KNOWLEDGE-AWARE EVOLUTION FOR TASK-FREE STREAMING FEDERATED CONTINUAL LEARNING WITH ARBITRARY CLASS OVERLAP

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Continual Learning (FCL) leverages inter-client collaboration to balance new knowledge acquisition and old knowledge retention, assuming explicit task boundaries. However, existing FCL methods struggle in streaming scenarios where ephemerally accessible data chunks exhibit arbitrary class overlap and lack task identifiers, inducing knowledge conflicts of the same class and preventing continual inference on local cumulative classes. To address this, we propose FedKACE with three synergistic components: 1) a gradient-responsive replay scheme that utilizes the ratio of squared L2 gradient norms of buffer to new samples to achieve client-specific balance between acquisition and retention; 2) a holistic buffer maintenance strategy that retains boundary-critical samples to enhance knowledge retention under class overlap; 3) an adaptive local-to-global inference model switching mechanism for continual inference on local cumulative classes. Experiments under diverse class overlap settings, along with theoretical analysis, demonstrate the effectiveness of FedKACE.

1 INTRODUCTION

Federated learning (FL) McMahan et al. (2017) enables multiple clients to collaboratively train models without sharing raw data, yet its static data assumption leads to severe performance degradation in dynamic class-incremental scenarios or under distribution shifts Li et al. (2020). To address this, federated continual learning (FCL) Yang et al. (2024) extends FL to dynamic environments using continual learning mechanisms, balancing new knowledge acquisition and old knowledge retention via cross-client collaboration, assuming explicit task boundaries Wang et al. (2024a).

However, in real-world settings, edge clients with constrained storage receive sequential data chunks exhibiting arbitrary class overlap and lacking task identifiers, where continuous data influx forces each chunk to be discarded after limited iterations to avoid memory overflow Kairouz et al. (2021). This reality violates two core assumptions of traditional FCL methods: First, arbitrary class overlap without task identifiers renders knowledge conflicts of the same class unresolvable Kirkpatrick et al. (2016), breaking the premises of both mutually exclusive class sets across tasks and identical class sets across clients. Second, ephemeral data access limits iterations per chunk and necessitates local inference on all seen classes Mai et al. (2022), contradicting the reliance on task identifiers and static intra-task data spanning multiple FL rounds. These violations expose key limitations of existing FCL methods Wang et al. (2024a): sample replay methods struggle to balance old and new samples of the same class in constrained buffers; generative replay methods suffer from poor generation quality under restricted iterations; and parameter isolation methods overallocate parameters and inflate model size in the absence of explicit task boundaries. Such limitations collectively impede practical FCL deployment.

To address these limitations, we tackle unique technical challenges in streaming FCL, where clients receive sample-disjoint data chunks that are accessible solely during their respective FL rounds, yet the class sets of these chunks exhibit arbitrary overlap. This setting poses three core challenges: (1) disentangling new and old samples of the same class without task boundaries; (2) preserving boundary-critical knowledge in constrained buffers under class overlap; and (3) performing continual inference on local cumulative classes without task identifiers. Preserving the low training

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.

complexity of sample replay, we propose the **Fed**erated **K**nowledge-**A**ware **C**ontinual **E**volution (FedKACE) framework. First, a gradient-responsive replay scheme addresses (1) by dynamically adjusting replay loss weights based on the ratio of the squared L2 gradient norms of buffer samples to those of new samples, achieving client-specific balance between knowledge acquisition and retention. Second, a holistic buffer maintenance strategy addresses (2) by jointly optimizing the informativeness, decision boundary significance, and feature space dispersion of buffer samples through a two-stage screening process to preserve boundary-critical knowledge under class overlap. Finally, an adaptive inference switching mechanism addresses (3). Motivated by the distinct convergence dynamics of local and global models, it determines when to switch the inference model by monitoring the rate of change in the discrepancy between the global model's accuracy and prediction confidence, both evaluated on the local buffer, thereby improving client-side inference performance.

Our contributions are as follows:

1. We address streaming FCL where data chunks exhibit arbitrary class overlap and lack task identifiers, requiring local continual inference on all encountered classes after each FL round.

2. We propose FedKACE, integrating a gradient-responsive replay scheme for disentangling old and new knowledge of the same class, a holistic buffer maintenance strategy for preserving boundary-critical samples under class overlap, and an adaptive switching mechanism for continual inference.

3. Experiments across diverse overlap settings, along with theoretical analysis, validate FedKACE.

2 Related Works

2.1 Federated Continual Learning

Federated Continual Learning (FCL) extends FL to adapt to dynamic data distributions while balancing stability and plasticity. Existing FCL methods fall into five types: Rehearsal (Sample Replay) methods Li et al. (2024); Serra & Buettner (2025); Qi et al. (2025) mitigate forgetting by storing key samples but struggle to balance the representation of old and new samples for overlapping classes; Generative Replay methods Tran et al. (2024); Rong et al. (2025) avoid raw data storage yet are constrained by generation quality and computational costs; Parameter Isolation methods Wang et al. (2024b); Salami et al. (2025) reduce storage needs but require explicit task boundaries and inflate the model size; Regularization methods Usmanova et al. (2021); Li et al. (2025b) require no extra storage but assume similar class sets across clients at each time step; Dynamic Collaboration methods Wu et al. (2025); Yang et al. (2025) incur no additional storage overhead yet depend on clients' class sets jointly cover all possible classes per FL round. Moreover, most FCL methods assume explicit task identifiers, full data availability for each task throughout its multiple FL rounds, and mutually exclusive class sets across tasks. These premises inherently fail in streaming settings characterized by ephemeral data, absent task boundaries, and arbitrary class overlap, hindering their performance.

2.2 Online Continual Learning

Online Continual Learning (OCL), at the intersection of continual and online learning, processes streaming data restricted to task-specific classes within each task Mai et al. (2022). Existing OCL methods fall into four types: Rehearsal (Sample Replay) methods Tong et al. (2025); Ye & Bors (2025) retain prior knowledge by storing key samples but are highly dependent on buffer capacity; Gradient Optimization methods Elsayed & Mahmood (2024); Wu et al. (2024); Ash et al. (2020) balance stability and plasticity by employing specialized update trajectories and thereby incur high computational costs and poor scalability; Module Enhancement methods Zhou et al. (2025); Liu et al. (2025) improve adaptation to non-stationary environments by introducing task-specific modules but require explicit task identifiers; Regularization methods Michel et al. (2024); Wang et al. (2024c) preserve prior knowledge by constraining parameter updates yet heavily rely on initialization from pretrained models. Moreover, while recent task-free OCL methods have relaxed explicit boundary assumptions, they remain limited in two regimes. First, deployed independently, they overfit locally biased streams under constrained buffers, lacking cross-client collaboration to counteract client drift. Second, naively coupled with FL and relying exclusively on global models, they suffer from cross-client heterogeneity and fail to adapt to local streaming chunks, forfeiting adaptive local knowledge retention.

3 NOTATION AND PROBLEM FORMULATION

In the streaming FCL setting, we assume static client data within each FL round. Consider a T-round FL process with a client set $\mathcal{K} = \{1, \dots, K\}$. In each FL round $t \in [T]$, each client $k \in \mathcal{K}$ receives a new data chunk $\mathcal{D}_k^t = \{(x_i, y_i)\}_{i=1}^{n_k^t}$ that is accessible solely during the t-th FL round, where n_k^t is the number of samples, x_i is an input and y_i is its label from the class set $\mathcal{C}_k^t$ of $\mathcal{D}_k^t$. Moreover, for each client k, data chunks are disjoint across FL rounds (i.e., $\mathcal{D}_k^t \cap \mathcal{D}_k^{t'} = \emptyset$ for all $t \neq t'$), yet class sets allow for arbitrary overlap (i.e., $0 \leq |\mathcal{C}_k^t \cap \mathcal{C}_k^{<t}| \leq \min(|\mathcal{C}_k^t|, |\mathcal{C}_k^{<t}|)$), where $\mathcal{C}_k^{<t} = \bigcup_{\tau=1}^{t-1} \mathcal{C}_k^\tau$ includes all classes encountered up to FL round $t-1$.

At the global level, let $\mathcal{C}_g = \bigcup_{k=1}^{K} \bigcup_{t=1}^{T} \mathcal{C}_k^t$ denote the global class set with size $|\mathcal{C}_g| = C_{\max}$, we set each client $k \in \mathcal{K}$ to encounter all $C_{\max}$ classes across all T FL rounds with heterogeneous class trajectories, resulting in arbitrary overlap among the local cumulative class sets $\mathcal{C}_k^{\leq t} = \bigcup_{\tau=1}^{t} \mathcal{C}_k^\tau$ for any intermediate FL round $t < T$ (i.e., $0 \leq |\mathcal{C}_k^{\leq t} \cap \mathcal{C}_{k'}^{\leq t}| \leq \min(|\mathcal{C}_k^{\leq t}|, |\mathcal{C}_{k'}^{\leq t}|)$ for all $k \neq k'$), thereby introducing complementary knowledge among clients.

