that balance their contributions. *These requirements are coupled: adapting the weights using gradient agreement requires historical-gradient estimates that remain informative throughout local training.* A private replay buffer provides only a limited view of the federation's history. Sharing replay gradients can broaden this view (Keshri et al., 2025), but gradients evaluated at the shared model can become inaccurate as a client's local model changes. Recomputing the shared replay gradients at every local iterate would require repeated communication with other clients. FEDGRIT addresses this difficulty through shared historical-gradient reconstruction.

3.1 RECONSTRUCTING HISTORICAL GRADIENTS DURING LOCAL TRAINING

To adjust historical gradients as the local model moves away from the shared model, we use the first-order approximation

$$\nabla f_i(\theta_{K,m}^{t,e}) \approx \nabla f_i(\theta_K^t) + \boldsymbol{H}_{K,i}^t(\theta_{K,m}^{t,e} - \theta_K^t), \tag{3}$$

where $\boldsymbol{H}_{K,i}^t := \nabla^2 f_i(\theta_K^t)$ is the federation-wide task-i Hessian at the shared model. The curvature term estimates how the historical gradient changes as client m moves away from θ_K^t. The exact gradient and Hessian in this approximation are unavailable because the full historical datasets are no longer accessible.

We estimate these quantities using participating clients' private replay buffers. During training on task K, let $\hat{f}_{i,m}$ denote the empirical loss on client m's retained samples from task i. For the

participating-client set $\mathcal{S}_K^t$, define the shared replay loss $\widehat{f}_i^t(\theta) := \frac{1}{N}\sum_{m\in\mathcal{S}_K^t}\widehat{f}_{i,m}(\theta)$. Its gradient at θ_K^t is obtained by averaging the replay gradients uploaded by participating clients. It estimates the federation-wide gradient using their retained data. Let $\tilde{\boldsymbol{H}}_{K,i}^t$ denote the compact curvature approximation constructed from the same clients' replay losses, as detailed in Section 3.3. Each client then reconstructs the historical gradient as

$$\nabla\tilde{f}_i(\theta_{K,m}^{t,e}) := \nabla\widehat{f}_i^t(\theta_K^t) + \tilde{\boldsymbol{H}}_{K,i}^t(\theta_{K,m}^{t,e} - \theta_K^t). \tag{4}$$

The shared gradient incorporates history from participating clients, while the curvature term adjusts it to the receiving client's model. Setting the curvature term to zero recovers reuse of the shared gradient. Accuracy depends on replay coverage, client sampling, curvature approximation, and deviation from the shared model.

3.2 Adaptive Task Weighting and Local Updates

Reconstruction supplies historical-gradient estimates at the client's current local model. We use these estimates together with the current-task gradient to adjust task weights according to their norms and mutual alignment. Following gradient-based multi-objective optimization (Yang et al., 2023), we select a minimum-norm convex combination:

$$\boldsymbol{w}_m^{t,e,*} \in \underset{\substack{\boldsymbol{w}\geq\mathbf{0}\\ \sum_{i=1}^K w_i=1}}{\arg\min} \left\| \sum_{i=1}^{K-1} w_i\nabla\tilde{f}_i(\theta_{K,m}^{t,e}) + w_K\nabla f_{K,m}(\theta_{K,m}^{t,e};\xi_{K,m}^{t,e}) \right\|^2. \tag{5}$$

Because the current-task weight is also optimized, the rule adjusts both the relative emphasis among historical tasks and the balance between historical and current learning.

For fixed gradient estimates, Eq. (5) is a convex quadratic problem in K weights, approximately solved by Frank–Wolfe (Appendix C.2 (Frank & Wolfe, 1956)). We write the updates using its exact minimizer $\boldsymbol{w}_m^{t,e,*}$ for the analysis:

$$\tilde{\mathbf{d}}_{K,m}^{t,e,*} = \sum_{i=1}^{K-1} w_{i,m}^{t,e,*}\nabla\tilde{f}_i(\theta_{K,m}^{t,e}) + w_{K,m}^{t,e,*}\nabla f_{K,m}(\theta_{K,m}^{t,e};\xi_{K,m}^{t,e}), \tag{6}$$

$$\theta_{K,m}^{t,e+1} = \theta_{K,m}^{t,e} - \eta_L\tilde{\mathbf{d}}_{K,m}^{t,e,*}. \tag{7}$$

The shared summaries remain fixed within a round, while reconstruction and weighting are repeated at every local step. The curvature correction adjusts the historical-gradient estimates as the local model moves, and the weights respond to the resulting relationships among current and historical gradients. Both computations are performed locally without further cross-client communication. For the first task, the historical sum is empty and $w_{1,m}^{t,e,*} = 1$.

3.3 Sharing Compact Curvature Information

Historical gradients can be averaged directly, but a full client Hessian contains d^2 entries. FedGRIT therefore compresses curvature before upload and recompresses the aggregate before broadcast.

Client-side compression. For each historical task $i < K$, participating client m approximates the Hessian of its replay loss at the shared model:

$$\nabla^2\widehat{f}_{i,m}(\theta_K^t) \approx \tilde{\boldsymbol{Q}}_{K,i,m}^t\tilde{\boldsymbol{\Lambda}}_{K,i,m}^t(\tilde{\boldsymbol{Q}}_{K,i,m}^t)^\top, \tag{8}$$

Here, $\widetilde{\boldsymbol{Q}}_{K,i,m}^t \in \mathbb{R}^{d\times r}$ has orthonormal columns, and $\widetilde{\boldsymbol{\Lambda}}_{K,i,m}^t \in \mathbb{R}^{r\times r}$ is diagonal with the corresponding signed eigenvalue estimates. We use Lanczos with Hessian-vector products to approximate the r eigenpairs with largest eigenvalue magnitudes, without forming the full Hessian (Lanczos, 1950; Chen et al., 2021). The procedure is given in Algorithm 2 in Appendix C.1. Each client uploads these factors together with its replay gradient $\nabla\widehat{f}_{i,m}(\theta_K^t)$.

Server-side aggregation and recompression. Averaging rank-r client approximations can produce rank up to $\min(Nr, d)$, so broadcasting all collected factors would make each recipient's curvature

Algorithm 1 FEDGRIT for task K with per-round summary refresh

Require: Initial model θ_K^0, clients' current-task datasets and private replay buffers, $T, E, N, r, \eta_L, \eta_G$
Ensure: Shared model θ_K^T and updated private replay buffers

for $t = 0, \dots, T-1$ **do**
  Server samples N clients as $\mathcal{S}_K^t$ and sends them θ_K^t.
  if $K > 1$ **then**
    Each participating client computes and uploads replay gradients and Lanczos curvature factors at θ_K^t for tasks $i = 1, \dots, K-1$.
    Server averages the replay gradients and recompresses each task's curvature using DSHA.
    Server broadcasts the shared gradients and curvature factors.
  end if
  for each client $m \in \mathcal{S}_K^t$ in parallel **do**
    Initialize $\theta_{K,m}^{t,0} = \theta_K^t$.
    for $e = 0, \dots, E-1$ **do**
      Reconstruct historical-task gradients using Eq. (4) when $K > 1$.
      Draw $\xi_{K,m}^{t,e}$ and compute the current-task gradient.
      For $K > 1$, approximately solve Eq. (5) using Frank-Wolfe; for $K = 1$, set $w_{1,m}^{t,e,*} = 1$.
      Use the computed weights to form the direction and update the local model according to Eqs. (6)-(7).
    end for
    Upload $\Delta_{K,m}^t = \theta_{K,m}^{t,E} - \theta_K^t$.
  end for
  Server updates θ_K^{t+1} using Eq. (10).
end for
Clients update their fixed-size private replay buffers.

payload grow with the number of participants. Distributed Spectral Hessian Aggregation (DSHA) recompresses this aggregate by retaining at most r eigenpairs with the largest eigenvalue magnitudes, preserving their signs:

$$\tilde{\boldsymbol{H}}_{K,i}^t := \boldsymbol{U}_{K,i}^t \boldsymbol{\Gamma}_{K,i}^t (\boldsymbol{U}_{K,i}^t)^\top \approx \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \tilde{\boldsymbol{Q}}_{K,i,m}^t \tilde{\boldsymbol{\Lambda}}_{K,i,m}^t (\tilde{\boldsymbol{Q}}_{K,i,m}^t)^\top. \tag{9}$$

The columns of $\boldsymbol{U}_{K,i}^t \in \mathbb{R}^{d \times r}$ are orthonormal, and $\boldsymbol{\Gamma}_{K,i}^t \in \mathbb{R}^{r \times r}$ contains the retained signed eigenvalues. The QR-based construction and DSHA pseudocode are provided in Appendix C.3.

The server broadcasts the averaged replay gradient $\nabla \widehat{f}_i^t(\theta_K^t)$ and the factors $(\boldsymbol{U}_{K,i}^t, \boldsymbol{\Gamma}_{K,i}^t)$ for each historical task. Clients apply these factors directly in Eq. (4), without forming a $d \times d$ matrix, at a cost of $\mathcal{O}(dr)$ per historical task per local step. For $r \geq 1$, each client receives $\mathcal{O}((K-1)dr)$ scalar entries per historical-summary refresh. Recompression makes this per-client payload independent of N, although it still grows with the number of historical tasks.

Complete round protocol. Algorithm 1 summarizes FEDGRIT with historical summaries refreshed at the beginning of every round. After exchanging summaries at θ_K^t, participating clients perform E adaptive local steps as described in Section 3.2, without further cross-client queries. Each client uploads $\Delta_{K,m}^t = \theta_{K,m}^{t,E} - \theta_K^t$, and the server updates

$$\theta_K^{t+1} = \theta_K^t + \eta_G \Delta_K^t, \qquad \Delta_K^t := \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \Delta_{K,m}^t. \tag{10}$$

At each task boundary, clients update their fixed-size private buffers using retained samples and samples from the completed task. Raw samples remain local throughout the procedure.

4 THEORETICAL ANALYSIS

We analyze how new task training affects cumulative historical loss and how the adaptive updates approach weighted stationarity across the tasks seen so far.

4.1 ASSUMPTIONS

Assumption 1 (Bounded Gradients). *For every task $i \in [K]$ and client $m \in [M]$, the gradient is bounded by a constant $D \geq 0$, i.e., $\mathbb{E}[\|\nabla f_{i,m}(\theta)\|] \leq D, \forall \theta \in \mathbb{R}^d$.*

Assumption 2 (L-smooth). *For every task $i \in [K]$, client $m \in [M]$, and all $\theta, \theta' \in \mathbb{R}^d$, there exists a constant $L > 0$ such that $\|\nabla f_{i,m}(\theta) - \nabla f_{i,m}(\theta')\| \leq L\|\theta - \theta'\|$.*

Assumption 3 (Unbiased Stochastic Estimation). *For every task $i \in [K]$, client $m \in [M]$, the stochastic gradient estimation is unbiased, i.e., $\mathbb{E}_t[\nabla f_{i,m}(\theta, \xi)] = \nabla f_{i,m}(\theta), \forall \xi \in \mathcal{T}_{i,m}$.*

Assumption 4 (Bounded Local Variance). *For every task $i \in [K]$, client $m \in [M]$, there always exists a constant $\sigma_L > 0$, such that the variance of stochastic gradient is bounded by $\mathbb{E}_t \|\nabla f_{i,m}(\theta; \xi) - \nabla f_{i,m}(\theta)\|^2 \leq \sigma_L^2$.*

Assumption 5 ((L, β)-Bounded Global Heterogeneity). *For every task $i \in [K]$, client $m \in [M]$, and all $\theta, \theta' \in \mathbb{R}^d$, there exists a constant $\beta > 0$, such that, the global heterogeneity of the local gradient is bounded by $\|\nabla f_i(\theta) - \nabla f_{i,m}(\theta')\|^2 \leq L^2\|\theta - \theta'\|^2 + \beta\sigma_G^2$.*

Assumptions 1- 4 control gradient magnitude, smoothness, and stochastic estimation. Assumption 5 states that the global-local gradient discrepancy is jointly controlled by the model deviation and client heterogeneity. When $\theta = \theta'$, Assumption 5 reduces to the conventional bounded global variance assumption (assumption 3 in Yang et al. (2021)), and it also reduces to the standard Lipschitz continuous assumption as Assumption 2 when specialized to i.i.d. case.

4.2 BACKWARD KNOWLEDGE TRANSFER ANALYSIS

We compare the cumulative historical loss during task-K training with its value at the task-start model $\theta_K^0 = \theta_{K-1}^T$. The following result bounds its increase under an alignment condition on the local update directions.

Theorem 1 (BWT under partial participation). *Suppose for each $\tau \in [t]$, local epoch $e \in [E]$, and client $m \in [M]$, the update direction $\tilde{\mathbf{d}}_{K,m}^{\tau,e}$ is aligned with earlier-task gradient at task-K start, i.e., $\langle \nabla f_{1:K-1}(\theta_K^0), \tilde{\mathbf{d}}_{K,m}^{\tau,e} \rangle \geq \epsilon \|\nabla f_{1:K-1}(\theta_K^0)\| \cdot \|\tilde{\mathbf{d}}_{K,m}^{\tau,e}\|, \epsilon \in (0, 1)$, and local and global learning rates satisfy $\eta_L \leq \min\{\frac{1}{3(K-1)EL}, \frac{2\epsilon\|\nabla f_{1:K-1}(\theta_K^0)\|}{DLEt\sqrt{7tEN}}\}$ and $\eta_G \leq \frac{1}{K-1}$. Then, for every $t \geq 1$, the server-updated model θ_K^t generated by the algorithm satisfies:*

$$\mathbb{E}_t\left[f_{1:K-1}\left(\theta_K^t\right)\right] \leq f_{1:K-1}\left(\theta_K^0\right) + \frac{2\epsilon^2\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|^2}{7\left(K-1\right)Et^2NLD^2}\left(\sigma_L^2 + \frac{\left(M-N\right)\beta\sigma_G^2}{N\left(M-1\right)}\right). \tag{11}$$

Implications. Eq. (11) bounds historical-loss deterioration relative to the task-start model, separating local stochasticity from sampling-related client heterogeneity. Increasing N reduces both the common $1/N$ factor and the sampling factor. Full participation eliminates the heterogeneity contribution but not necessarily the stochastic contribution. The guarantee requires alignment with the task-start historical gradient, and its dependence on t and E must be interpreted together with the corresponding step-size restrictions. It controls cumulative historical loss, not improvement on every earlier task.

Remark 1 (Relation to prior backward-transfer guarantees). *Lin et al. (2022) show that, under gradient alignment and a suitable step-size restriction, learning a new task does not increase an earlier-task loss (Theorem 2). Our alignment condition follows the same principle, applied to adaptive local update directions and the cumulative historical loss in a federated setting. From a complementary perspective, Shi & Wang (2023) bound the error on a previous domain by the historical model's error plus a prediction-disagreement term (Lemma 3.2). Equation (11) also compares historical performance before and after new-task learning, but expresses the correction through stochasticity and client sampling in the federated updates. Xu & Li (2026) studies historical-loss retention under server-proximal aggregation, with an explicit stochastic correction proportional to σ_L^2 in its BWT bound. Here, the guarantee concerns adaptive local directions and additionally separates the contribution of $\beta\sigma_G^2$ through the factor $(M-N)/(N(M-1))$. This distinction identifies how partial participation enters the historical-loss bound for heterogeneous clients.*

4.3 WEIGHTED TASK-UNIFORM CONVERGENCE OF FEDGRIT

The preceding BWT analysis concerns changes in historical loss. We now examine stationarity across the tasks seen so far. Since FEDGRIT adjusts task weights separately at each client's local steps, we evaluate the resulting combination of federation-wide task gradients: $\nabla \tilde{f}_{1:K}(\theta_K^t) \triangleq \sum_{i=1}^{K} w_i^{t,*} \nabla f_i(\theta_K^t)$, where $w_i^{t,*} \triangleq \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*}$ is the average weight assigned to task i across participating clients and local steps in round t. These weights are nonnegative and sum to one, so the resulting gradient combination provides a weighted stationarity criterion for tasks $1, \ldots, K$. A zero value certifies Pareto stationarity, though individual task gradients need not vanish.

Theorem 2 (Weighted Task-Uniform Convergence of FEDGRIT)**.** *Suppose Assumptions 1-5 hold, local and global learning rates η_L and η_G satisfy $\eta_L \leq \frac{1}{3(K-1)EL}, K \geq 3$, and $\eta_G \eta_L \leq \frac{1}{6EL}$. Then, the outputs of* FEDGRIT *satisfy*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_t \left\| \nabla \tilde{f}_{1:K}\left(\theta_K^t\right) \right\|^2 \leq \frac{8\left(f_{1:K}^0 - f_{1:K}^*\right)}{\eta_G \eta_L EKT} + 4\eta_G \eta_L EL\sigma_L^2 + \frac{4(M-N)}{N(M-1)} \eta_G \eta_L \beta EL \sigma_G^2 + 80\left[1 + \Lambda_{1:K-1} + \Lambda_{1:K}\right] \eta_L^2 E^2 D^2 L^2 + 4\beta\sigma_G^2, \tag{12}$$

where $f_{1:K}^0 \triangleq f_{1:K}\left(\theta_K^0\right)$, $f_{1:K}^* \triangleq f_{1:K}\left(\theta_K^*\right)$, $\Lambda_K \triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_t \left\| \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \right\|^2$, *and* $\Lambda_{1:K-1} \triangleq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_t \left\| \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(1 - w_{K,m}^{t,e,*}\right) \right\|^2$.

Implications. Eq. (12) separates optimization progress from the effects of stochasticity, client sampling, and local updates. The first term depends on the initial optimality gap $f_{1:K}^0 - f_{1:K}^*$ and decreases with T for fixed admissible learning rates and local-step count E. The second term captures local stochastic variance σ_L^2, while the third captures the contribution of client heterogeneity under partial participation. Its factor $(M-N)/(N(M-1))$ decreases as more clients participate and vanishes when $N = M$.

The local-update term scales with $\eta_L^2 E^2 D^2 L^2$ and depends on the weight moments $\Lambda_{1:K-1}$ and Λ_K. These quantities summarize the total weight assigned to historical tasks and to the current task, respectively. Since $w_K^{t,*} \in [0,1]$, their sum lies in $[1/2, 1]$. Thus, adaptive weight allocation changes the coefficient of this term within a bounded range. Choosing η_L proportional to $1/E$ controls its explicit growth with the number of local steps; increasing E alone does not necessarily tighten the overall bound. The remaining term $4\beta\sigma_G^2$ reflects the global-local gradient discrepancy allowed by Assumption 5. Unlike the sampling-related heterogeneity term, it persists under full participation and is not reduced by the learning rate alone. For fixed K, η_G, and problem parameters, choosing $\eta_L = \Theta(1/(EL\sqrt{T}))$ reduces the other terms at rates $O(T^{-1/2})$ or faster, leaving a stationarity bound of $4\beta\sigma_G^2 + O(T^{-1/2})$. The result therefore quantifies how optimization and local training interact with the heterogeneity tolerance in the bound.

Remark 2 (Relation to prior convergence analyses)**.** *Han et al. (2023) and Keshri et al. (2025) analyze adaptive updates that balance a pooled historical objective with the current objective, including stationarity guarantees for the historical objective. Here, historical tasks retain separate weights, and the quantity $\nabla \tilde{f}_{1:K}(\theta_K^t)$ tracks the federation-wide gradient combination induced by the client-specific local weighting decisions. Xu & Li (2026) establishes a convergence bound for the cumulative objective $f_{1:K}$ under a bounded inter-task gradient difference assumption. Its bound explicitly includes the initial historical-gradient norm $\|\nabla f_{1:K-1}(\theta_K^0)\|^2$. Eq. (12) instead evaluates stationarity under adaptive task weights and makes their historical–current allocation explicit through $\Lambda_{1:K-1}$ and Λ_K. The two results concern different stationarity criteria. The distinguishing feature here is the analysis of a global task-gradient combination induced by local adaptive decisions, with separate terms for partial participation and persistent global-local gradient discrepancy.*

Corollary 1 (Convergence Rate)**.** *With a constant learning rate η_G and a local learning rate $\eta_L = \mathcal{O}(\frac{1}{\sqrt{KT}EL})$, and if $\beta = \mathcal{O}(\eta_L)$, when the length of the task sequence K is fixed, the weighted task-uniform convergence rate satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_t \left\| \nabla \tilde{f}_{1:K}\left(\theta_K^t\right) \right\|^2 = \mathcal{O}\left(\frac{1}{\sqrt{T}} + \frac{\sigma_L^2}{\sqrt{T}} + \frac{\sigma_G^2}{E\sqrt{T}}\right). \tag{13}$$

Table 1: Final average accuracy (ACC, %) and backward transfer (BWT, pp) on four benchmarks. Results are averaged over three seeds. Bold denotes the best reported mean in each column.

	CIFAR-100		ImageNet100		Digit-10		VLCS	
Method	ACC (%) ↑	BWT (pp) ↑	ACC (%) ↑	BWT (pp) ↑	ACC (%) ↑	BWT (pp) ↑	ACC (%) ↑	BWT (pp) ↑
FedAvg	43.21 ± 1.07	-19.09 ± 1.87	28.94 ± 0.66	-25.65 ± 0.66	61.83 ± 1.22	-43.72 ± 0.98	53.33 ± 0.74	-18.81 ± 0.21
FedProx	43.25 ± 2.42	-20.92 ± 2.21	31.55 ± 1.83	-24.01 ± 2.08	60.55 ± 1.40	-45.68 ± 0.73	44.99 ± 0.32	-22.69 ± 0.88
FSMGDA	34.00 ± 0.52	-11.75 ± 0.80	21.93 ± 2.01	-14.35 ± 3.89	67.31 ± 0.84	-21.14 ± 0.44	53.78 ± 2.29	-5.01 ± 1.80
FedCMOO	45.09 ± 2.93	-11.37 ± 1.64	31.04 ± 2.12	-11.82 ± 2.59	69.46 ± 0.80	-20.01 ± 0.41	54.79 ± 0.61	-4.26 ± 0.85
FLwF-2T	34.01 ± 1.21	-17.66 ± 2.31	29.14 ± 1.26	-12.49 ± 2.33	68.31 ± 0.70	-21.00 ± 0.35	51.07 ± 0.29	-7.69 ± 0.49
FedSSI	44.77 ± 0.55	-10.07 ± 0.09	32.52 ± 0.50	-10.38 ± 0.59	52.37 ± 1.45	-17.80 ± 2.14	49.22 ± 3.20	-9.16 ± 1.91
SPECIAL	43.17 ± 1.69	-10.07 ± 1.61	31.97 ± 0.68	-13.98 ± 1.04	66.01 ± 2.31	-22.09 ± 3.13	53.61 ± 0.98	-5.04 ± 0.37
FedPOCL	47.09 ± 2.29	-10.17 ± 1.43	42.01 ± 2.40	-12.42 ± 1.40	75.00 ± 2.51	$\mathbf{-16.47 \pm 0.87}$	56.31 ± 3.09	-5.07 ± 1.19
FedAGC	41.81 ± 0.73	-12.78 ± 0.29	33.81 ± 3.25	-10.99 ± 2.51	62.71 ± 2.90	-21.49 ± 2.20	52.74 ± 1.88	-3.39 ± 0.78
FedBECAME	42.13 ± 1.77	-11.98 ± 2.37	30.90 ± 2.08	-21.87 ± 2.08	63.89 ± 1.90	-22.37 ± 4.31	51.92 ± 3.21	-5.09 ± 1.94
C-FLAG	46.82 ± 2.71	-13.05 ± 2.73	36.75 ± 1.83	-11.99 ± 2.32	69.21 ± 1.36	-19.80 ± 2.44	53.10 ± 1.97	-3.92 ± 1.38
FEDGRIT (ours)	$\mathbf{54.66 \pm 0.15}$	$\mathbf{-6.07 \pm 0.29}$	$\mathbf{43.31 \pm 0.89}$	$\mathbf{-6.92 \pm 0.93}$	$\mathbf{80.12 \pm 0.93}$	-18.56 ± 0.68	$\mathbf{57.03 \pm 0.46}$	$\mathbf{-2.28 \pm 0.23}$

Corollary 1 makes explicit the role of persistent client heterogeneity. For a fixed nonzero β, FEDGRIT converges to a heterogeneity-dependent stationarity neighborhood, with the remaining optimization and stochastic terms decaying as $\mathcal{O}(T^{-1/2})$ or faster. Under the additional condition $\beta = \mathcal{O}(\eta_L)$, this residual also vanishes, yielding an overall $\mathcal{O}(T^{-1/2})$ weighted task-uniform stationarity rate. This has the same dependence on the number of communication rounds as nonconvex federated multi-objective optimization guarantees (Yang et al., 2023; Askin et al., 2025), although those results consider simultaneous access to the objectives and use different stationarity criteria.

5 EXPERIMENTS

We evaluate FEDGRIT on class- and domain-incremental benchmarks and examine its performance and contributions of shared historical-gradient reconstruction and adaptive task weighting. All experiments are implemented in PyTorch with two NVIDIA Quadro RTX 8000 GPUs. Additional sensitivity studies of task weights, curvature approximations, and runtime are provided in Appendix F.

Setup. We use CIFAR-100 (Krizhevsky et al., 2009) and ImageNet100 (Russakovsky et al., 2015) for class-incremental learning, each divided into five tasks of 20 classes. For domain-incremental learning, Digit-10 and VLCS (Torralba & Efros, 2011) each comprise four sequential domain tasks. Unless otherwise stated, we use $M = 16$ clients, $N = 8$ participants per round, and $T = 50$ communication rounds per task. Each participating client performs three local epochs per round. Client data follow Dirichlet partitions with $\alpha = 1$ for the class-incremental benchmarks and $\alpha = 0.5$ for the domain-incremental benchmarks. We use ResNet-18 (He et al., 2016), batch size 32, learning rates $\eta_L = 0.001$ and $\eta_G = 1$, and a private replay buffer of 400 examples per client. Main results are averaged over three seeds, and component and sensitivity studies use CIFAR-100. Dataset construction, implementation details, and baseline adaptations are described in Appendix F.2.