Formally, we define the model $\theta = \{\phi, h\}$, where ϕ is the feature extractor and h is the linear layer of fixed output dimension $C_{\max}$. In FL round t, client k activates the output dimensions of its local model $\theta_k^t = \{\phi_k^t, h_k^t\}$ or the global model $\theta_g^t = \{\phi_g^t, h_g^t\}$ corresponding to $\mathcal{C}_k^{\leq t}$ through masking. The objective for each client $k \in \mathcal{K}$ in the t-th FL round is to optimize the inference model $\tilde{\theta}_k^t = \{\tilde{\phi}_k^t, \tilde{h}_k^t\}$ (i.e., θ_k^t or θ_g^t) for the local cumulative class set $\mathcal{C}_k^{\leq t}$, formalized as:

$$\min_{\tilde{\theta}_k^t} \mathcal{F}(\tilde{\theta}_k^t) = \sum_{c \in \mathcal{C}_k^{\leq t}} \mathop{\mathbb{E}}_{x \sim \mathcal{P}_{g,c}} \left[l\left(f_{\tilde{\theta}_k^t}(x), c \right) \right] + \mathcal{R}_k^t(\tilde{\theta}_k^t) \tag{1}$$

where $f_{\tilde{\theta}_k^t}(x) = \tilde{h}_k^t(\tilde{\phi}_k^t(x))$ is the prediction function, $\mathcal{P}_{g,c}$ is the global class-conditional data distribution for class c, l is the loss function, and $\mathcal{R}_k^t(\cdot)$ is a federated regularization term applied to $\tilde{\theta}_k^t$ to enhance global generalization. See Appendix A.4 for a detailed comparison with FCL/OFCL.

4 METHODOLOGY

4.1 GRADIENT-RESPONSIVE REPLAY SCHEME (GRRS)

In streaming FCL, heterogeneous data distributions across clients lead to divergent parameter update directions and magnitudes, rendering fixed knowledge retention strategies unable to adapt to varying client learning dynamics. Due to the lack of task identifiers, we propose that each client treats samples in its representative local buffer as a unified historical task. In FL round t, client k receives a new data chunk $\mathcal{D}_k^t$ with class set $\mathcal{C}_k^t$ and accesses its previous local buffer $\mathcal{M}_k^{t-1}$ with historical class set $\mathcal{C}_k^{<t} = \bigcup_{\tau=1}^{t-1} \mathcal{C}_k^\tau$, exhibiting arbitrary class overlap such that $0 \leq |\mathcal{C}_k^t \cap \mathcal{C}_k^{<t}| \leq \min(|\mathcal{C}_k^t|, |\mathcal{C}_k^{<t}|)$. For each epoch $j \in [1, J]$, given the local model $\theta_k^{t,j-1}$ to be updated and the cross-entropy loss function $l(\cdot;\cdot)$, the task loss $L_{k,\text{task}}^t(\theta_k^{t,j-1})$ on $\mathcal{D}_k^t$ and the replay loss $L_{k,\text{rep}}^t(\theta_k^{t,j-1})$ on $\mathcal{M}_k^{t-1}$ are:

$$L_{k,\text{task}}^t(\theta_k^{t,j-1}) = \sum_{(x,y) \in \mathcal{D}_k^t} l\left(f_{\theta_k^{t,j-1}}(x)[\mathcal{C}_k^t]; y \right) \tag{2}$$

$$L_{k,\text{rep}}^t(\theta_k^{t,j-1}) = \sum_{(x,y) \in \mathcal{M}_k^{t-1}} l\left(f_{\theta_k^{t,j-1}}(x)[\mathcal{C}_k^{<t}]; y \right) \tag{3}$$

where $[\mathcal{C}_k^t]$ and $[\mathcal{C}_k^{<t}]$ denote the output masks that restrict predictions to the respective classes. Given an adaptive replay loss weight $\lambda_k^{t,j}$, the total training loss in epoch j is:

$$L_{k,\text{total}}^t(\theta_k^{t,j-1}) = L_{k,\text{task}}^t(\theta_k^{t,j-1}) + \lambda_k^{t,j} \cdot L_{k,\text{rep}}^t(\theta_k^{t,j-1}) \tag{4}$$

Then, client k update $\theta_k^{t,j} = \theta_k^{t,j-1} - \eta \cdot \nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j-1})$ with learning rate η, completing the j-th epoch. For the adaptive replay loss weight λ, given that the squared L2 norms of gradients quantify parameter change magnitudes Li et al. (2023), we propose that client k computes the ideal weight

$\tilde{\lambda}_k^{t,j+1}$ in epoch $j+1$ as the ratio of the squared L2 norms of the replay and task loss gradients, evaluated at $\theta_k^{t,j-1}$ during epoch j:

$$\tilde{\lambda}_k^{t,j+1} = \|\nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j-1})\|_2^2 / \|\nabla_\theta L_{k,\text{task}}^t(\theta_k^{t,j-1})\|_2^2 \tag{5}$$

This weight naturally increases to strengthen retention constraints when buffer gradients dominate, as a large replay gradient signals that the model has drifted from the old-task optimum and forgetting is underway; conversely, it decreases to relax them when new data gradients prevail, freeing parameter capacity for new knowledge acquisition. However, computing the full-model gradients for both replay and task losses incurs two full backward passes. Following Ash et al. (2020), we use the output layer h as a proxy, and define the actual weight $\lambda_k^{t,j+1}$ for epoch $j+1$ as:

$$\lambda_k^{t,j+1} = \|\nabla_h L_{k,\text{rep}}^t(\theta_k^{t,j-1})\|_2^2 / \|\nabla_h L_{k,\text{task}}^t(\theta_k^{t,j-1})\|_2^2 \tag{6}$$

Replacing full backward passes with output-layer computations reduces overhead, since output-layer gradients directly reflect the sensitivity of classification boundaries to new versus old knowledge. Gradient norms are computed as running averages of batch gradients within each epoch. In the first epoch, each client $k \in \mathcal{K}$ initializes its local model $\theta_k^{t,0} = \theta_g^{t-1}$ and sets $\lambda_k^{t,1} = 1$.

Building on the Gradient-Responsive Replay Scheme, we present the following Theorem 1:

Theorem 1 (Local Saddle-Point Convergence of GRRS) *Under standard assumptions (see Appendix B.1), let $\theta_k^{t,0}$ be the initial model for client k in FL round t. After J local epochs via the Gradient-Responsive Replay Scheme, $\theta_k^{t,J}$ converges to the ideal saddle point $(\theta_k^{t,*}, \lambda_k^{t,*})$ within a truncation error bound:*

$$\mathbb{E}[L_{k,total}^t(\theta_k^{t,J}|\lambda_k^{t,J})] \le L_{k,total}^t(\theta_k^{t,*}|\lambda_k^{t,*}) + \mathcal{E}_{k,bound}^t(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2) \tag{7}$$

where the ideal saddle point $(\theta_k^{t,*}, \lambda_k^{t,*})$ satisfies the minimax equilibrium property for any $\theta \in \Omega_\theta$ and $\lambda \in \Omega_\lambda$:

$$L_{k,total}^t(\theta_k^{t,*}|\lambda) \le L_{k,total}^t(\theta_k^{t,*}|\lambda_k^{t,*}) \le L_{k,total}^t(\theta|\lambda_k^{t,*}) \tag{8}$$

$\mathcal{E}_{k,bound}^t\left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2\right)$ *is an error bound functional detailed in Appendix B.3.*

The proof of Theorem 1 is in Appendix B.3.

4.2 Holistic Buffer Maintenance Strategy (HBMS)

In streaming FCL without task identifiers, client buffer constraints pose particular challenges in the presence of class overlap, impeding the balanced representation of identical classes across data chunks. Guided by kernel spectral boundary theory and information value assessment, we design a two-stage buffer maintenance strategy to maximize boundary-critical knowledge retention.