Baselines. We compare with conventional FL methods FedAvg (McMahan et al., 2017) and FedProx (Li et al., 2020); multi-objective FL methods FSMGDA (Yang et al., 2023) and FedC-MOO (Askin et al., 2025), adapted to sequential training with replay; and FCL methods FLwF-2T (Usmanova et al., 2022), FedSSI (Li et al., 2025b), FedAGC (Zhang et al., 2025), SPECIAL (Xu & Li, 2026), and C-FLAG (Keshri et al., 2025). We also evaluate our federated adaptations of POCL (Wu et al., 2024b) and BECAME (Li et al., 2025a), denoted FedPOCL and FedBECAME.

Metrics. Let $A_{i,j}$ denote test accuracy on task j after learning task i. We report final average accuracy (ACC) and backward transfer (BWT):

$$\mathrm{ACC} = \frac{1}{K}\sum_{j=1}^{K} A_{K,j}, \qquad \mathrm{BWT} = \frac{1}{K-1}\sum_{j=1}^{K-1}(A_{K,j} - A_{j,j}). \tag{14}$$

Higher values are better for both metrics, and negative BWT indicates an average decline on earlier tasks relative to their performance immediately after learning. ACC is reported in percent, while BWT and differences between reported accuracies are measured in percentage points (pp).

Main results. Table 1 shows that FEDGRIT achieves the highest mean final accuracy on all four benchmarks. Compared with FedPOCL, the strongest baseline by mean ACC on each dataset, the gains are 7.57 and 1.30 pp on CIFAR-100 and ImageNet100, and 5.12 and 0.72 pp on Digit-10 and

Table 2: Historical-gradient source and weighting on CIFAR-100. Local replay uses each client's retained data at its current model; shared reconstruction uses the aggregated gradients and curvature summaries of FEDGRIT. Fixed weights are shared across clients, whereas adaptive weights are recomputed during local training. The final row is the full FEDGRIT method.

Historical gradients	Weighting	ACC (%) ↑	BWT (pp) ↑
Local replay	Fixed	35.27	−17.18
Local replay	Adaptive	47.57	−12.69
Shared reconstruction	Fixed	39.54	−11.84
Shared reconstruction	Adaptive	**54.81**	**−6.04**

VLCS. The improvements extend across both incremental settings, with the largest gains on CIFAR-100 and Digit-10. FEDGRIT also achieves the highest BWT on both class-incremental benchmarks and VLCS. On Digit-10, however, FedPOCL has better BWT (−16.47 pp vs. −18.56 pp), despite its lower final accuracy (75.00% vs. 80.12%). These metrics capture complementary aspects of continual learning: *final accuracy measures performance after learning the task sequence, whereas BWT measures changes relative to each method's own performance immediately after learning an earlier task.* Both methods adapt task weights, but our federated POCL implementation uses local replay, whereas FEDGRIT uses shared reconstructed history. Table 2 further examines the contributions of the historical-gradient source and adaptive weighting.

Historical-gradient source and adaptive weighting. Table 2 compares fixed and adaptive weights using either local replay gradients or shared reconstructed gradients. Adaptive weighting improves ACC by 12.30 pp with local replay and by 15.27 pp with shared reconstruction. Conversely, replacing local replay with shared reconstruction improves ACC by 4.27 pp under fixed weights and by 7.24 pp under adaptive weights. *Both components therefore contribute, and the gain from shared reconstruction is larger when weights are adapted.*

The full method also achieves the highest BWT among these variants, improving it by 6.65 pp over adaptive weighting with local replay. Together with the corresponding 7.24 pp ACC gain, this result shows that *adaptive weighting benefits from using shared reconstructed gradients rather than relying on each client's buffer.*

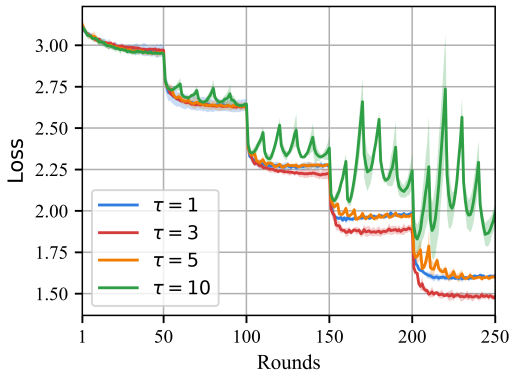


Figure 2: Loss of FEDGRIT on CIFAR-100 for curvature-refresh interval τ. Historical-task curvature summaries are refreshed every τ communication rounds.

Curvature refresh frequency. Figure 2 shows pronounced loss oscillations when the refresh interval reaches 10 rounds, while shorter intervals are more stable. This illustrates the sensitivity to outdated curvature information, and evaluating the overall efficiency trade-off additionally requires communication and runtime measurements.

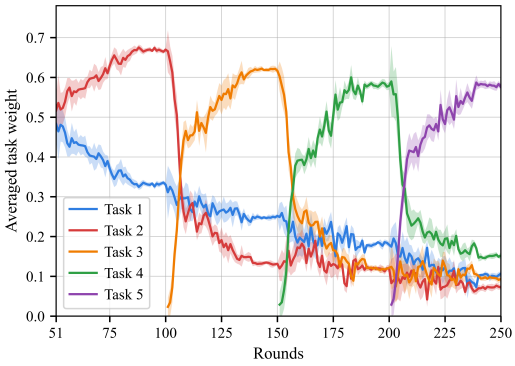


Figure 3: Evolution of averaged task weights in FEDGRIT on CIFAR-100 over communication rounds.

Adaptive task weights. Figure 3 illustrates how FEDGRIT reallocates task weights during sequential training. Within the displayed stages, the current-task weight generally increases and exceeds the individual historical-task weights later in training. Its weight drops when the next task arrives, while historical-task weights continue to evolve without a fixed ordering. These trajectories illustrate adaptive allocation both between current and historical tasks and among historical tasks.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, software implementation and debugging, literature identification, reference formatting, and feedback on the presentation and evaluation of experimental comparisons. We did not use generative AI tools to produce research data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research investigates knowledge retention and weighted task-uniform convergence in Federated Continual Learning settings. By reconstructing previous-task gradients through compact curvature-information exchange and adaptively tuning task weights during local optimization, FedGRIT enables each client to determine an effective update direction that balances the learning of new knowledge and the retention of previously acquired knowledge. Our framework focuses on the optimization and information-sharing challenges in FCL rather than the inherent structures and potential risks of deep neural networks. Therefore, FEDGRIT should be deployed with appropriate backbone models and adapted to the specific requirements of different applications and environments.

REPRODUCIBILITY STATEMENT

All experimental configurations are detailed in Section 5 and Appendix F, and all assumptions and theoretical proofs are listed in Section 4 and Appendix E. The source code and data used for our experiments will be made publicly available upon publication.

REFERENCES

Baris Askin, Pranay Sharma, Gauri Joshi, and Carlee Joe-Wong. Federated communication-efficient multi-objective optimization. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Sara Babakniya, Zalan Fabian, Chaoyang He, Mahdi Soltanolkotabi, and Salman Avestimehr. A data-free approach to mitigate catastrophic forgetting in federated class incremental learning for vision tasks. *Advances in Neural Information Processing Systems*, 36:66408–66425, 2023.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning, 2019.

Tyler Chen, Thomas Trogdon, and Shashanka Ubaru. Analysis of stochastic lanczos quadrature for spectrum approximation. In *International Conference on Machine Learning*, pp. 1728–1739. PMLR, 2021.

Erik Daxberger, Siddharth Swaroop, Kazuki Osawa, Rio Yokota, Richard E Turner, José Miguel Hernández-Lobato, and Mohammad Emtiyaz Khan. Improving continual learning by accurate gradient reconstructions of the past. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Seungyub Han, Yeongmo Kim, Taehyun Cho, and Jungwoo Lee. On the convergence of continual learning with adaptive methods. In *Uncertainty in Artificial Intelligence*, pp. 809–818. PMLR, 2023.

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

Satish Kumar Keshri, Nazreen Shah, and Ranjitha Prasad. On the convergence of continual federated learning using incrementally aggregated gradients. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 5068–5076. PMLR, 03–05 May 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwińska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of research of the National Bureau of Standards*, 45(4): 255–282, 1950.

Mei Li, Yuxiang Lu, Qinyan Dai, Suizhi Huang, Yue Ding, and Hongtao Lu. BECAME: Bayesian continual learning with adaptive model merging. In *Forty-second International Conference on Machine Learning*, 2025a.

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

Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12820–12829, 2024.

Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. FedSSI: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *Forty-second International Conference on Machine Learning*, 2025b.

Yichen Li, Haozhao Wang, Hang Su, Yulong Li, Xiaoquan Yi, Yankai Jiang, Chuang Zhao, Imran Razzak, and Ruixuan Li. Cross-task calibration for asynchronous federated continual learning. In *Forty-third International Conference on Machine Learning*, 2026a.

Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Converge faster, talk less: Hessian-informed federated zeroth-order optimization. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 40, pp. 2935–2947, 2017. doi: 10.1109/TPAMI.2017.2773081.

Yan-Shuo Liang and Wu-Jun Li. Adaptive plasticity improvement for continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7816–7825, 2023.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Beyond not-forgetting: Continual learning with backward knowledge transfer. *Advances in Neural Information Processing Systems*, 35:16165–16177, 2022.

Hao Liu and Huaping Liu. Continual learning with recursive gradient optimization. In *International Conference on Learning Representations*, 2022.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6467–6476, 2017.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

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

Subarnaduti Paul, Lars-Joel Frey, Roshni Ramanna Kamath, Kristian Kersting, and Martin Mundt. Masked autoencoders are efficient continual federated learners. In Vincenzo Lomonaco, Stefano Melacci, Tinne Tuytelaars, Sarath Chandar, and Razvan Pascanu (eds.), *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 70–85. PMLR, 29 Jul–01 Aug 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017.

Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting. *Advances in Neural Information Processing Systems*, 31, 2018.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Donald Shenaj, Marco Toldo, Alberto Rigon, and Pietro Zanuttigh. Asynchronous federated continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5055–5063, 2023. doi: 10.1109/CVPRW59228.2023.00534.

Haizhou Shi and Hao Wang. A unified approach to domain incremental learning with memory: Theory and algorithm. *Advances in Neural Information Processing Systems*, 36:15027–15059, 2023.

Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-iid data, 2019.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Anastasiia Usmanova, François Portet, Philippe Lalanda, and German Vega. Federated continual learning through distillation in pervasive computing. In *2022 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 86–91. IEEE, 2022.

Chenshen Wu, Luis Herranz, Xialei Liu, Joost Van De Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in neural information processing systems*, 31, 2018.

Hongda Wu and Ping Wang. Fast-convergent federated learning with adaptive weighting. *IEEE Transactions on Cognitive Communications and Networking*, 7(4):1078–1088, 2021.

Yichen Wu, Long-Kai Huang, Renzhen Wang, Deyu Meng, and Ying Wei. Meta continual learning revisited: Implicitly enhancing online hessian approximation via variance reduction. In *The Twelfth International Conference on Learning Representations*, 2024a.

Yichen Wu, Hong Wang, Peilin Zhao, Yefeng Zheng, Ying Wei, and Long-Kai Huang. Mitigating catastrophic forgetting in online continual learning by modeling previous task interrelations via pareto optimization. In *Forty-first international conference on machine learning*, 2024b.

Abudukelimu Wuerkaixi, Sen Cui, Jingfeng Zhang, Kunda Yan, Bo Han, Gang Niu, Lei Fang, Changshui Zhang, and Masashi Sugiyama. Accurate forgetting for heterogeneous federated continual learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Longtao Xu and Jian Li. Server-proximal aggregation for federated domain-incremental learning under partial participation: Task-uniform convergence and backward transfer. In *Forty-third International Conference on Machine Learning*, 2026.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-IID federated learning. In *International Conference on Learning Representations*, 2021.

Haibo Yang, Zhuqing Liu, Jia Liu, Chaosheng Dong, and Michinari Momma. Federated multi-objective learning. *Advances in neural information processing systems*, 36:39602–39625, 2023.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, pp. 12073–12086. PMLR, 2021.

Chengchao Zhang, Fanhua Shang, Hongying Liu, Liang Wan, and Wei Feng. Fedagc: Federated continual learning with asymmetric gradient correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3841–3850, 2025.

Zhao Zhang, Yong Zhang, Da Guo, Shuang Zhao, and Xiaolin Zhu. Communication-efficient federated continual learning for distributed learning system with non-iid data. *Science China Information Sciences*, 66(2):122102, 2023.