In FL round t, client k updates the global model θ_g^{t-1} using the new data chunk $\mathcal{D}_k^t$ and its buffer $\mathcal{M}_k^{t-1}$ (with capacity M) to obtain the local model $\theta_k^{t,J}$, and then constructs its new buffer $\mathcal{M}_k^t$ from $\mathcal{M}_k^{t-1} \cup \mathcal{D}_k^t$ using $\theta_k^{t,J}$ for the class set $\mathcal{C}_k^{\le t} = \mathcal{C}_k^{<t} \cup \mathcal{C}_k^t$: First, client k computes the normalized logit vector $\hat{g}_k^t(x) = f_{\theta_k^{t,J}}(x) / \|f_{\theta_k^{t,J}}(x)\|_2$ for each sample $x \in \mathcal{M}_k^{t-1} \cup \mathcal{D}_k^t$, and next:

Computing Kernel Diversity and Class Metrics: Maximizing the minimum distance of candidate samples to the current buffer enhances feature dispersion and prioritizes boundary samples to ensure significance and informativeness, reducing the kernel matrix condition number and hypothesis space complexity. To achieve this, we first define a Gaussian kernel function $K(x, x_i)$ that captures the spatial distribution of $\hat{g}_k^t(x)$ via its smoothness and local sensitivity:

$$K(x, x_i) = \exp\left(-\beta \|\hat{g}_k^t(x) - \hat{g}_k^t(x_i)\|_2^2\right) \tag{9}$$

where β controls the kernel decay rate, defined as $|\mathcal{M}_k^{t-1}|^{2/d}$ for effective class dimension $d = |\mathcal{C}_k^{\le t}|$ if $\mathcal{M}_k^{t-1} \neq \emptyset$, and 1 otherwise. Then, client k computes the kernel diversity score (DS) for each $x \in \mathcal{M}_k^{t-1} \cup \mathcal{D}_k^t$ with respect to $\mathcal{M}_k^{t-1}$:

$$\text{DS}(x) \underset{\text{objective}}{\triangleq} -\log\left(\max_{x_i \in \mathcal{M}_k^{t-1}} K(x, x_i)\right) \underset{\text{computed}}{\propto} \min_{x_i \in \mathcal{M}_k^{t-1}} \|\hat{g}_k^t(x) - \hat{g}_k^t(x_i)\|_2^2 \tag{10}$$

If $\mathcal{M}_k^{t-1} = \emptyset$, $\mathrm{DS}(x) = 0$ due to the absence of historical samples. Then, for each sample $x \in \mathcal{M}_k^{t-1} \cup \mathcal{D}_k^t$ with hard label c_x, if $c_x \in \mathcal{C}_k^{<t}$, client k uses $K(x, x_i)$ to compute the conditional predictive distributions before and after its inclusion in the buffer, by reusing the pre-computed pairwise distances $\|\hat{g}_k^t(x) - \hat{g}_k^t(x_i)\|_2^2$ obtained when computing DS:

$$\bar{p}_{\mathcal{M}_k^{t-1}}(c_x \mid x) = A(x, c_x)/B(x); \quad \bar{p}_{\mathcal{M}_k^{t-1} \cup \{x\}}(c_x \mid x) = \frac{A(x, c_x) + K(x, x) \cdot p(c_x \mid x)}{B(x) + K(x, x)} \tag{11}$$

where $A(x, c_x) = \sum_{x_i \in \mathcal{M}_k^{t-1}} K(x, x_i) \cdot p(c_x \mid x_i)$ with $p(c_x \mid x)$ being the predicted probability for class c_x under $\theta_k^{t,J}$, and $B(x) = \sum_{x_i \in \mathcal{M}_k^{t-1}} K(x, x_i)$. If $\mathcal{M}_k^{t-1} = \emptyset$ or new classes $c_x \in \mathcal{C}_k^t \setminus \mathcal{C}_k^{<t}$, both $\bar{p}_{\mathcal{M}_k^{t-1}}(c_x \mid x)$ and $\bar{p}_{\mathcal{M}_k^{t-1} \cup \{x\}}(c_x \mid x)$ are set to $p(c_x \mid x)$ due to the lack of prior information. Accordingly, client k computes the Information-Diversity Value (IDV) and Consistency-Diversity Value (CDV) for each sample $x \in \mathcal{M}_k^{t-1} \cup \mathcal{D}_k^t$. When $c_x \in \mathcal{C}_k^{<t}$, IDV and CDV are defined as:

$$\mathrm{IDV}(x \mid c_x) = -\log \bar{p}_{\mathcal{M}_k^{t-1}}(c_x \mid x) + \lambda_1 \cdot \mathrm{DS}(x) \tag{12}$$

where the adaptive factors $\lambda_1 = \log |\mathcal{M}_k^{t-1}|/(|\mathcal{M}_k^{t-1}|)^{1/2}$ and $\lambda_2 = 1/(|\mathcal{M}_k^{t-1}|)^{1/2}$ prioritize diversity for small buffers and informativeness for large buffers, with $\lambda_1 = \lambda_2 = 0$ if $\mathcal{M}_k^{t-1} = \emptyset$. If $\mathcal{M}_k^{t-1} = \emptyset$ or new classes $c_x \in \mathcal{C}_k^t \setminus \mathcal{C}_k^{<t}$, we set $\mathrm{IDV}(x \mid c_x) = -\log p(c_x \mid x) + \lambda_1 \cdot \mathrm{DS}(x)$ and $\mathrm{CDV}(x \mid c_x) = 0$ due to the lack of prior information.

Maintaining Buffer via Two-Stage: In our method, buffer capacity M is evenly distributed to form subsets $\mathcal{M}_{k,c}^t$ for each class $c \in \mathcal{C}_k^{\leq t}$, each of size $\lfloor M/|\mathcal{C}_k^{\leq t}| \rfloor$, where samples are selected from the candidate set $X_c = \{x \in \mathcal{M}_k^{t-1} \cup \mathcal{D}_k^t \mid y_x = c\}$ of class c: first filtering high-uncertainty candidates via IDV, then retaining boundary-critical ones via CDV, as IDV and CDV target distinct objectives with incompatible scales. For old classes $c \in \mathcal{C}_k^{<t}$, client k selects the top-$\min\{2|\mathcal{M}_{k,c}^t|, |X_c|\}$ samples from X_c ranked by $\mathrm{IDV}(x|c)$ to form a candidate set $\tilde{\mathcal{M}}_{k,c}^t$, and then retains the top-$|\mathcal{M}_{k,c}^t|$ samples ranked by $\mathrm{CDV}(x|c)$ as the buffer $\mathcal{M}_{k,c}^t$. For new classes $c \in \mathcal{C}_k^t \setminus \mathcal{C}_k^{<t}$ where $\mathrm{CDV}(x \mid c) = 0$ for all $x \in X_c$, client k forms $\mathcal{M}_{k,c}^t$ by selecting $|\mathcal{M}_{k,c}^t|$ samples via information-weighted sampling without replacement, with probability $P(x) = \exp(\mathrm{IDV}(x|c))/\sum_{x' \in X_c} \exp(\mathrm{IDV}(x'|c))$. Finally, client k constructs the complete buffer $\mathcal{M}_k^t = \bigcup_{c \in \mathcal{C}_k^{\leq t}} \mathcal{M}_{k,c}^t$.

Building on the Holistic Buffer Maintenance Strategy, we present the following Theorem 2:

Theorem 2 (Spectral Efficiency of HBMS) *Given $M > 2C_{\max}$ and standard assumptions (see Appendix B.1), let $C_{\mathrm{rand}} \triangleq 1$ be the spectral efficiency constant of class-wise random sampling, and let $C_\kappa \triangleq \inf_{\theta \in \Omega_\theta, \hat{g} \in \mathcal{G}} \exp(-\beta(\theta, \hat{g}) \Delta\rho(\theta, \hat{g}))$ (depending only on M and $C_{\max}$) be that of our Holistic Buffer Maintenance Strategy, where $\Delta\rho(\theta, \hat{g}) \triangleq \rho_{\mathrm{rand}} - \rho_{\mathrm{HBMS}}$ is the spectral radius gap between the two methods. Then, $0 < C_\kappa < C_{\mathrm{rand}}$ holds for all $k \in \mathcal{K}$ and $t \in [T]$.*

The proof of Theorem 2 is in Appendix B.4.

4.3 Adaptive Inference Model Switching Mechanism

In FL, local and global models exhibit distinct performance trajectories on local data: local models outperform the global model in early FL rounds but overfit local data (or local buffer) more rapidly while the global model achieves better generalization via global regularization over time. Inspired by Futami & Fujisawa (2024), we propose that each client $k \in \mathcal{K}$ defaults to the local model for inference and then adaptively switches to the global model, based on the rate of change in the discrepancy between the global model's average accuracy $\mathrm{ACC}_{k,\mathrm{BF}}^t$ and the average prediction probability of the ground-truth class $\mathrm{PROB}_{k,\mathrm{BF}}^t$, both computed on the representative local buffer $\mathcal{M}_k^t$ after client k receives the global model θ_g^t from the server at the end of FL round t:

$$\mathrm{ACC}_{k,\mathrm{BF}}^t = \frac{1}{|\mathcal{M}_k^t|} \sum_{(x,y) \in \mathcal{M}_k^t} \mathbb{I}\left(\hat{y}_g^t(x) = y\right); \quad \mathrm{PROB}_{k,\mathrm{BF}}^t = \frac{1}{|\mathcal{M}_k^t|} \sum_{(x,y) \in \mathcal{M}_k^t} z_g^t(x)[y] \tag{13}$$

where $z_g^t(x)$ and $\hat{y}_g^t(x)$ denote the prediction probability and the predicted label for input x generated by θ_g^t, respectively. Then, we define the generalization gap GAP_k^t between $\mathrm{ACC}_{k,\mathrm{BF}}^t$ and

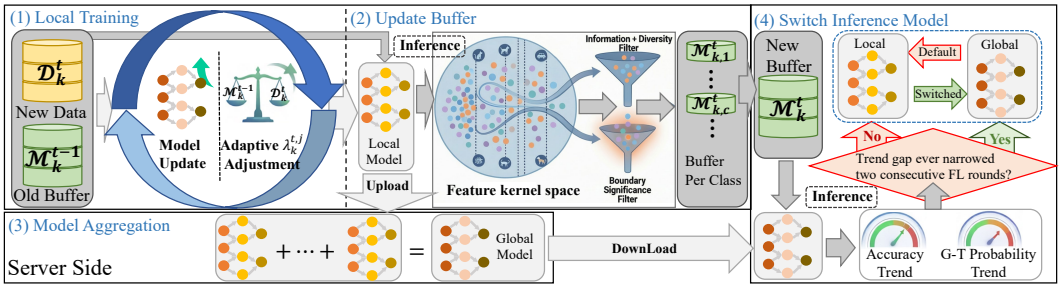

Figure 1: Illustration of the FedKACE. In FL round t, client k trains locally via the **Gradient-Responsive Replay Scheme** (Sec. 4.1), then updates its buffer via the **Holistic Buffer Maintenance Strategy** (Sec. 4.2), subsequently uploads the local model for aggregation, and finally evaluates the new global model via the **Adaptive Inference Model Switching Mechanism** (Sec. 4.3).