A Related Work

A.1 Continual Learning

Continual Learning aims to enable a model to sequentially learn knowledge from evolving tasks while mitigating catastrophic forgetting of previously learned knowledge. Existing approaches can be categorized into replay-based, regularization-based, and architecture-based methods. Replay-based methods maintain a limited memory of selected previous samples and jointly incorporate them with the current data during training (Chaudhry et al., 2019; Rebuffi et al., 2017; Wu et al., 2018). Regularization-based methods preserve previous predictions through knowledge distillation and keep the model not to deviate from the previous optimal model (Kirkpatrick et al., 2017; Li & Hoiem, 2017; Ritter et al., 2018; Li et al., 2025a). Architecture-based methods mitigate catastrophic forgetting by expanding the network or task-specific parameters to isolate the information learned from different tasks (Mallya & Lazebnik, 2018; Liang & Li, 2023).

A.2 Federated Continual Learning

Federated Continual Learning can be regarded as an expansion of conventional continual learning in the federated setting, which enables clients to sequentially train the model cooperatively, so the main types of FCL can also be categorized into these three types (Yoon et al., 2021; Usmanova et al., 2022; Babakniya et al., 2023; Wuerkaixi et al., 2024). However, a key distinction between the two settings lies in information communication, as the communication between clients and the server in FCL needs to account for limited bandwidth and communication costs (Wu & Wang, 2021; Zhang et al., 2023; Li et al., 2024; Paul et al., 2025). In addition, FCL introduces task-level asynchrony besides the communication asynchrony inherited from conventional FL, where different clients may simultaneously train on different tasks due to the heterogeneous task sequences of different clients. FedSpace addresses this asynchrony through prototype aggregation, representation alignment and a modified aggregation policy, enabling knowledge sharing among different clients (Shenaj et al., 2023), and C^2-AFCL introduces a two-stage orthogonal calibration mechanism to mitigate both client and task drift caused by asynchronous client updates (Li et al., 2026a).

A.3 Hessian-informed Continual Learning

Second-order information has been widely explored in CL to preserve previously acquired knowledge. For all Hessian-based methods, the problem which can not be overlooked is how to find the Hessian matrix. Since forming the full Hessian causes unnecessary computation and storage costs, recent studies focuses on using more efficient approximation approaches to recover the Hessian information while training. The Fisher information matrix is one of the most widely used approximation methods, whose diagonal entries are commonly employed to measure the importance of parameters, but it only captures parameter-wise information while ignoring the correlations among parameters, leading to substantial approximation errors (Kirkpatrick et al., 2017; Shoham et al., 2019; Liu & Liu, 2022). As an alternative to the Fisher information matrix, Daxberger et al. (2023) uses the generalized Gauss–Newton (GGN) matrix to provide a more accurate curvature approximation for previous-gradient reconstruction. Another challenge of utilizing Hessian information in CL setting is to prevent the model parameters from deviating significantly from the point at which the Hessian was computed. Wu et al. (2024a) and Li et al. (2026b) employ iterative strategies to update the Hessian to improve the accuracy of approximation by maintaining up-to-date curvature information while approximating. However, effectively maintaining and updating Hessian information in distributed learning settings remains largely underexplored.

B Notation List

The detailed notation is list in table

Table 3: Notations and Terminologies.

Notation	Description
M	Number of all clients
N	Number of participated clients per round
T	Number of global epochs
E	Number of local updates
K	Number of tasks
$[n]$	Set of integers $\{1, \ldots, n\}$
θ_i^t	Global model at round t in task i
$\theta_{i,m}^t$	Local model on client m at round t in task i
$\theta_{i,m}^{t,e}$	Local model on client m at epoch e of round t in task i
$w_{i,m}^{t,e,*}$	Tuned weight for the gradient of task i on client m of epoch e on round t
$\mathbf{d}_{i,m}^{t,e}$	Update direction on client m of epoch e on round t
$\Delta_{i,m}^t$	Update vector on client m of round t in task i
f_i	Loss function of task i
$f_{1:K}$	Loss function over task 1 to task K
∇f_i	the gradient of f_i
$\nabla \widehat{f}_{1:K}$	Weighted gradient over all tasks
$\mathcal{T}$	Set of all tasks, i.e., $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$
$\mathcal{T}_i$	Dataset of the task i
$\mathcal{T}_{i,m}$	Dataset of client m on task i
$\mathcal{S}_i^t$	The index of participants at round t in task i
$\boldsymbol{w}_m^{t,e,*}$	Set of tuned weights on client m of epoch e on round t
$\|\cdot\|^2$	ℓ_2-norms

C Algorithms

C.1 Lanczos Algorithm

Lanczos Algorithm is an iterative method for efficiently approximating the dominant eigenvalues and eigenvectors of a large symmetric matrix. The necessary for a matrix to be decomposed via Lanczos algorithm is the existence of a operator $\mathcal{H}(\boldsymbol{q}) : \boldsymbol{q} \mapsto \boldsymbol{H}\boldsymbol{q}$, and the automatic differentiation for $\nabla_\theta[\nabla_\theta \widehat{f}_{i,m}(\theta)^\top \boldsymbol{q}]\Big|_{\theta=\theta_K^t}$ enables the product of the Hessian matrix and any vector $\boldsymbol{q} \in \mathbb{R}^d$ to be computed, which guarantees the original Hessian matrix to be decomposed via Lanczos Algorithm.

Algorithm 2 Hessian-Lanczos Decomposition Algorithm

Input: Operator $\mathcal{H}(\boldsymbol{q}) : \boldsymbol{q} \mapsto \boldsymbol{H}\boldsymbol{q}$, target rank r, Lanczos steps $k \geq r$
Initialize: $\boldsymbol{q}_1 \in \mathbb{R}^{d\times 1}$ with $\|\boldsymbol{q}_1\| = 1$, $\boldsymbol{q}_0 = \boldsymbol{0} \in \mathbb{R}^{d\times 1}$, $\beta_0 = 0$
for $j = 1, \ldots, k$ **do**
  $\boldsymbol{w} \leftarrow \boldsymbol{H}\boldsymbol{q}_j - \beta_{j-1}\boldsymbol{q}_{j-1}$
  $\alpha_j \leftarrow \boldsymbol{q}_j^T \boldsymbol{w}$
  $\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha_j \boldsymbol{q}_j$
  Reorthogonalize $\boldsymbol{w}$ against $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_j\}$
  $\beta_j \leftarrow \|\boldsymbol{w}\|$, $\boldsymbol{q}_{j+1} \leftarrow \boldsymbol{w}/\beta_j$
end for
Construct $\boldsymbol{Q}_k = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k]$ and tridiagonal $\boldsymbol{T}_k$ from $\{\alpha_j, \beta_j\}$
Decompose $\boldsymbol{T}_k$ as $\boldsymbol{T}_k = \boldsymbol{V}\boldsymbol{\Theta}\boldsymbol{V}^T$
Select the top-r eigenpairs $(\boldsymbol{V}_r, \boldsymbol{\Theta}_r)$
Reconstruct $\widetilde{\boldsymbol{Q}} \leftarrow \boldsymbol{Q}_k \boldsymbol{V}_r$, $\widetilde{\boldsymbol{\Lambda}} \leftarrow \boldsymbol{\Theta}_r$
return $\widetilde{\boldsymbol{Q}}_{d\times r}, \widetilde{\boldsymbol{\Lambda}}_{r\times r}$

C.2 FRANK-WOLFE ALGORITHM

In order for the Frank-Wolfe Algorithm to be self-contained, we restate the algorithm in Wu et al. (2024b) here.

Algorithm 3 Frank-Wolfe Algorithm for Solving Eq. (5)

Input: The gradients set $\boldsymbol{G} = [\nabla \tilde{f}_1(\theta_{K,m}^{t,e})^\top, \ldots, \nabla \tilde{f}_{K-1}(\theta_{K,m}^{t,e})^\top, \nabla f_{K,m}(\theta_{K,m}^{t,e}; \xi_{K,m}^{t,e})^\top]$
Initialize: $\boldsymbol{w} = [\frac{1}{K}, \ldots, \frac{1}{K}]$
1: Compute $\boldsymbol{P} = \boldsymbol{G}^\top \boldsymbol{G}$
2: **repeat**
3: $\quad \boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha} \in \{\boldsymbol{\alpha}^\top \mathbf{1} = 1, \boldsymbol{\alpha} \succ 0\}} \boldsymbol{\alpha}^\top \boldsymbol{P} \boldsymbol{w}$
4: $\quad L = \boldsymbol{w} + \eta(\boldsymbol{\alpha} - \boldsymbol{w})$
5: $\quad \eta = \arg\min_{\eta \in [0,1]} L^\top \boldsymbol{P} L$
6: $\quad \boldsymbol{w} \leftarrow (1 - \eta)\boldsymbol{w} + \eta \boldsymbol{\alpha}$
7: **until** $\eta \sim 0$ or reaching the maximum iteration number
8: **return** The final computed weights $\boldsymbol{w}$

C.3 DISTRIBUTED SPECTRAL HESSIAN AGGREGATION ALGORITHM

Algorithm 4 Distributed Spectral Hessian Aggregation (DSHA)

Require: $\boldsymbol{A}_{K,i}^t$, $\boldsymbol{B}_{K,i}^t$, target rank r
1: Compute an economy QR factorization $\boldsymbol{A}_{K,i}^t = \boldsymbol{D}_{K,i}^t \boldsymbol{R}_{K,i}^t$.
2: Form $\boldsymbol{C}_{K,i}^t = \boldsymbol{R}_{K,i}^t \boldsymbol{B}_{K,i}^t (\boldsymbol{R}_{K,i}^t)^\top$.
3: Retain at most r eigenpairs of $\boldsymbol{C}_{K,i}^t$ with largest eigenvalue magnitudes in $(\boldsymbol{V}_{K,i}^t, \boldsymbol{\Gamma}_{K,i}^t)$, preserving the eigenvalue signs.
4: Set $\boldsymbol{U}_{K,i}^t = \boldsymbol{D}_{K,i}^t \boldsymbol{V}_{K,i}^t$.
5: **return** $(\boldsymbol{U}_{K,i}^t, \boldsymbol{\Gamma}_{K,i}^t)$.

D LEMMAS

Lemma 1. *Suppose f is twice differentiable and L-smooth, let $\boldsymbol{H} = \nabla^2 f(x)$, $\theta \in \mathbb{R}^d$, and $\tilde{\boldsymbol{H}} = \boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^T$ as the rank-r truncated eigendecomposition of $\boldsymbol{H}$, we have:*

$$\left\| \tilde{\boldsymbol{H}} \theta \right\|^2 \leq L^2 \|\theta\|^2 . \tag{15}$$

Proof. Since f is L-smooth and twice differentiable, then we have

$$\|\boldsymbol{H}\|_2 = \left\| \nabla^2 f(x) \right\|_2 \leq L. \tag{16}$$

Since $\boldsymbol{H}$ is symmetric, its spectral norm is given by the largest absolute eigenvalue as

$$\|\boldsymbol{H}\|_2 = \max_{i \in [d]} |\lambda_i(\boldsymbol{H})|. \tag{17}$$

Therefore, every eigenvalue $\lambda_i, i \in [r]$, retained in the diagonal elements of $\boldsymbol{V}$ satisfies

$$|\lambda_i| \leq L, \tag{18}$$

which means that $\|\boldsymbol{V}\|_2 = \max_{i \in [r]} |\lambda_i| \leq L$. Since $\boldsymbol{U}$ is orthonormal matrix, its columns are orthonormal, then we have $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_r$, which means $\|\boldsymbol{U}\|_2 = \left\|\boldsymbol{U}^T\right\|_2 = 1$. Therefore, for any $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned} \left\| \tilde{\boldsymbol{H}} \theta \right\| &= \left\| \boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^T \theta \right\| \\ &\leq \|\boldsymbol{U}\|_2 \|\boldsymbol{V}\|_2 \|\boldsymbol{U}^T\|_2 \|\theta\| \end{aligned}$$

$$\leq L\|\theta\|, \tag{19}$$

and we can further have $\|\tilde{\boldsymbol{H}}\theta\|^2 \leq L^2\|\theta\|^2$, which completes the proof.