$\mathrm{PROB}^t_{k,\mathrm{BF}}$ and its rate of change $\Delta\mathrm{GAP}^t_k$:

$$\mathrm{GAP}^t_k = \max(0, \mathrm{ACC}^t_{k,\mathrm{BF}} - \mathrm{PROB}^t_{k,\mathrm{BF}}); \quad \Delta\mathrm{GAP}^t_k = \mathrm{GAP}^t_k - \mathrm{GAP}^{t-1}_k, \quad \forall t > 1 \tag{14}$$

When $\Delta\mathrm{GAP}^t_k < 0$, the gap between $\mathrm{ACC}^t_{k,\mathrm{BF}}$ and $\mathrm{PROB}^t_{k,\mathrm{BF}}$ is decreasing on the local buffer, indicating that the global model has converged sufficiently to adapt to client k's local data distribution. To ensure robustness, client k switchs its inference model to the global model after this condition holds for two consecutive FL rounds, with the switching FL round t^{sw}_k defined as:

$$t^{\mathrm{sw}}_k = \min\left\{t \mid \Delta\mathrm{GAP}^t_k < 0 \wedge \Delta\mathrm{GAP}^{t-1}_k < 0\right\} \tag{15}$$

When $t < t^{\mathrm{sw}}_k$, client k uses the local model θ^t_k for inference, while for $t \geq t^{\mathrm{sw}}_k$, it uses the global model θ^t_g for inference and stops computing $\mathrm{ACC}^t_{k,\mathrm{BF}}$ and $\mathrm{PROB}^t_{k,\mathrm{BF}}$ to reduce computational cost. The subsequent Theorem 3 proves the necessity of switching the inference model.

4.4 FedKACE Framework

Based on the above, we propose the FedKACE framework: In each FL round t, each client $k \in \mathcal{K}$ updates the previous global model θ^{t-1}_g on its buffer $\mathcal{M}^{t-1}_k$ and new data chunk $\mathcal{D}^t_k$ via the Gradient-Responsive Replay Scheme and obtains $\theta^{t,J}_k$. Then, client k updates its buffer to $\mathcal{M}^t_k$ from $\mathcal{D}^t_k \cup \mathcal{M}^{t-1}_k$ via the Holistic Buffer Maintenance Strategy. Next, client k uploads $\theta^{t,J}_k$ for server aggregation, yielding the global model $\theta^t_g = \frac{1}{K}\sum_{k=1}^K \theta^{t,J}_k$. Finally, using the Adaptive Inference Model Switching Mechanism, client k evaluates θ^t_g on $\mathcal{M}^t_k$ to decide whether to switch its inference model from $\theta^{t,J}_k$ to θ^t_g, thereby concluding the t-th FL round.

Building on the FedKACE framework,we present the following Theorem 3:

Theorem 3 (Regret Gap between Global and Local Models in FedKACE) *Under our problem formulation and standard assumptions (see Appendix B.1), let the optimal model for client k under the local cumulative data distribution $p^{\leq t}_k$ be $f^{t,*}_{k,\mathrm{cum}} \triangleq \arg\min_{\theta\in\Omega_\theta} \mathbb{E}_{p^{\leq t}_k}[\mathcal{R}(f_\theta)]$, and define $\mathrm{Regret}^t_k(\theta) \triangleq \mathbb{E}_{p^{\leq t}_k}\left[\mathcal{R}(f_\theta) - \mathcal{R}(f^{t,*}_{k,\mathrm{cum}})\right]$ as the regret of model θ on client k at FL round t, the regret gap between the local model $\theta^{t,J}_k$ and the global model θ^t_g satisfies:*

$$\mathop{\mathbb{E}}_{k\in\mathcal{K}}\left[\mathrm{Regret}^t_k(\theta^t_g) - \mathrm{Regret}^t_k(\theta^{t,J}_k)\right] \leq \frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu) - \frac{\mu}{2K}\sum_{k=1}^K \|\theta^{t,J}_k - \theta^t_g\|_2^2 \tag{16}$$

where $\Delta^{\mathrm{TV}}(t,\nu)$ globally bounds the cross-client TV distance $\|p^{\leq t}_k - p^{\leq t}_{k'}\|_{\mathrm{TV}}$ for all $k, k' \in \mathcal{K}$ at FL round t, decaying to a constant as t increases, with definition detailed in Appendix B.7. $\nu = n_{\min}/n_{\max}$ is the ratio of the minimum to the maximum per-round data size, defined as $n_{\min} \triangleq \min_{k\in\mathcal{K},\tau\leq T}(n^\tau_k)$ and $n_{\max} \triangleq \max_{k\in\mathcal{K},\tau\leq T}(n^\tau_k)$. For small t, $\Delta^{\mathrm{TV}}(t,\nu)$ dominates, indicating that the local model $\theta^{t,J}_k$ is superior; for large t, the negative model discrepancy term dominates, favoring the global model θ^t_g.

The proof of Theorem 3 is in Appendix B.7.

5 EXPERIMENTS

5.1 EXPERIMENT SETTINGS

Datasets: On CIFAR100 Krizhevsky & Hinton (2009) and ImageNet100 (first 100 classes of ILSVRC2012 Krizhevsky et al. (2012)), we conduct experiments with $C_{\max} = 100$ global classes, total $K = 10$ clients and total $T = 100$ FL rounds. Before training, each client independently and randomly permutes all 100 classes into a cyclic list, and pre-defines its per-round class sets via a sliding window of size 5 with stride $5 - O$, where $O \in \{5, 4, 2, 0\}$ (i.e., **O**verlap) denotes the class overlap between adjacent FL rounds. When $O = 5$, the window remains fixed for 5 consecutive FL rounds and then advances by 5 classes, simulating static class sets in OCL. For $O \in \{4, 2, 0\}$, the window strides by $1, 3, 5$ classes per FL round along the cyclic list with wrap-around, thereby modeling varying class overlap across consecutive FL rounds. In each FL round, every client trains for $J = 20$ epochs on 5 window-determined classes using 100 samples (i.e., (x, y) pairs) per class without explicit task identifiers. The local data are non-overlapping for each client across all its FL rounds, but the finite dataset size causes cross-client sample overlap.

Evaluation Metric: As the forgetting rate is inapplicable without task identifiers, we adopt accuracy (ACC) and regret (REG) from online learning as metrics. After each FL round t, each client computes its accuracy on the standard test set of its classes encountered thus far, and the ACC^t is averaged across all clients. To compute regret, we establish a Centralized baseline as an upper bound, where clients independently train newly initialized models for $J = 20$ epochs on all samples from FL rounds 1 to t without global aggregation. The regret for an evaluated method is then defined as $\text{REG}^t_{\text{method}} = \text{ACC}^t_{\text{Centralized}} - \text{ACC}^t_{\text{method}}$, denoting the accuracy gap between the method and the Centralized baseline. The overall evaluation metrics for a method are Average Accuracy (AA), defined as $\text{AA}_{\text{method}} = \frac{1}{T}\sum_{t=1}^{T} \text{ACC}^t_{\text{method}}$, and Average Regret (AR), defined as $\text{AR}_{\text{method}} = \frac{1}{T}\sum_{t=1}^{T} \text{REG}^t_{\text{method}}$.

Baseline Methods: We evaluate seven baselines: the classical FL method FedAVG McMahan et al. (2017); five FCL methods, including the parameter isolation-based TFCL Wang et al. (2024b), the dynamic collaboration-based DCFCL Yang et al. (2025), and three sample replay-based methods, namely Re-Fed Li et al. (2024), OFCL Serra & Buettner (2025) and FedCBDR Qi et al. (2025); and the Centralized method as an upper bound, with Re-Fed, OFCL, FedCBDR and our FedKACE employing a local buffer capacity of $M = 1000$. All methods use ResNet-18 He et al. (2016) as model. We exclude generative replay and prompt-based FCL methods, as they underperform sample replay-based methods Li et al. (2025a). The remaining FCL baselines use the FL round index as the explicit task identifier.

More details are provided in Appendix A.