□

Lemma 2. *The curvature matrix $\tilde{\boldsymbol{H}}^t_{K,i}$ used by clients to reconstruct the gradients satisfies that $\|\tilde{\boldsymbol{H}}^t_{K,i}\theta\|^2 \leq L^2\|\theta\|^2, \forall\theta \in \mathbb{R}^d$.*

Proof. According to Lemma 1, the updated curvature information $(\tilde{\boldsymbol{Q}}^t_{K,i,m}, \tilde{\boldsymbol{\Lambda}}^t_{K,i,m})$ satisfies $\|\tilde{\boldsymbol{Q}}^t_{K,i,m}\tilde{\boldsymbol{\Lambda}}^t_{K,i,m}(\tilde{\boldsymbol{Q}}^t_{K,i,m})^\top\|^2 \leq L^2$, then we have

$$\|\bar{\boldsymbol{H}}^t_{K,i}\|^2 = \left\|\frac{1}{N}\sum_{m\in\mathcal{S}^t_K}\tilde{\boldsymbol{Q}}^t_{K,i,m}\tilde{\boldsymbol{\Lambda}}^t_{K,i,m}(\tilde{\boldsymbol{Q}}^t_{K,i,m})^\top\right\|^2 \leq L^2, \tag{20}$$

which means that the ℓ^2-norm of recovered Hessian matrix is still bounded by L^2. Therefore, according to Lemma 1, we also have

$$\left\|\tilde{\boldsymbol{H}}^t_{K,i}\theta\right\|^2 = \left\|\boldsymbol{U}^t_{K,i}\boldsymbol{\Gamma}^t_{K,i}(\boldsymbol{U}^t_{K,i})^\top\theta\right\|^2 \leq L^2\|\theta\|^2, \tag{21}$$

where $(\boldsymbol{U}^t_{K,i}, \boldsymbol{\Gamma}^t_{K,i})$ are the results of truncated EVD of $\boldsymbol{H}^t_{K,i}$ with rank-r. □

Lemma 3. *If the local learning rate satisfies $\eta_L \leq \frac{1}{3(k-1)EL}, \forall k \geq 2$, we have the following results:*

$$\left\|\theta^{t,e}_{k,m} - \theta^t_k\right\|^2 \leq 20\eta_L^2E^2D^2. \tag{22}$$

Proof.

$$\begin{aligned}
&\left\|\theta^{t,e}_{k,m} - \theta^t_K\right\|^2 \\
=&\left\|\theta^{t,e-1}_{k,m} - \theta^t_K - \eta_L\left(\sum_{i=1}^{k-1} w^{t,e-1,*}_{i,m}\left(\nabla f_i(\theta^t_k) + \tilde{\boldsymbol{H}}_i\left(\theta^{t,e-1}_{k,m} - \theta^t_K\right)\right)\right.\right. \\
&\left.\left.+w^{t,e,*}_{K,m}\nabla f_{k,m}(\theta^{t,e-1}_{k,m};\xi^{t,e-1}_{k,m})\right)\right\|^2 \\
=&\left(1+\frac{1}{2E-1}\right)\left\|\theta^{t,e-1}_{k,m} - \theta^t_k\right\|^2 + 4E\eta_L^2\left\|\sum_{i=1}^{k} w^{t,e-1,*}_{i,m}\nabla f_i(\theta^t_K)\right\|^2 \\
&+4E\eta_L^2\left\|\sum_{i=1}^{k-1} w^{t,e-1,*}_{k,m}\tilde{\boldsymbol{H}}_i\left(\theta^{t,e-1}_{k,m} - \theta^t_k\right)\right\|^2 \\
\leq&\left(1+\frac{1}{2E-1}\right)\left\|\theta^{t,e-1}_{k,m} - \theta^t_k\right\|^2 + 4\eta_L^2ED^2 + 4\eta_L^2E\left(k-1\right)^2L^2\left\|\theta^{t,e-1}_{k,m} - \theta^t_k\right\|^2 \\
=&\left(1+\frac{1}{2E-1}+4\eta_L^2E\left(k-1\right)^2L^2\right)\left\|\theta^{t,e-1}_{k,m} - \theta^t_k\right\|^2 + 4\eta_L^2ED^2.
\end{aligned} \tag{23}$$

When $\eta_L \leq \frac{1}{3(k-1)EL}$, we have $1+\frac{1}{2E-1}+4\eta_L^2E\left(k-1\right)^2L^2 \leq 1+\frac{1}{E-1}$. Then, we have

$$\left\|\theta^{t,e}_{k,m} - \theta^t_k\right\|^2 \leq \left(1+\frac{1}{E-1}\right)\left\|\theta^{t,e-1}_{k,m} - \theta^t_k\right\|^2 + 4\eta_L^2EK^2D^2. \tag{24}$$

Unrolling the recursion, we have

$$\left\|\theta^{t,e}_{k,m} - \theta^t_k\right\|^2 \leq \sum_{\tau=0}^{e-1}\left(1+\frac{1}{E-1}\right)^\tau\left(4\eta_L^2ED^2\right) \leq 20\eta_L^2E^2D^2. \tag{25}$$

□

Lemma 4. *Assume we randomly select N clients from all M clients, and the set $\mathcal{S}$ is the set of the client index. If $\varepsilon_n \in \mathbb{R}^d, n \in [M]$ satisfies $\mathbb{E}\left\|\varepsilon_n\right\|^2 \leq \sigma^2$ and $\sum_{n=1}^M \varepsilon_n = 0$, then*

- *if $\mathbb{E}\left[\varepsilon_n\right] = 0, \forall n \in [M]$, we have:*

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n\in\mathcal{S}}\varepsilon_n\right\|^2 \leq \frac{\sigma^2}{N}. \tag{26}$$

- *if $\mathbb{E}\left[\varepsilon_n\right] \neq 0, \forall n \in [M]$, we have:*

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n\in\mathcal{S}}\varepsilon_n\right\|^2 \leq \frac{M-N}{N\left(M-1\right)}\sigma^2. \tag{27}$$

Proof. Since $\sum_{n\in\mathcal{S}}\varepsilon_n = \sum_{n=1}^M \mathbb{I}\{n\in\mathcal{S}\}\varepsilon_n$, we can expand as:

$$\begin{aligned}
\mathbb{E}\left\|\frac{1}{N}\sum_{n\in\mathcal{S}}\varepsilon_n\right\|^2 =& \mathbb{E}\left\|\frac{1}{N}\sum_{n=1}^M \mathbb{I}\{n\in\mathcal{S}\}\varepsilon_n\right\|^2 \\
=& \frac{1}{N^2}\sum_{n=1}^M \mathbb{E}\left[\mathbb{I}\{n\in\mathcal{S}\}\right]\mathbb{E}\left\|\varepsilon_n\right\|^2 + \frac{1}{N^2}\sum_{m\neq n}\mathbb{E}\left[\mathbb{I}\{m,n\in\mathcal{S}\}\right]\mathbb{E}\left\langle\varepsilon_n,\varepsilon_m\right\rangle \\
=& \frac{1}{MN}\sum_{n=1}^M \mathbb{E}\left\|\varepsilon_n\right\|^2 + \frac{N-1}{MN\left(M-1\right)}\sum_{m\neq n}\mathbb{E}\left\langle\varepsilon_n,\varepsilon_m\right\rangle
\end{aligned} \tag{28}$$

Since $\sum_{n=1}^M \varepsilon_n = 0$, then we have

$$\left\|\sum_{n=1}^M \varepsilon_n\right\|^2 = \left\langle\sum_{n=1}^M \varepsilon_n, \sum_{m=1}^M \varepsilon_m\right\rangle = \sum_{n=1}^M\left\|\varepsilon_n\right\|^2 + \sum_{m\neq n}\left\langle\varepsilon_n,\varepsilon_m\right\rangle = 0, \tag{29}$$

where we can get the relation that $\sum_{m\neq n}\left\langle\varepsilon_n,\varepsilon_m\right\rangle = -\sum_{n=1}^M\left\|\varepsilon_n\right\|^2$. Then, we have:

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n\in\mathcal{S}}\varepsilon_n\right\|^2 = \left(\frac{1}{MN} - \frac{N-1}{MN\left(M-1\right)}\right)\sum_{n=1}^M\mathbb{E}\left\|\varepsilon_n\right\|^2 \tag{30}$$

$$= \frac{M-N}{MN\left(M-1\right)}\sum_{n=1}^M\mathbb{E}\left\|\varepsilon_n\right\|^2 \tag{31}$$

$$\leq \frac{M-N}{N\left(M-1\right)}\sigma^2. \tag{32}$$

If $\mathbb{E}\left[\varepsilon_n\right] = 0, \forall n \in [M]$, we have:

$$\mathbb{E}\left\|\frac{1}{N}\sum_{n\in\mathcal{S}}\varepsilon_n\right\|^2 = \frac{1}{MN}\sum_{n=1}^M\mathbb{E}\left\|\varepsilon_n\right\|^2 \leq \frac{1}{N}\sigma^2. \tag{33}$$

This completes the proof. □

Lemma 5. *By denoting $\bar{\mathbf{d}}_{K,m}^{t,e} = \mathbb{E}_t[\tilde{\mathbf{d}}_{K,m}^{t,e}]$, we can bound the norm of the difference between $\tilde{\mathbf{d}}_{K,m}^{t,e}$ and $\mathbb{E}_t[\mathbf{d}_{K,m}^{t,e}]$ as*

$$\mathbb{E}_t\left\|\tilde{\mathbf{d}}_{K,m}^{t,e} - \bar{\mathbf{d}}_{K,m}^{t,e}\right\|^2 \leq \sigma_L^2 + \frac{M-N}{N(M-1)}\beta\sigma_G^2. \tag{34}$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} - \bar{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
=&\mathbb{E}_t \left\| \sum_{i=1}^{K-1} w_{i,m}^{t,e,*} \left(\frac{1}{N} \sum_{n \in \mathcal{S}_K^t} \nabla f_{i,n} \left(\theta_K^t; \xi_{i,n}^t \right) + \tilde{\boldsymbol{H}}_i^t \left(\theta_{K,m}^{t,e} - \theta_{K,m}^t \right) \right) + w_{K,m}^{t,e,*} \nabla f_{K,m} \left(\theta_{K,m}^{t,e}; \xi_{K,m}^{t,e} \right) \right. \\
&\left. - \sum_{i=1}^{K-1} w_{i,m}^{t,e,*} \left(\nabla f_i \left(\theta_K^t \right) + \tilde{\boldsymbol{H}}_i^t \left(\theta_{K,m}^{t,e} - \theta_{K,m}^t \right) \right) - w_{K,m}^{t,e,*} \nabla f_{K,m} \left(\theta_{K,m}^{t,e} \right) \right\|^2 \\
=&\mathbb{E}_t \left\| \sum_{i=1}^{K-1} w_{i,m}^{t,e,*} \left(\frac{1}{N} \sum_{n \in \mathcal{S}_K^t} \left(\nabla f_{i,n} \left(\theta_K^t; \xi_{i,n}^t \right) - \nabla f_{i,n} \left(\theta_K^t \right) + \nabla f_{i,n} \left(\theta_K^t \right) - \nabla f_i \left(\theta_K^t \right) \right) \right) \right. \\
&\left. + w_{K,m}^{t,e,*} \left(\nabla f_{K,m} \left(\theta_{K,m}^{t,e}; \xi_{K,m}^{t,e} \right) - \nabla f_{K,m} \left(\theta_{K,m}^{t,e} \right) \right) \right\|^2 . \qquad (35)
\end{aligned}
$$

Since

$$
\begin{aligned}
&\mathbb{E}_t \left\| \frac{1}{N} \sum_{n \in \mathcal{S}_K^t} \left(\nabla f_{i,n} \left(\theta_K^t; \xi_{i,n}^t \right) - \nabla f_{i,n} \left(\theta_K^t \right) + \nabla f_{i,n} \left(\theta_K^t \right) - \nabla f_i \left(\theta_K^t \right) \right) \right\|^2 \\
\leq &\mathbb{E}_t \left\| \frac{1}{N} \sum_{n \in \mathcal{S}_K^t} \left(\nabla f_{i,n} \left(\theta_K^t; \xi_{i,n}^t \right) - \nabla f_{i,n} \left(\theta_K^t \right) \right) \right\|^2 \\
&+ \mathbb{E}_t \left\| \frac{1}{N} \sum_{n \in \mathcal{S}_K^t} \left(\nabla f_{i,n} \left(\theta_K^t \right) - \nabla f_i \left(\theta_K^t \right) \right) \right\|^2 \\
\leq &\frac{\sigma_L^2}{N} + \frac{M-N}{N(M-1)} \beta \sigma_G^2, \qquad (36)
\end{aligned}
$$

where the second inequality holds due to Lemma 4 based on the truth that

- we have $\mathbb{E}_t \left[\nabla f_{i,n} \left(\theta_K^t; \xi_{i,n}^t \right) - \nabla f_{i,n} \left(\theta_K^t \right) \right] = 0, \mathbb{E}_t \| \nabla f_{i,n} (\theta_K^t; \xi_{i,n}^t) - \nabla f_{i,n} (\theta_K^t) \|^2 \leq \sigma_L^2, \forall n \in \mathcal{S}_K^t$ based on Assumption 4.
- we have $\mathbb{E}_t \left[\nabla f_{i,n} \left(\theta_K^t \right) - \nabla f_i \left(\theta_K^t \right) \right] \neq 0, \mathbb{E}_t \| \nabla f_{i,n} \left(\theta_K^t \right) - \nabla f_i \left(\theta_K^t \right) \|^2 \leq \beta \sigma_G^2, \forall n \in \mathcal{S}_K^t$ based on Assumption 5.

Therefore, we have

$$
\mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} - \bar{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \leq \sigma_L^2 + \frac{M-N}{N(M-1)} \beta \sigma_G^2. \qquad (37)
$$

□

E Complete Proofs

E.1 Proof of Theorem 1

Proof. We can expand $\mathbb{E}_t \left[f_{1:K-1} \left(\theta_K^t \right) \right]$ as

$$
\begin{aligned}
&\mathbb{E}_t \left[f_{1:K-1} \left(\theta_K^t \right) \right] \\
\leq &f_{1:K-1} \left(\theta_K^0 \right) + \left\langle \nabla f_{1:K-1} \left(\theta_K^0 \right), \mathbb{E}_t \left[\theta_K^t - \theta_K^0 \right] \right\rangle + \frac{(K-1) L}{2} \mathbb{E}_t \left\| \theta_K^t - \theta_K^0 \right\|^2 \\
= &f_{1:K-1} \left(\theta_K^0 \right) - \eta_G \eta_L \left\langle \nabla f_{1:K-1} \left(\theta_K^0 \right), \mathbb{E}_t \left[\frac{1}{N} \sum_{\tau=0}^{t-1} \sum_{m \in \mathcal{S}_K^\tau} \sum_{e=0}^{E-1} \tilde{\mathbf{d}}_{K,m}^{\tau,e} \right] \right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{(K-1)L}{2}\eta_G^2\eta_L^2\mathbb{E}_t\left\|\frac{1}{N}\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\tilde{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\\
=&f_{1:K-1}\left(\theta_K^0\right)\underbrace{-\eta_G\eta_L\left\langle\nabla f_{1:K-1}\left(\theta_K^0\right),\frac{1}{N}\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\rangle}_{T_1}\\
&\underbrace{+\frac{(K-1)L}{2}\eta_G^2\eta_L^2\mathbb{E}_t\left\|\frac{1}{N}\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\tilde{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2}_{T_2}.
\end{aligned}
\tag{38}
$$

Term T_1 can be bounded by

$$
\begin{aligned}
T_1=&-\eta_G\eta_L\left\langle\nabla f_{1:K-1}\left(\theta_K^0\right),\frac{1}{M}\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\rangle\\
\leq&-\frac{\eta_G\eta_L}{M}\cdot\epsilon\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|\cdot\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|\right].
\end{aligned}
\tag{39}
$$

Term T_2 can be bounded as

$$
\begin{aligned}
T_2\leq&\frac{(K-1)L}{2}\cdot\frac{\eta_G^2\eta_L^2Et}{N}\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\tilde{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\right]\\
\leq&\underbrace{\frac{(K-1)L}{2}\cdot\frac{\eta_G^2\eta_L^2Et}{N}\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\right]}_{T_{21}}\\
&+\underbrace{\frac{(K-1)L}{2}\cdot\frac{\eta_G^2\eta_L^2Et}{N}\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}-\tilde{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\right]}_{T_{22}}.
\end{aligned}
\tag{40}
$$

Since $\eta_L\leq\frac{2\epsilon\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|}{DLEt\sqrt{7tEN}}$ and $\eta_G\leq\frac{1}{K-1}$, we have

$$
\begin{aligned}
T_{21}=&\frac{\eta_G\eta_L}{N}\cdot(K-1)\eta_G\cdot\frac{\eta_LLEt}{2}\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\right]\\
\leq&\frac{\eta_G\eta_L}{N}\cdot\frac{\epsilon\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|}{D\sqrt{7tEN}}\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\right]\\
\leq&\frac{\eta_G\eta_L}{N}\cdot\epsilon\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|\cdot\frac{\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2}{\sqrt{\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2}}\\
\leq&\frac{\eta_G\eta_L}{N}\cdot\epsilon\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|\cdot\sqrt{\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2}\\
\leq&\frac{\eta_G\eta_L}{N}\cdot\epsilon\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|\cdot\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|\right],
\end{aligned}
\tag{41}
$$

where the second inequality holds due to the truth that

$$\begin{aligned}
&\mathbb{E}_t \left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2 \\
=&\mathbb{E}_t \left\|\sum_{i=1}^{K-1} w_{i,m}^{t,e,*}\left(\nabla f_i\left(\theta_K^t\right)+\tilde{\boldsymbol{H}}_i^t\left(\theta_{K,m}^{\tau,e}-\theta_{K,m}^t\right)\right)+w_{K,m}^{t,e,*}\nabla f_{K,m}\left(\theta_{K,m}^{\tau,e}\right)\right\|^2 \\
\leq&2\mathbb{E}_t \left\|\sum_{i=1}^{K-1} w_{i,m}^{t,e,*}\nabla f_i\left(\theta_K^t\right)+w_{K,m}^{t,e,*}\nabla f_{K,m}\left(\theta_{K,m}^{\tau,e}\right)\right\|^2 \\
&+2\mathbb{E}_t \left\|\sum_{i=1}^{K-1} w_{i,m}^{t,e,*}\tilde{\boldsymbol{H}}_i^\tau\left(\theta_{K,m}^{\tau,e}-\theta_{K,m}^\tau\right)\right\|^2 \\
\leq&2D^2+2L^2\cdot 20\eta_L^2E^2D^2 \\
\leq&2D^2+\frac{40D^2}{9\left(K-1\right)^2} \\
\leq&7D^2,
\end{aligned} \tag{42}$$

where the second inequality comes from Assumption 1 and Lemma 3, and the third inequality holds when $\eta_L \leq \frac{1}{3(K-1)EL}$.

According to Lemma 5, term T_{22} can be bounded as

$$\begin{aligned}
T_{22} =&\frac{\left(K-1\right)L}{2}\cdot\frac{\eta_G^2\eta_L^2Et}{N}\left[\sum_{\tau=0}^{t-1}\sum_{m\in\mathcal{S}_K^\tau}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\bar{\mathbf{d}}_{K,m}^{\tau,e}-\tilde{\mathbf{d}}_{K,m}^{\tau,e}\right\|^2\right] \\
\leq&\frac{2\epsilon^2\left\|\nabla f_{1:K-1}\left(\theta_K^0\right)\right\|^2}{7\left(K-1\right)Et^2NLD^2}\left(\sigma_L^2+\frac{\left(M-N\right)\beta\sigma_G^2}{N\left(M-1\right)}\right).
\end{aligned} \tag{43}$$

Therefore, we have

$$\mathbb{E}_t\left[f_{1:K-1}\left(\theta_K^t\right)\right] \leq f_{1:K-1}\left(\theta_K^0\right)+\frac{2\epsilon^2\left\|\nabla f_{1:K}\left(\theta_K^0\right)\right\|^2}{7\left(K-1\right)Et^2NLD^2}\left(\sigma_L^2+\frac{\left(M-N\right)\beta\sigma_G^2}{N\left(M-1\right)}\right). \tag{44}$$

□

E.2 Proof of Theorem 2

Proof. According to the update rule, we have

$$\theta_K^{t+1}=\theta_K^t+\frac{\eta_G}{N}\sum_{m\in\mathcal{S}_K^t}\Delta_{K,m}^t=\theta_K-\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbf{d}_{K,m}^{t,e}, \tag{45}$$

where $\mathbf{d}_{K,m}^{t,e}=\sum_{i=1}^{K-1}w_{i,m}^{t,e,*}(\nabla f_i(\theta_K^t)+\tilde{\boldsymbol{H}}_i^t(\theta_{K,m}^{t,e}-\theta_K^t))+w_{K,m}^{t,e,*}\nabla f_{K,m}(\theta_{K,m}^{t,e};\xi_{K,m}^{t,e})$.

Due to the smoothness in Assumption 2, we have

$$\begin{aligned}
\mathbb{E}_t\left[f_{1:K}\left(\theta_K^{t+1}\right)\right] \leq& f_{1:K}\left(\theta_K^t\right)+\underbrace{\mathbb{E}_t\left\langle\nabla f_{1:K}\left(\theta_K^t\right),-\frac{\eta_G\eta_L}{M}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E}\mathbf{d}_{K,m}^{t,e}\right\rangle}_{T_3} \\
&+\frac{KL}{2}\underbrace{\mathbb{E}_t\left\|-\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E}\mathbf{d}_{K,m}^{t,e}\right\|^2}_{T_4}.
\end{aligned} \tag{46}$$

For term T_3, we have

$$\begin{aligned}
T_3 =&\mathbb{E}_t \left\langle \nabla f_{1:K}\left(\theta_K^t\right), -\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E}\left(\sum_{i=1}^{K-1} w_{i,m}^{t,e,*}\left(\nabla f_i(\theta_K^t)+\tilde{\boldsymbol{H}}_i^t\left(\theta_{K,m}^{t,e}-\theta_K^t\right)\right)\right.\right.\\
&\left.\left.+w_{K,m}^{t,e,*}\nabla f_{K,m}\left(\theta_{K,m}^{t,e};\xi_{K,m}^{t,e}\right)\right)\right\rangle\\
=&\mathbb{E}_t \left\langle \nabla f_{1:K}\left(\theta_K^t\right), -\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\left(\sum_{i=1}^{K-1} w_{i,m}^{t,e,*}\left(\nabla f_i(\theta_K^t)+\tilde{\boldsymbol{H}}_i^t\left(\theta_{K,m}^{t,e}-\theta_K^t\right)\right)\right.\right.\\
&\left.\left.+w_{K,m}^{t,e,*}\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right)\right)\right\rangle. \qquad (47)
\end{aligned}$$

By defining $\tilde{\mathbf{d}}_{K,m}^{t,e}=\sum_{i=1}^{K-1} w_{i,m}^{t,e,*}(\nabla f_i(\theta_K^t)+\tilde{\boldsymbol{H}}_i^t(\theta_{K,m}^{t,e}-\theta_K^t))+w_{K,m}^{t,e,*}\nabla f_{K,m}(\theta_{K,m}^{t,e})$, we have

$$\begin{aligned}
T_3 =&\mathbb{E}_t \left\langle \nabla f_{1:K}\left(\theta_K^t\right), -\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle\\
=&\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\langle \nabla f_{1:K}\left(\theta_K^t\right), -\tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle\\
=&\underbrace{\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\langle \nabla f_{1:K}\left(\theta_K^t\right)-\sum_{i=1}^{K-1}\nabla f_i\left(\theta_K^t\right)-\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right), -\tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle}_{T_{31}}\\
&\underbrace{-\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\langle \sum_{i=1}^{K-1}\nabla f_i\left(\theta_K^t\right)+\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right), \tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle}_{T_{32}}. \qquad (48)
\end{aligned}$$

For term T_{31}, we have

$$\begin{aligned}
T_{31} =&\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\langle \nabla f_{1:K}\left(\theta_K^t\right)-\sum_{i=1}^{K-1}\nabla f_i\left(\theta_K^t\right)-\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right), -\tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle\\
=&\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\langle \sum_{i=1}^{K}\nabla f_i\left(\theta_K^t\right)-\sum_{i=1}^{K-1}\nabla f_i\left(\theta_K^t\right)-\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right), -\tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle\\
=&\frac{\eta_G\eta_L}{N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\langle \nabla f_K\left(\theta_K^t\right)-\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right), -\tilde{\mathbf{d}}_{K,m}^{t,e}\right\rangle\\
\leq&\frac{\eta_G\eta_L}{2N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\nabla f_K\left(\theta_K^t\right)-\nabla f_{K,m}\left(\theta_{K,m}^{t,e}\right)\right\|^2+\frac{\eta_G\eta_L}{2N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\tilde{\mathbf{d}}_{K,m}^{t,e}\right\|^2\\
\leq&\frac{\eta_G\eta_L}{2N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\left(L^2\mathbb{E}_t\left\|\theta_K^t-\theta_{K,m}^{t,e}\right\|^2+\beta\sigma_G^2\right)+\frac{\eta_G\eta_L}{2N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\tilde{\mathbf{d}}_{K,m}^{t,e}\right\|^2\\
\leq&10\eta_G\eta_L^3E^3L^2D^2+\frac{\eta_G\eta_LE}{2}\beta\sigma_G^2+\frac{\eta_G\eta_L}{2N}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\mathbb{E}_t\left\|\tilde{\mathbf{d}}_{K,m}^{t,e}\right\|^2, \qquad (49)
\end{aligned}$$

where the first inequality comes from the truth that $\langle a,b\rangle\leq\frac{1}{2}\|a\|^2+\frac{1}{2}\|b\|^2$, the second inequality holds due to Assumption 2, and the third inequality comes from Lemma 3.