5.2 BASELINE EXPERIMENTS

Under the setup above, Table 1 shows results for baseline methods and FedKACE, while Figure 2 illustrates accuracy trends of all methods across all $T = 100$ FL rounds. All baseline methods exhibited performance limitations in streaming FCL setting: FedAVG lacks mechanisms to handle dynamic class evolution, consequently failing to retain historical knowledge; TFCL's parameter

Table 1: Experiment Results of Baseline Methods

Method	CIFAR100								ImageNet100							
	O=5		O=4		O=2		O=0		O=5		O=4		O=2		O=0	
	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓
FedAVG	5.99	45.55	4.36	44.10	4.18	32.27	4.39	30.80	5.34	48.91	4.39	46.88	3.58	36.00	4.31	33.01
TFCL	3.58	47.95	3.12	45.35	1.63	34.83	1.36	33.83	3.51	50.74	3.18	48.08	1.78	37.80	1.53	35.79
DCFCL	4.10	47.44	3.59	44.88	1.85	34.61	1.54	33.65	4.01	50.24	3.67	47.59	2.03	37.55	1.74	35.58
Re-Fed	8.61	42.92	7.29	41.08	8.61	27.84	10.12	25.07	6.74	47.50	5.59	45.67	7.28	32.30	9.65	27.67
OFCL	16.91	34.63	16.41	32.05	12.44	24.01	13.40	21.79	15.98	38.26	15.90	35.37	16.54	23.04	16.59	20.73
FedCBDR	21.91	29.62	20.73	27.73	18.44	18.01	19.09	16.10	19.01	35.23	18.90	32.37	18.53	21.05	18.25	19.07
FedKACE	**26.59**	**24.95**	**26.74**	**21.73**	**22.03**	**14.43**	**20.96**	**14.23**	**22.82**	**31.43**	**24.39**	**26.87**	**22.28**	**17.30**	**21.65**	**15.67**
Centralized	51.53	0.00	48.47	0.00	36.45	0.00	35.19	0.00	54.25	0.00	51.27	0.00	39.58	0.00	37.32	0.00

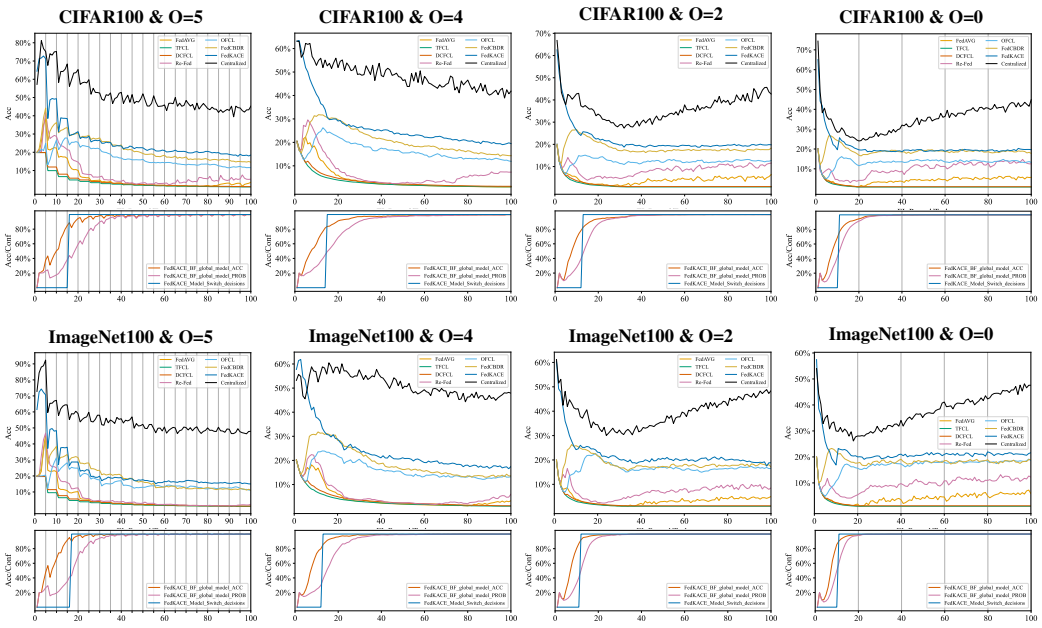

Figure 2: Trends in Method Accuracy and FedKACE Buffer Metrics on Baseline Experiments.

isolation mechanism fails to integrate inter-class knowledge across FL rounds, fragmenting model knowledge; DCFCL's coalitional affinity game fails to reach stable equilibria due to incomplete class coverage, undermining knowledge retention; Re-Fed's static sample importance estimation mechanism largely fails to adapt to continually shifting distributions, reducing historical sample representativeness; OFCL's Bregman Information-based uncertainty estimation partially fails under persistent distribution drift, compromising the selection of per-class representative samples; FedCBDR's feature encryption-reconstruction mechanism via matrix decomposition slightly fails to maintain unbiased representations, hindering intra-class feature discrimination. In contrast, FedKACE eliminates dependency on explicit task boundaries through its three synergistic mechanisms, thereby achieving the highest average accuracy and the lowest average regret.

5.3 Ablation Studies

We conduct ablation studies under the $O = 5$ and $O = 0$ settings. For the Adaptive Inference Model Switching Mechanism, we evaluate: FedKACE_AS1(AS1), which triggers switching when $\Delta(\text{gap}_k^t) < 0$ is first satisfied; FedKACE_AS2(AS2) and FedKACE_AS3(AS3), which exclusively use the global and local models, respectively. For the Gradient-Responsive Replay Scheme, we examine: FedKACE_AS4(AS4), which fixes $\lambda_k^{t,j} = 1$; and FedKACE_AS5(AS5), which uses the full model gradient to compute $\lambda_k^{t,j}$. For the Holistic Buffer Maintenance Strategy, we analyze: FedKACE_AS6(AS6), which removes CDV and uses only IDV-weighted sampling; and FedKACE_AS7(AS7), which replaces HBMS with class-balanced random sampling. For the complete FedKACE framework, we test: FedKACE_AS8(AS8), which preserving adaptive inference model switching but fixing $\lambda_k^{t,j} = 1$ and using class-balanced random sampling; and LocalKACE(LKC), which in-

Table 2: Experiment Results of Ablation Study

Method	Cifar100				ImageNet100			
	O=5		O=0		O=5		O=0	
	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓
AS1	25.42	26.12	**21.03**	**14.17**	22.05	32.20	**21.72**	**15.60**
AS2	23.20	28.34	19.82	15.37	20.21	34.04	20.79	16.53
AS3	24.23	27.30	16.81	18.38	22.03	32.22	16.17	21.14
AS4	25.83	25.70	20.32	14.87	21.40	32.85	20.24	17.08
AS5	**26.63**	**24.90**	21.00	14.19	**22.89**	31.36	**21.67**	15.65
AS6	26.54	25.00	20.24	14.95	22.23	32.02	20.98	16.33
AS7	26.47	25.06	20.21	14.99	22.34	31.91	20.18	17.14
AS8	24.72	26.81	17.75	17.44	19.17	35.08	15.52	21.80
LKC	18.78	32.75	13.17	22.02	18.23	36.02	11.47	25.85
FKC	26.59	24.95	20.96	14.23	22.82	31.43	21.65	15.67
CTR	51.53	0.00	35.19	0.00	54.25	0.00	37.32	0.00

Table 3: AS Variants' Average local Buffer Kernel Condition Numbers in Final 75 FL Rounds

Method	Cifar100		ImageNet100	
	O=5	O=0	O=5	O=0
FedKACE_AS6	1575.6	1055.4	1638.6	396.8
FedKACE_AS7	1637.3	1126.8	1885.4	478.4
FedKACE	1593.4	1085.5	1711.6	421.2

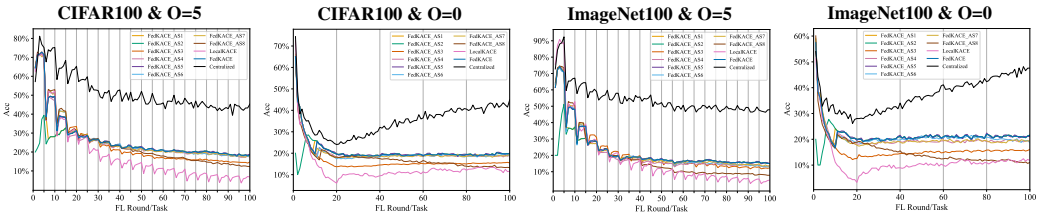

Figure 3: Trends in Method Accuracy on Ablation Studies.