For term T_{32}, we have

$$T_{32} = \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(\mathbb{E}_t \left\langle \sum_{i=1}^{K-1} \tilde{\boldsymbol{H}}_i^t \left(\theta_{K,m}^{t,e} - \theta_{K,m}^t \right), \tilde{\mathbf{d}}_{K,m}^{t,e} \right\rangle \right.$$
$$\left. - \mathbb{E}_t \left\langle \sum_{i=1}^{K-1} \nabla f_i \left(\theta_K^t \right) + \nabla f_{K,m} \left(\theta_{K,m}^{t,e} \right) + \sum_{i=1}^{K-1} \tilde{\boldsymbol{H}}_i^t \left(\theta_{K,m}^{t,e} - \theta_{K,m}^t \right), \tilde{\mathbf{d}}_{K,m}^{t,e} \right\rangle \right) \tag{50}$$

Since $\tilde{\mathbf{d}}_{K,m}^{t,e}$ is a general solution in the convex hull of the vectors $\boldsymbol{g}_{K,m}^{t,e} := \{\nabla f_i(\theta_K^t) + \tilde{\boldsymbol{H}}_i(\theta_{K,m}^{t,e} - \theta_{K,m}^t)\}_{i=1}^{K-1} \cup \{\nabla f_{K,m}(\theta_{K,m}^{t,e})\}$, any vector $\boldsymbol{t}$ in the convex hull satisfies that $\left\langle \boldsymbol{t}, \tilde{\mathbf{d}}_{K,m}^{t,e} \right\rangle \geq \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2$ Yang et al. (2023). Therefore, we further have

$$\begin{aligned}
T_{32} \leq & - \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \sum_{i=1}^{K} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
& + \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \sum_{i=1}^{K-1} \left(\frac{1}{2} \left\| \tilde{\boldsymbol{H}}_i^t \right\|^2 \cdot \left\| \theta_{K,m}^{t,e} - \theta_{K,m}^t \right\|^2 + \frac{1}{2} \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \right) \\
\leq & - \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \sum_{i=1}^{K} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
& + \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(\frac{K-1}{2} L^2 \cdot 20 E^2 \eta_L^2 D^2 + \frac{1}{2} \sum_{i=1}^{K-1} \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \right) \\
= & - \frac{\eta_G \eta_L K}{2N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 + 10 (K-1) \eta_G \eta_L^3 E^3 D^2 L^2,
\end{aligned} \tag{51}$$

where the second inequality holds due to Lemma 3 and Lemma 2

Plugging Eq. (49) and Eq. (51) back to Eq. (48), we have

$$\begin{aligned}
T_3 \leq & 10 \eta_G \eta_L^3 E^3 L^2 D^2 + \frac{\eta_G \eta_L}{2N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
& - \frac{\eta_G \eta_L K}{2N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 + 10 (K-1) \eta_G \eta_L^2 E^3 D^2 L^2 \\
= & \left(\frac{1}{2} - \frac{K}{2} \right) \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 + 10 \eta_G K \eta_L^3 E^3 D^2 L^2.
\end{aligned} \tag{52}$$

For term T_4, we have

$$\begin{aligned}
T_4 = & \mathbb{E}_t \left\| - \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E} \mathbf{d}_{K,m}^{t,e} \right\|^2 \\
\leq & \frac{\eta_G^2 \eta_L^2 E}{M} \sum_{m=1}^{M} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \mathbf{d}_{K,m}^{t,e} \right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{\eta_G^2 \eta_L^2 E}{M} \sum_{m=1}^{M} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \mathbf{d}_{K,m}^{t,e} - \tilde{\mathbf{d}}_{K,m}^{t,e} + \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
&\leq \frac{\eta_G^2 \eta_L^2 E}{M} \sum_{m=1}^{M} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \mathbf{d}_{K,m}^{t,e} - \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 + \frac{\eta_G^2 \eta_L^2 E}{M} \sum_{m=1}^{M} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
&\leq \eta_G^2 \eta_L^2 E^2 \left(\sigma_L^2 + \frac{M-N}{N(M-1)} \beta \sigma_G^2 \right) + \frac{\eta_G^2 \eta_L^2 E}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2, \qquad (53)
\end{aligned}
$$

where the last inequality holds due to Lemma 5.

Plugging Eq. (52) and Eq. (53) back to Eq. (46), we have

$$
\begin{aligned}
\mathbb{E}_t \left[f_{1:K} \left(\theta_K^{t+1} \right) \right] \leq f_{1:K} \left(\theta_K^t \right) - \left(\frac{K}{2} - \frac{1}{2} - \frac{\eta_G \eta_L K E L}{2} \right) \frac{\eta_G \eta_L}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \\
+ 10 \eta_G \eta_L^3 K E^3 D^2 L^2 + \frac{\eta_G^2 \eta_L^2 K E^2 L}{2} \left(\sigma_L^2 + \frac{M-N}{N(M-1)} \beta \sigma_G^2 \right). \qquad (54)
\end{aligned}
$$

By setting $\frac{K}{2} - \frac{1}{2} - \frac{\eta_G \eta_L K E L}{2} \geq \frac{K}{4}$, that is, $\eta_G \eta_L \leq \frac{1}{6EL}$, we have

$$
\begin{aligned}
\frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 \leq & \frac{4 \left(f_{1:K} \left(\theta_K^t \right) - f_{1:K} \left(\theta_K^{t+1} \right) \right)}{\eta_G \eta_L K} + 40 \eta_L^2 K E^3 D^2 L^2 \\
& + 2 \eta_G \eta_L E^2 L \left(\sigma_L^2 + \frac{M-N}{N(M-1)} \beta \sigma_G^2 \right) \qquad (55)
\end{aligned}
$$

By defining $\nabla \tilde{f}_{1:K}(\theta_K^t) = \sum_{i=1}^{K} \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \frac{1}{E} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*} \nabla f_i(\theta_K^t)$, we have

$$
\begin{aligned}
&\mathbb{E}_t \left\| E \nabla \tilde{f}_{1:K}(\theta_K^t) \right\|^2 = \mathbb{E}_t \left\| \sum_{i=1}^{K} \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*} \nabla f_i(\theta_K^t) \right\|^2 \\
&\leq 2 \mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 + \underbrace{2 \, \mathbb{E}_t \left\| \sum_{i=1}^{K} \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*} \nabla f_i(\theta_K^t) - \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2}_{T_3}. \qquad (56)
\end{aligned}
$$

As for the term T_3, we have

$$
\begin{aligned}
T_3 = & \mathbb{E}_t \left\| \sum_{i=1}^{K-1} \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*} \tilde{\boldsymbol{H}}_i^t \left(\theta_K^t - \theta_{K,m}^{t,e} \right) \right. \\
& \left. + \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \left(\nabla f_K \left(\theta_K^t \right) - \nabla f_{K,m} \left(\theta_{K,m}^{t,e} \right) \right) \right\|^2 \\
\leq & \underbrace{2 \, \mathbb{E}_t \left\| \sum_{i=1}^{K-1} \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*} \tilde{\boldsymbol{H}}_i^t \left(\theta_K^t - \theta_{K,m}^{t,e} \right) \right\|^2}_{T_{31}}
\end{aligned}
$$

$$+ 2\,\underbrace{\mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \left(\nabla f_K \left(\theta_K^t \right) - \nabla f_{K,m} \left(\theta_{K,m}^{t,e} \right) \right) \right\|^2}_{T_{32}}. \tag{57}$$

According to Lemma 2, we have:

$$\begin{aligned} T_{31} \leq & \mathbb{E}_t \left\| \sum_{i=1}^{K-1} \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{i,m}^{t,e,*} \right\|^2 \cdot L^2 \cdot 20 \eta_L^2 E^2 D^2 \\ = & \mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(1 - w_{K,m}^{t,e,*} \right) \right\|^2 \cdot 20 \eta_L^2 E^2 D^2 L^2. \end{aligned} \tag{58}$$

Meanwhile, we can also bound term T_{32} as

$$\begin{aligned} T_{32} = & \mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \left(\nabla f_{K,m} \left(\theta_K^t \right) - \nabla f_{K,m} \left(\theta_{K,m}^{t,e} \right) \right) \right\|^2 \\ \leq & \mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \right\|^2 \cdot \left(20 \eta_L^2 E^2 D^2 L^2 + \beta \sigma_G^2 \right) \\ \leq & \mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \right\|^2 \cdot 20 \eta_L^2 E^2 D^2 L^2 + \beta E \sigma_G^2 \end{aligned} \tag{59}$$

Plugging Eq. (58) and Eq. (59) back to Eq. (57), we have

$$\begin{aligned} T_3 \leq & 40 \left(\mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(1 - w_{K,m}^{t,e,*} \right) \right\|^2 + \mathbb{E}_t \left\| \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \right\|^2 \right) \eta_L^2 E^2 D^2 L^2 \\ & + 2 \beta E \sigma_G^2. \end{aligned} \tag{60}$$

Plugging Eq. (60) back to Eq. (56), we have

$$\begin{aligned} \mathbb{E}_t \left\| \nabla \tilde{f}_{1:K} \left(\theta_K^t \right) \right\|^2 \leq & \frac{2}{E} \cdot \frac{1}{N} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left\| \tilde{\mathbf{d}}_{K,m}^{t,e} \right\|^2 + 80 \left(\mathbb{E}_t \left\| \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(1 - w_{K,m}^{t,e,*} \right) \right\|^2 \right. \\ & \left. + \mathbb{E}_t \left\| \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \right\|^2 \right) \eta_L^2 E^2 D^2 L^2 + 4 \beta E \sigma_G^2 \\ \leq & \frac{8 \left(f_{1:K} \left(\theta_K^t \right) - f_{1:K} \left(\theta_K^{t+1} \right) \right)}{\eta_G \eta_L E K} + 80 \eta_L^2 E^2 D^2 L^2 + 4 \eta_G \eta_L E L \left(\sigma_L^2 + \frac{M - N}{N (M - 1)} \beta \sigma_G^2 \right) \\ & + 80 \left(\mathbb{E}_t \left\| \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} \left(1 - w_{K,m}^{t,e,*} \right) \right\|^2 + \mathbb{E}_t \left\| \frac{1}{EN} \sum_{m \in \mathcal{S}_K^t} \sum_{e=0}^{E-1} w_{K,m}^{t,e,*} \right\|^2 \right) \eta_L^2 E^2 D^2 L^2 + 4 \beta \sigma_G^2. \end{aligned} \tag{61}$$

Rearranging and summing from $t = 0, \ldots, T-1$, we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_t \left\| \nabla \tilde{f}_{1:K} \left(\theta_K^t \right) \right\|^2 \leq \frac{8 \left(f_{1:K}^0 - f_{1:K}^* \right)}{\eta_G \eta_L E K T} + \frac{4 \eta_G \eta_L E L \sigma_L^2}{N} + \frac{4 (M - N)}{N (M - 1)} \eta_G \eta_L \beta E L \sigma_G^2$$

$$+ 80\left[1+\Lambda_{1:K-1}+\Lambda_{1:K}\right]\eta_L^2 E^2 D^2 L^2 + 4\beta\sigma_G^2, \tag{62}$$

where $f_{1:K}^0 \triangleq f_{1:K}\left(\theta_K^0\right)$, $f_{1:K}^* \triangleq f_{1:K}\left(\theta_K^*\right)$, $\Lambda_K \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_t\left\|\frac{1}{EN}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1} w_{K,m}^{t,e,*}\right\|^2$, and $\Lambda_{1:K-1} \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_t\left\|\frac{1}{EN}\sum_{m\in\mathcal{S}_K^t}\sum_{e=0}^{E-1}\left(1-w_{K,m}^{t,e,*}\right)\right\|^2$. □

F Details of Experiments

F.1 Datasets

We estimate the performance of FedGRIT and baselines on two class-incremental datasets and two domain-incremental datasets:

- Class-incremental datasets:
  - **CIFAR100** consists of 100 object classes with $50,000$ training images and $10,000$ test images. We randomly divide the 100 classes into 5 disjoint tasks, with each task containing 20 classes.
  - **ImageNet100** is a subset of ImageNet containing 100 object classes, with $50,000$ training images and $5,000$ test images. Similar to CIFAR100, we randomly divide the 100 classes into 5 disjoint tasks, with 20 classes assigned to each task.
- Domain-incremental datasets:
  - **Digit-10** consists of 10 classes across 4 domains: MNIST, USPS, SVHN, and EMNIST, with a total of $380,548$ images. Similarly, each domain is treated as an individual task, resulting in 4 sequential tasks.
  - **VLCS** consists of 5 classes across 4 domains: VOC2007, LabelMe, Caltech101, and SUN09, with a total of $7,486$ images. We treat each domain as an individual task, resulting in 4 sequential tasks.

F.2 Detailed Configurations

Our experiments contain two parts: the main experiment based on 4 datasets and the ablation study.

- The main experiment focuses on evaluating the performance of FedGRIT in the four datasets and comparing it with other baselines in two metrics, i.e., ACC and BWT. In the main experiment, the local learning rate η_L is initialized as 0.001 and decayed with the rate of 0.96 after each 5 epochs, and the global learning rate η_G is fixed at 1.0. We consider the number of clients to be 16 and the number of participants to be 8 each round. We set communication rounds $T = 50$ and local epochs $E = 3$ for all datasets. We set the dirichlet parameter $\alpha = 0.5$ for two domain-incremental datasets and $\alpha = 1.0$ for two class-incremental datasets. The size of the buffer for each client is fixed at 400. The maximum iterations for Frank Wolfe and Lanczos algorithms are fixed at 10.
- In the ablation study, we conduct a series of experiments to evaluate the performance of FedGRIT under different settings on CIFAR100. In the ablation study of the updating frequency, we set the update interval $\rho = \{1, 3, 5, 10\}$. In the ablation study of buffer size, we set buffers with different sizes $\{100, 400, 700, 1000\}$. In the ablation study of the truncated rank-r, we set $r = \{1, 3, 5, 10\}$. In the ablation study on the straggler robustness, we set the ratio $\rho = \{0.25, 0.5, 0.75, 1.0\}$. In the client scaling and participation ratio, we consider two configurations as $N/M = \{8/8, 8/16, 8/32, 8/64\}$ and $N/M = \{2/8, 4/16, 8/32, 16/64\}$. All ablation studies follow the same basic experimental settings as the main experiments, with only the configuration corresponding to the ablation component being varied.