Table 4: Experimental Results of FedKACE with Varying Buffer Capacity M and Client Count K

Method	CIFAR100								ImageNet100							
	O=5		O=4		O=2		O=0		O=5		O=4		O=2		O=0	
	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓	AA↑	AR↓
K10_M500	24.60	26.94	25.19	23.27	19.53	16.93	18.99	16.20	21.21	33.03	23.26	28.00	21.67	17.91	20.11	17.21
K10_M1000	26.59	24.95	26.74	21.73	22.03	14.43	20.96	14.23	22.82	31.43	24.39	26.87	22.28	17.30	21.65	15.67
K10_M2000	27.72	23.82	28.21	20.26	23.15	13.31	23.82	11.37	23.98	30.27	25.43	25.84	24.46	15.12	24.19	11.13
K20_M1000	27.07	24.46	28.23	20.24	23.05	13.41	21.88	13.32	23.74	30.51	24.74	26.53	23.30	16.28	22.69	14.62
Centralized	51.53	0.00	48.47	0.00	36.45	0.00	35.19	0.00	54.25	0.00	51.27	0.00	39.58	0.00	37.32	0.00

corporates all FedKACE components except federated aggregation, relying solely on local training.

Table 2 shows ablation results (FedKACE as FKC, Centralized as CTR), while Figure 3 illustrates accuracy trends of all variants, and Table 3 reports the average buffer kernel condition numbers for AS5, AS6, and FedKACE over the last 75 FL rounds. AS1 achieves marginal improvement in $O = 0$ but suffers significant degradation in $O = 5$ due to premature switching before model stabilization. AS2 and AS3 underperform by only using global or local models for inference, respectively, forfeiting adaptive model selection benefits. AS4, which fixes $\lambda_k^{t,j} = 1$, cannot accommodate dynamically evolving knowledge and degrades performance. AS5 computes $\lambda_k^{t,j}$ using full model gradients, yielding minimal gains while increasing gradient FLOPs per sample from 12.8K to 35.68M ($\sim 2780\times$ on CIFAR) and 51.2K to 1823.57M ($\sim 35600\times$ on ImageNet). AS6 only uses IDV-weighted sampling, slightly reducing buffer kernel condition number but failing to identify decision boundary-critical samples, consequently compromising performance. AS7 uses class-balanced random sampling, yielding the highest kernel condition number and poor accuracy due to inefficient buffer management. Finally, AS8 and LKC both underperform, demonstrating the necessity of integrating the components of FedKACE.

5.4 Extend Research

We analyze FedKACE under varying buffer capacities M (default 1000) and client counts K (default 10). Experimental results in Table 4 confirm that while increasing either M or K improves performance, expanding M yields greater gains than adding K (e.g., $M = 2000, K = 10$ outperforms $M = 1000, K = 20$ for an identical $K \times M$ product), indicating more substantial gains from buffer expansion than client addition.

6 Conclusion

In this paper, we address streaming FCL where data chunks lack task identifiers and exhibit arbitrary class overlap, posing three unique challenges: resolving knowledge conflicts of the same class without task boundaries, preserving boundary-critical knowledge in constrained buffers, and performing local continual inference on cumulative classes. To address these, we propose FedKACE with three synergistic components: First, a gradient-responsive replay scheme uses the squared L2 gradient norm ratio of buffer to new samples, balancing client-specific knowledge acquisition and retention. Second, a holistic buffer maintenance strategy employs a two-stage screening process to jointly optimize the informativeness, boundary significance, and feature dispersion of buffer samples, enhancing knowledge retention under class overlap. Finally, an adaptive mechanism monitors the global model's performance on the local buffer to determine when to switch the inference model from local to global, improving client-side continual inference. Theoretical analysis and experiments under diverse overlap settings validate FedKACE.

REFERENCES

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 451–459, 2011.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

Mohamed Elsayed and A. Rupam Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for expected calibration error. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, 2016.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4), 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 1106–1114, 2012.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Mach. Learn.*, 106(1):93–117, 2017.

Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Process. Mag.*, 37(3):50–60, 2020.

Yichen Li, Qunwei Li, Haozhao Wang, Ruixuan Li, Wenliang Zhong, and Guannan Zhang. Towards efficient replay in federated incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 12820–12829. IEEE, 2024.

Yichen Li, Yuying Wang, Jiahua Dong, Haozhao Wang, Yining Qi, Rui Zhang, and Ruixuan Li. Resource-constrained federated continual learning: What does matter? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025b.

Sihao Liu, Yibo Yang, Xiaojie Li, David A. Clifton, and Bernard Ghanem. Enhancing online continual learning with plug-and-play state space model and class-conditional mixture of discretization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20502–20511. Computer Vision Foundation / IEEE, 2025.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Nicolas Michel, Maorong Wang, Ling Xiao, and Toshihiko Yamasaki. Rethinking momentum knowledge distillation in online continual learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Zhuang Qi, Ying-Peng Tang, Lei Meng, Han Yu, Xiaoxiao Li, and Xiangxu Meng. Class-wise balancing data replay for federated class-incremental learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Krishnan Raghavan and Prasanna Balaprakash. Formalizing the generalization-forgetting trade-off in continual learning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17284–17297, 2021.

Xuankun Rong, Jianshu Zhang, Kun He, and Mang Ye. CAN: leveraging clients as navigators for generative replay in federated continual learning. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025.

Riccardo Salami, Pietro Buzzega, Matteo Mosconi, Jacopo Bonato, Luigi Sabetta, and Simone Calderara. Closed-form merging of parameter-efficient modules for federated continual learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Giuseppe Serra and Florian Buettner. Federated continual learning goes online: Uncertainty-aware memory management for vision tasks and beyond. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Ruilin Tong, Yuhang Liu, Javen Qinfeng Shi, and Dong Gong. Coreset selection via reducible loss in continual learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Minh-Tuan Tran, Trung Le, Xuan-May Le, Mehrtash Harandi, and Dinh Phung. Text-enhanced data-free approach for federated class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 23870–23880. IEEE, 2024.

Anastasiia Usmanova, François Portet, Philippe Lalanda, and Germán Vega. A distillation-based approach integrating continual learning and federated learning for pervasive services. *CoRR*, abs/2109.04197, 2021.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5362–5383, 2024a.

Qiang Wang, Bingyan Liu, and Yawen Li. Traceable federated continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 12872–12881. IEEE, 2024b.

Xinrui Wang, Chuanxing Geng, Wenhai Wan, Shao-Yuan Li, and Songcan Chen. Forgetting, ignorance or myopia: Revisiting key challenges in online continual learning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024c.

Feng Wu, Siwei Feng, Yuanlu Chen, and Libang Zhao. Personalized federated class-incremental learning through critical parameter transfer. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025.

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, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Danni Yang, Zhikang Chen, Sen Cui, Mengyue Yang, Ding Li, Abudukelimu Wuerkaixi, Haoxuan Li, Jinke Ren, and Mingming Gong. Decentralized dynamic cooperation of personalized models for federated continual learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xin Yang, Hao Yu, Xin Gao, Hao Wang, Junbo Zhang, and Tianrui Li. Federated continual learning via knowledge fusion: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(8): 3832–3850, 2024.

Fei Ye and Adrian G. Bors. Online task-free continual learning via dynamic expansionable memory distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 20512–20522. Computer Vision Foundation / IEEE, 2025.

Yuhao Zhou, Yuxin Tian, Jindi Lv, Mingjia Shi, Yuanxi Li, Qing Ye, Shuhao Zhang, and Jiancheng Lv. Ferret: An efficient online continual learning framework under varying memory constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 4850–4861. Computer Vision Foundation / IEEE, 2025.

A Datasets, Baseline Methods and Experimental Setup

A.1 Details of Datasets

We conduct experiments on public datasets. Given the finite dataset size, we allow identical samples to be accessed by different clients across different FL rounds. Specifically, across all T rounds, each sample is accessed by at most K distinct clients, limiting its total system entries to no more than K. In contrast, traditional FCL methods employ static data partitioning where a single client exclusively holds each sample, but since each task spans far more than K FL rounds, that client repeatedly trains on the same sample, yielding a usage count far exceeding K. Thus, although our setting relaxes data exclusivity, it still imposes a stricter constraint than the traditional FCL setting by strictly limiting each sample's overall reuse frequency.

CIFAR100: The CIFAR100 dataset Krizhevsky & Hinton (2009) is a real-world image classification dataset consisting of RGB images organized into 20 superclasses with 5 classes each, totaling 100 classes, where each class contains 500 training images and 100 test images. In our experimental setup, for each class assigned to a client in an FL round, the client randomly selects 100 unaccessed training images from the 500 available samples per class, with no sample repetition across its FL rounds. After each FL round, each client evaluates its accuracy using the standard test set (all 100 images per class) from all its classes encountered thus far.

ImageNet100: The ImageNet100 dataset consists of the first 100 classes from the ILSVRC2012 dataset Krizhevsky et al. (2012), with 1300 samples per class, where we designate the first 1000 samples as training images and the remaining 300 as test images. Prior to training, each client independently and randomly selects a fixed set of 100 samples from the 300 test samples per class to construct its unique standard test set. In our experimental setup, for each class assigned to a client in an FL round, the client randomly selects 100 unaccessed training images from the 1000 available samples per class, with no sample repetition across its FL rounds. After each FL round, each client evaluates its accuracy using the standard test set (all 100 images per class) from all classes encountered thus far.

A.2 Details of Baseline Methods

We select various FL and FCL methods for comparison, where the FL round index is provided as an explicit task identifier for each FCL method to adapt them to our setting.