In addition, the hyperparameters of baselines are listed in Table 4, and unspecified hyperparameters follow the settings reported in the original papers.

Table 4: Hyperparameters for baselines

Baselines	Hyperparameters
FedProx	$\lambda = 0.5$
FLwF-2T	$\alpha = 0.5, \beta = 0.25$
FedSSI	$\lambda = 0.5$
FedAGC	$\gamma = 0.5, \lambda = 0.01$
SPECIAL	$\lambda = 0.4$ for Digit-10 and VLCS, $\lambda = 0.3$ for CIFAR100 and ImageNet100

F.3 Additional Ablation Studies

In order to observe the evolution of model performance on the whole task sequence, we further define the average all-task accuracy for task i by

$$\bar{\text{ACC}}_i = \frac{1}{K}\sum_{j=1}^{K} A_{i,j}, \tag{63}$$

where $A_{i,j}$ represents the test accuracy on task j after learning task i. Note that this metric is evaluated over all K tasks, including both seen and unseen tasks at task i.

Ablation on the buffer size. Buffer size determines the number of previous samples that a client can store to be used to reconstruct the gradients. Figure 4 shows that a larger buffer size provides richer previous-information and leads to better performance. However, the performance degradation remains marginal even when the buffer size is reduced to 400, indicating that FedGRIT does not heavily rely on storing a large number of previous samples and it can still effectively extract and aggregate curvature information even based on a small number of previous samples.

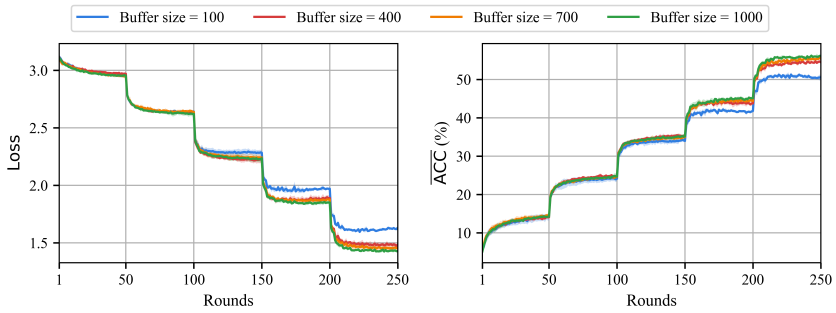


Figure 4: Loss and ACC (%) of FedGRIT with different buffer sizes

Ablation on the rank-r in truncated EVD. The value of r determines the number of eigendirections retained for Hessian reconstruction. A larger r preserves curvature information along more directions, but it would also introduce additional non-dominant information and noise that can impair the model's generalization ability. Figure 5 shows that the best performance is achieved when the truncation rank r is set to 3, indicating that the dominant Hessian information can be effectively captured by only a few leading eigendirections, and excessive eigendirections may introduce additional client-specific variations to degrade the overall performance.

Straggler robustness. Curvature information exchange requires all participants to upload their previous-task information in our default setting. However, the bandwidth and hardware constraints may prevent stragglers from transmitting their information at every round. Therefore, we consider a restricted setting where only a fraction of participants upload their curvature information to be aggregated in the server. Figure 6 shows that the performance degrades only marginally when only 50% of the participants provide curvature information in each round, and the final performance remains competitive even when the fractions is further reduced to 25%, demonstrating the robustness of FedGRIT under more practical settings.

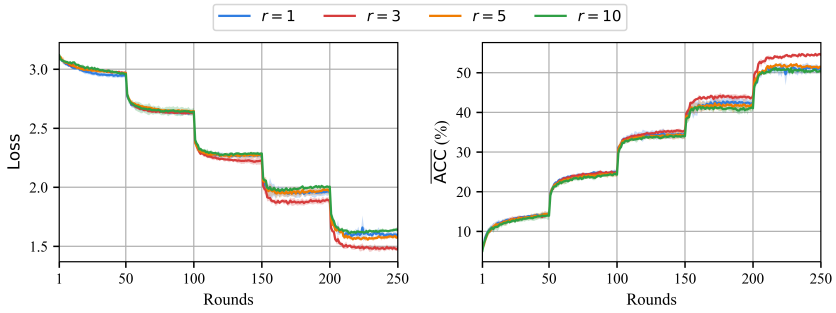


Figure 5: Loss and ACC (%) of FEDGRIT with different truncation rank r

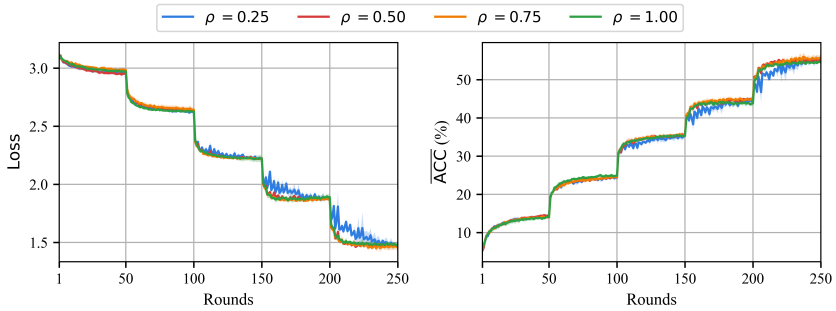


Figure 6: Loss and ACC (%) of FEDGRIT with different fractions of participants uploading the curvature information.

Client scaling and participation ratio. Figure7 and Figure 8 study how FEDGRIT scales with the participation ratio and client population when the entire dataset is distributed across 64 clients. With fixed participants $N = 8$, increasing the number of total clients M can expose the model to more client data to improve the performance. With fixed ratio $N/M = 1/4$, scaling up both M and N allows more clients to participate in each training round, enabling the model to benefit from more diverse local information and leading to improved training performance. This observation strengthens Theorem 2 and highlights the dominant effect of M on model performance.

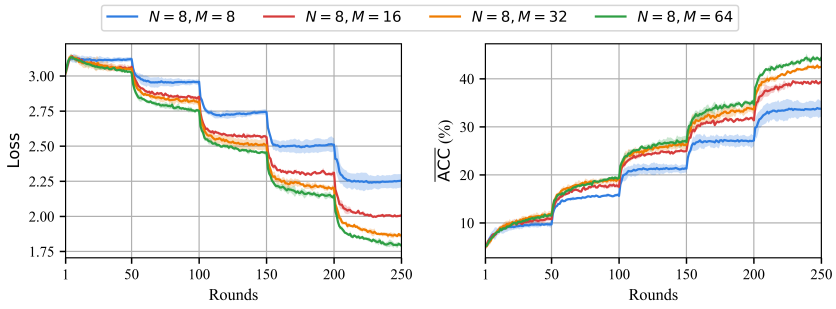


Figure 7: Loss and ACC (%) of FEDGRIT with fixed number of participants but different ratios.

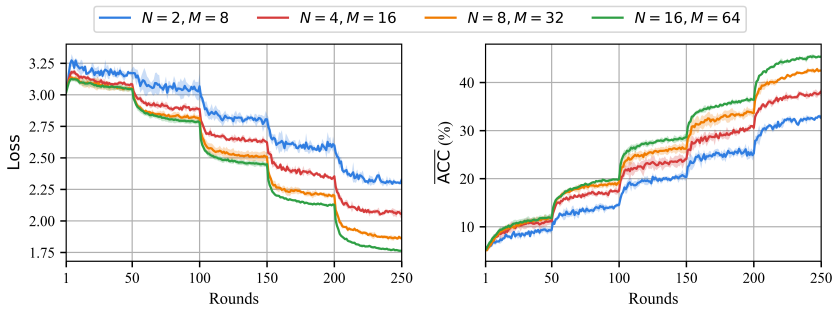


Figure 8: Loss and ACC (%) of FEDGRIT with fixed ratio but different clients population.

Different Hessian Approximation. We compare the Hessian approximation method between truncated EVD adopted in FEDGRIT with the Fisher-information-based method and Kronecker-factored approximate curvature (K-FAC), to evaluate the effectiveness of different Hessian approximation strategies. Figure 9 shows that the Fisher-information-based approximation can effectively retain previous information during the first few rounds after transitioning to a new task. However, the Fisher approximation may ignore the covariance of gradients which captures the interactions among parameters, limiting the accuracy of curvature approximation and degrading the performance. Although K-FAC captures richer parameter interactions, its approximation accuracy is constrained by the assumed Kronecker and layer-wise structures, which may contribute to insufficient characterization of the dominant curvature directions and affecting the stability of model training. In contrast, our truncated EVD mechanism retains the top-r dominant eigendirections, which jointly captures the important covariance and preserving the principal information of the original Hessian, thereby achieving stable performance across all tasks.

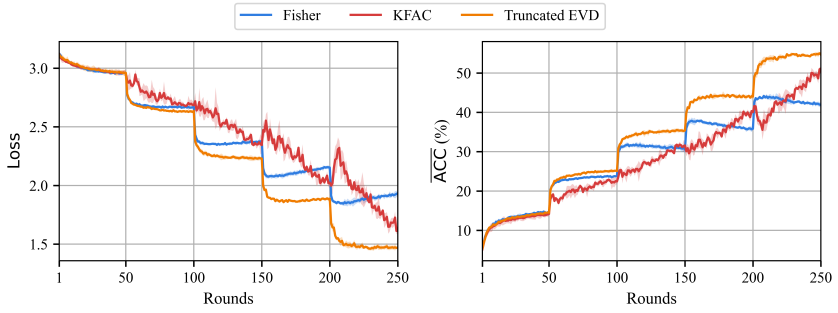


Figure 9: Different Hessian Approximation.

Table 5: Communication and runtime cost. We report the *rounds-to-best* and *time-per-round* over all tasks for each dataset. Lower is better for both metrics.

Method	CIFAR-100		ImageNet100		Digit-10		VLCS	
	Rounds(r)	Time(s)	Rounds(r)	Time(s)	Rounds(r)	Time(s)	Rounds(r)	Time(s)
FedAvg	226.7	**27.7**	221.7	**36.7**	55.7	**68.8**	90.7	**51.2**
FedProx	220.0	34.7	220.7	39.6	60.0	75.6	83.7	52.1
FSMGDA	245.7	33.4	240.7	42.7	107.7	75.3	88.7	53.7
FedCMOO	237.3	40.3	242.7	47.0	95.0	72.8	78.0	55.6
FLwF-2t	218.3	39.1	212.0	46.4	84.3	90.9	74.7	52.4
FedSSI	**162.0**	46.7	216.0	56.7	62.3	141.2	68.7	56.6
SPECIAL	228.3	28.0	199.7	36.3	64.0	69.8	**57.0**	52.7
FedPOCL	238.7	86.0	229.3	93.4	98.7	143.1	75.0	65.2
FedAGC	244.0	101.4	229.7	102.1	107.0	150.1	94.0	56.9
FedBECAME	242.3	53.0	221.3	78.4	106.0	102.6	60.7	77.7
C-FLAG	227.0	59.3	224.0	48.6	68.7	92.6	75.3	59.3
FEDGRIT (ours)	206.3	64.1	**195.7**	72.9	**41.0**	131.3	66.7	69.2

Table 6: Numerical efficiency comparison on CIFAR-100. Communication cost is measured per selected client per round, including both upload and download. Peak memory usage refers to peak allocated GPU memory.

Method	Communication cost (MB) ↓	Peak memory usage (GB) ↓
FedAvg	85.73	4.40
FedProx	85.73	4.35
FSMGDA	171.47	6.31
FedCMOO	214.23	6.22
FLwF-2T	85.73	5.39
FedSSI	85.73	22.06
SPECIAL	85.73	4.40
FedPOCL	85.73	4.74
FedAGC	128.60	8.51
FedBECAME	85.73	9.18
C-FLAG	257.19	12.65
FEDGRIT (ours)	347.88	18.52

F.4 COMPUTATION AND COMMUNICATION COST

We report the results of rounds-to-best and time-per-round for all algorithms in Table 5. Compared with other A-FCL methods, FEDGRIT performs client-specific weight adaptation, allowing each client to determine its own optimal update direction at each local epoch instead of following the same weights as other clients. This personalized adaptation strategy enables FEDGRIT to achieve the fewest rounds-to-best among all FCL and A-FCL methods. The numerical efficiency comparisons are provided in Table 6. The results show that methods performing information exchange and processing entirely on the client side incur communication costs comparable to FedAvg. In contrast, federated multi-objective learning methods require the transmission of multiple task gradients, FedAGC additionally downloads group models, and C-FLAG exchanges $\nabla g(x_t)$ and $\nabla f^{\dagger}(x_t)$ with the server to determine adaptive learning rates. Similarly, FedGRIT requires additional curvature-information exchange between clients and the server. These additional client-server interactions inevitably introduce communication overhead beyond standard FL.