FedAVG: FedAVG McMahan et al. (2017) is a representative federated learning method. In each FL round, the server first broadcasts the current global model to all clients; each client then performs local training initialized from this model and uploads its updated parameters to the server. Subsequently, the server aggregates all local parameters through weighted averaging to obtain the updated global model, where each weight equals the ratio of the client's local dataset size to the total dataset size across all clients.

TFCL: TFCL Wang et al. (2024b) is a parameter-isolation-based federated continual learning approach that addresses the limitations of conventional FCL methods that fail to account for task recurrence. At the onset of each task, TFCL decomposes the entire model into tagged sub-models through its Traceable Task Learning (TTL) module, thereby enabling the system to precisely locate and refine corresponding features from previous tasks when recurring tasks emerge; additionally, its Group-wise Knowledge Aggregation (GKA) module selectively federates recurring tasks by grouping similar tasks and aggregating their knowledge on the server side via client-server knowledge distillation across multiple FL rounds.

DCFCL: DCFCL Yang et al. (2025) is a decentralized federated continual learning method driven by dynamic cooperation that addresses catastrophic forgetting across temporal and spatial dimensions. In each FL round, clients form non-overlapping coalitions through a coalitional affinity game that quantifies the benefits derived from cooperation by evaluating gradient coherence and model similarity; subsequently, a cooperative equilibrium, where no alternative coalition yields greater benefits for any participant, is achieved via a merge-blocking algorithm integrated with dynamic cooperative evolution, thereby enabling personalized models to effectively balance new knowledge acquisition and prior knowledge retention under heterogeneous data distributions.

Re-Fed: Re-Fed Li et al. (2024) is a sample replay-based federated continual learning method. When a new task arrives, each client first updates a Personalized Informative Model (PIM) that incorporates both global and local information to compute sample importance. Clients quantify sample importance by calculating weighted accumulated gradient norms during PIM updates, and then cache the most valuable samples within their storage limits. Subsequently, clients train their local models using both the cached historical samples and the new task samples. During each FL round, the server aggregates the updated local models to form a new global model. When the next task arrives, clients again evaluate and store important samples from all previous tasks using the same PIM mechanism for replay in subsequent tasks.

OFCL: OFCL Serra & Buettner (2025) is a sample replay-based federated online continual learning method , where all clients process streaming data restricted to the same task-specific classes within each task. When a new task arrives, clients process streaming mini-batches of data and employ Bregman Information to estimate epistemic uncertainty at the sample level. Based on these estimates, clients selectively store representative samples with low epistemic uncertainty for each class into local buffers, and then train on both current and buffered historical data. Upon task transitions, clients refresh their buffers using the same uncertainty-based selection mechanism to retain knowledge for subsequent tasks.

FedCBDR: FedCBDR Qi et al. (2025) is a sample replay-based federated continual learning method. During each FL round of every task, clients employ a task-aware temperature scaling module to adaptively adjust the temperature of logits and reweight the learning objective at both the class and instance levels, thereby jointly training local models with historical samples and new task data; the server then aggregates these updated local models to construct a new global model. Upon task completion, clients obfuscate local features via ISVD and upload them to the server, which aggregates these matrices, performs SVD to reconstruct global representations of historical tasks while preserving privacy, and then conducts class-aware and importance-sensitive balanced sampling based on leverage scores to determine the indices of historical samples for retention, enabling clients to update their replay buffers accordingly.

A.3 Details of Experimental Setup

The following are the specific details of the experimental setup.

Model Configuration: To accommodate dataset-specific characteristics, randomly initialized ResNet-18 models He et al. (2016) are configured with an initial channel width of 16 for CIFAR-100 and 64 for ImageNet-100. All models employ an output layer with a fixed output dimension of 100, adapted to streaming classes via a dynamic masking mechanism: at FL round t, client k activates only the output neurons corresponding to its accumulated class set $\mathcal{C}_k^{\leq t} = \bigcup_{\tau=1}^{t} \mathcal{C}_k^{\tau}$, while masking predictions and setting gradients to zero for all other neurons.

Local Training Protocol: For all methods, during the local training phase of each FL round, clients employ the AdamW optimizer (weight decay = 0.001) with a batch size of 32 and a fixed number of $J = 20$ training epochs; the learning rate decays from 0.01 to 0 via a cosine annealing schedule where the decay period is synchronized with the local training epochs. For FedKACE, the adaptive inference model switching mechanism does not interfere with local training. Regarding hyperparameter configuration, all baseline FCL methods strictly adhere to the fixed hyperparameters specified in their original papers, whereas FedKACE adaptively adjusts its hyperparameters through data-driven mechanisms, thereby eliminating the need for manual tuning.

A.4 Assumption Comparison between FCL, OFCL and Streaming FCL

Table 5: Assumption comparison between FCL, OFCL and streaming FCL

Dimension	**FCL & OFCL**	**Streaming FCL (Ours)**
Task ID & Boundary	Explicit task ID required. Well-defined task boundaries.	No task ID. No task boundaries.
Temporal Constraint	One task spans multiple FL rounds. Full data available throughout the task duration (FCL). Continuous data stream supporting sufficient iterations per task (OFCL).	One data chunk per FL round. Ephemeral data access. Limited iterations (J epochs).
Class Space	Per-task class subset of global set $\mathcal{C}_g$. Identical across all clients. Mutually exclusive across tasks.	Global class set $\mathcal{C}_g$ known. Local class sets differ across clients. Arbitrary class overlap across chunks.
Inference	Task ID identifies target classes. Inference performed only upon task completion.	Cumulative inference on all seen classes. Inference performed after every FL round.
Buffer Management	Task-wise allocation. Intra-task reuse. Fixed class set per task.	Unified management. Must disentangle old vs. new samples of the same class. Operates under evolving class sets.

B THEOREM PROOF

This proof contains three theorems and three lemmas, where Lemma 1 supports the proof of Theorem 1, and Lemmas 2 and 3 support the proof of Theorem 3.

B.1 ASSUMPTIONS

Assumption 1 (Bounded Gradients of Local Loss Functions): For all clients $k = 1, \ldots, K$, all FL rounds $t = 1, \ldots, T$, and all $\theta \in \Omega_\theta$, the local loss functions $L^t_{k,\text{task}}(\theta)$ and $L^t_{k,\text{rep}}(\theta)$ are continuously differentiable, and there exist constants $L_1, L_2 > 0$ such that the Euclidean norms of their gradients are bounded by:

$$\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2 \leq L_1, \quad \|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2 \leq L_2 \tag{17}$$

Assumption 2 (Compactness of Parameter Spaces): Both the model parameter space Ω_θ and the replay weight space Ω_λ are compact sets.

Assumption 3 (Robbins-Monro Condition for Local Learning Rates): The sequence of local learning rates $\{\eta^{t,j}_k\}$ satisfies $\sum_{j=1}^{\infty} \eta^{t,j}_k = \infty$ and $\sum_{j=1}^{\infty} (\eta^{t,j}_k)^2 < \infty$. In practical implementations, a sufficiently large finite number of local iterations J is used to approximate this asymptotic condition, ensuring approximate convergence of the local model.

Assumption 4 (Compactness of the Range of the Normalized Logit Vector): The range of the normalized logit vector $\hat{g}(x)$ constitutes a compact set $\mathcal{G} \subseteq \mathbb{R}^d$, where the effective dimension $d = |\mathcal{C}_k^{\leq t}|$ is defined as the size of the class set.

Assumption 5 (Lipschitz Continuity and Boundedness of the Loss Function): The loss function $\ell(f(x), y)$ is L_{Lip}-Lipschitz continuous and is bounded within $[0, 1]$.

Assumption 6 (Uniformity of Samples per Class in the Buffer): In the t-th FL round, for any client k, the number of samples in the buffer for each class $c \in \mathcal{C}_k^{\leq t}$ is given by $|\mathcal{M}^t_{k,c}| = \Theta(M/|\mathcal{C}_k^{\leq t}|)$.

Assumption 7 (Finiteness of the Number of Global Classes): The total number of global classes is finite, as formalized by $\left|\bigcup_{k=1}^{K} \bigcup_{\tau=1}^{\infty} \mathcal{C}_k^\tau\right| = C_{\max} < \infty$.

Assumption 8 (Client Distribution Shift Constraint): For any client k, the data distributions p_k^t and p_k^{t-1} in consecutive FL rounds t and $t-1$ satisfy $D_{\text{TV}}(p_k^t, p_k^{t-1}) \leq C_{\max}/t^\alpha$, where D_{TV} is the total variation distance and $\alpha > 0$ denotes the decay rate of the distribution shift.

Assumption 9 (Static Generalization Bound for Buffer Sampling): For any client $k \in \mathcal{K}$ and any FL round t, assume the buffer $\mathcal{M}_k^t$ contains M samples. For any model $f \in \mathcal{F}$, the difference between the empirical risk on the buffer $\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f)$ and the expected risk under the buffer distribution $\mathcal{R}_{p_{\mathcal{M}_k^t}}(f)$ satisfies the following explicit bound:

$$\left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f) - \mathcal{R}_{p_{\mathcal{M}_k^t}}(f)\right| \leq \frac{C_{\text{rep}}}{\sqrt{M}} \tag{18}$$

where the constant $C_{\text{rep}} = \sqrt{2}\left(4L_{\text{Lip}}B + \sqrt{8\log(2/\delta)}\right)$, L_{Lip} is the Lipschitz constant of the loss function, B is the RKHS norm bound, and δ is the confidence parameter. This bound solely captures the statistical fluctuations induced by finite sampling.

Assumption 10 (L-smoothness of the local loss function): Let $L^t_{k,\text{total}}$ abstractly denote the local loss function for any client $k \in \mathcal{K}$ at FL round t, for any model parameters θ_a and θ_b, the following inequality holds:

$$L^t_{k,\text{total}}(\theta_a) \leq L^t_{k,\text{total}}(\theta_b) + \langle \nabla_\theta L^t_{k,\text{total}}(\theta_b), \theta_a - \theta_b \rangle + \frac{L}{2}\|\theta_a - \theta_b\|_2^2 \tag{19}$$

Assumption 11 (μ-strong convexity of the local loss function): Let $L^t_{k,\text{total}}$ abstractly denote the local loss function for any client $k \in \mathcal{K}$ at FL round t, for any model parameters θ_a and θ_b, the

following inequality holds:

$$L^t_{k,\text{total}}(\theta_a) \geq L^t_{k,\text{total}}(\theta_b) + \langle \nabla_\theta L^t_{k,\text{total}}(\theta_b), \theta_a - \theta_b \rangle + \frac{\mu}{2}\|\theta_a - \theta_b\|_2^2 \tag{20}$$

Assumption 12 (Bounded Local Gradient Variance): Let $L^t_{k,\text{total}}$ abstractly denote the local loss function for any client $k \in \mathcal{K}$ at FL round t, given that $\xi_k^{t,j}$ is uniformly sampled from the local data of client k at epoch j, the variance of the stochastic gradient satisfies the following bound:

$$\mathbb{E}\left[\|\nabla_\theta L^t_{k,\text{total}}(\theta_k^{t,j}, \xi_k^{t,j}) - \nabla_\theta L^t_{k,\text{total}}(\theta_k^{t,j})\|_2^2\right] \leq \sigma_k^2 \tag{21}$$

Assumption 13 (Bounded Local Gradient Norm): Let $L^t_{k,\text{total}}$ abstractly denote the local loss function for any client $k \in \mathcal{K}$ at FL round t, given that $\xi_k^{t,j}$ is uniformly sampled from the local data of client k at epoch j, the expected squared ℓ_2-norm of the stochastic gradient satisfies the following bound:

$$\mathbb{E}\left[\|\nabla_\theta L^t_{k,\text{total}}(\theta_k^{t,j}, \xi_k^{t,j})\|_2^2\right] \leq G^2 \tag{22}$$

B.2 PROOF OF LEMMA 1

Lemma 1 (Existence and Asymptotic Stability of the Ideal Saddle Point in Local Training) *Under standard assumptions (see Appendix B.1), let* $L^t_{k,total}(\theta^{t,j}_k, \lambda^{t,j}_k) = L^t_{k,task}(\theta^{t,j}_k) + \lambda^{t,j}_k L^t_{k,rep}(\theta^{t,j}_k)$ *be the total loss function of client* k *at the* j*-th iteration step of the* t*-th FL round, where the replay loss weight* $\lambda^{t,j+1}_k$ *is updated via full-model gradients as* $\lambda^{t,j+1}_k = \|\nabla_\theta L^t_{k,rep}(\theta^{t,j}_k)\|_2^2 / \|\nabla_\theta L^t_{k,task}(\theta^{t,j}_k)\|_2^2$*. Under these conditions, there exists an ideal saddle point* $(\theta^{t,*}_k, \lambda^{t,*}_k) \in \Omega_\theta \times \Omega_\lambda$ *that exactly satisfies the minimax equilibrium condition:*

$$L^t_{k,total}(\theta^{t,*}_k, \lambda) \le L^t_{k,total}(\theta^{t,*}_k, \lambda^{t,*}_k) \le L^t_{k,total}(\theta, \lambda^{t,*}_k), \quad \forall \theta \in \Omega_\theta, \lambda \in \Omega_\lambda \tag{23}$$

Furthermore, as the number of local training steps J *approaches infinity, the sequence of iterates generated by the aforementioned update rule asymptotically converges to this saddle point within its neighborhood.*

Proof:

We establish the existence of the ideal saddle point by leveraging the theoretical framework in Raghavan & Balaprakash (2021). According to Assumption 11, $L^t_{k,\text{total}}(\theta, \cdot)$ is convex with respect to λ. Furthermore, based on the specific construction of $L^t_{k,\text{total}}$, its second-order partial derivative with respect to λ is zero (i.e., $\partial^2_{\lambda\lambda} L^t_{k,\text{total}}(\theta, \lambda) = 0$), which implies that $L^t_{k,\text{total}}(\theta, \cdot)$ is actually an affine function with respect to λ. Additionally, Assumption 2 ensures that the domain $\Omega_\theta \times \Omega_\lambda$ is a compact set. These properties satisfy the prerequisites of Theorem 1 in Raghavan & Balaprakash (2021), guaranteeing the existence of a saddle point $(\Delta x^*, \theta^*)$ such that the value function $H(\cdot, \cdot)$ satisfies:

$$H(\Delta x, \theta^*) \le H(\Delta x^*, \theta^*) \le H(\Delta x^*, \theta), \quad \forall \theta \in \Omega_\theta, \Delta x \in \Omega_\lambda \tag{24}$$

In the context of this paper, by mapping the value function H to $L^t_{k,\text{total}}$, the variable Δx to λ, and keeping θ as θ, we can conclude that there exists $(\theta^{t,*}_k, \lambda^{t,*}_k) \in \Omega_\theta \times \Omega_\lambda$ satisfying:

$$L^t_{k,\text{total}}(\theta^{t,*}_k, \lambda) \le L^t_{k,\text{total}}(\theta^{t,*}_k, \lambda^{t,*}_k) \le L^t_{k,\text{total}}(\theta, \lambda^{t,*}_k), \quad \forall \theta \in \Omega_\theta, \lambda \in \Omega_\lambda \tag{25}$$

Next, we prove the asymptotic stability of the saddle point $(\theta^{t,*}_k, \lambda^{t,*}_k)$. By definition, the total loss function for local training is given by $L^t_{k,\text{total}}(\theta^{t,j}_k, \lambda^{t,j}_k) = L^t_{k,\text{task}}(\theta^{t,j}_k) + \lambda^{t,j}_k L^t_{k,\text{rep}}(\theta^{t,j}_k)$. Utilizing the first-order necessary condition at the saddle point $(\theta^{t,*}_k, \lambda^{t,*}_k)$, i.e., $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k, \lambda^{t,*}_k) = \mathbf{0}$, we obtain:

$$\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k) + \lambda^{t,*}_k \nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k) = \mathbf{0} \tag{26}$$

Furthermore, since $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k, \lambda^{t,*}_k) = \mathbf{0}$, its inner product with any vector is zero, which implies that $\langle \nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k), \nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k, \lambda^{t,*}_k) \rangle = 0$ and $\langle \nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k), \nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k, \lambda^{t,*}_k) \rangle = 0$. By substituting the expression of the total gradient and expanding the aforementioned inner products, we obtain:

$$\begin{cases} \|\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k)\|_2^2 + \lambda^{t,*}_k \langle \nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k), \nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k) \rangle = 0 \\ \langle \nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k), \nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k) \rangle + \lambda^{t,*}_k \|\nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2^2 = 0 \end{cases} \tag{27}$$

By eliminating the cross inner product term $\langle \nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k), \nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k) \rangle$, we obtain the gradient energy balance property at the saddle point:

$$\|\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k)\|_2^2 - (\lambda^{t,*}_k)^2 \|\nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2^2 = 0 \tag{28}$$

Substituting the optimal replay loss weight at the saddle point, $\lambda^{t,*}_k = \|\nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2^2 / \|\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k)\|_2^2$, into the above equation yields:

$$\|\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k)\|_2^6 = \|\nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2^6 \tag{29}$$

Combined with the non-negativity of the norm, the above equation implies that $\|\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k)\|_2^2 = \|\nabla_\theta L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2^2$, which yields a replay loss weight of $\lambda^{t,*}_k = 1$ at the saddle point. This indicates that in this dual-task setting, the saddle point is given by $(\theta^{t,*}_k, 1)$.

Next, considering that the replay loss weight λ is a function of θ, we define the quotient mapping $\mathcal{Q}(\theta) = \|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2 / \|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2$. Since $\|\nabla_\theta L^t_{k,\text{task}}(\theta^{t,*}_k)\|_2^2 > 0$ at the saddle point, we define $\mathcal{B}_\delta \triangleq \{\theta : \|\theta - \theta^{t,*}_k\|_2 \le \delta\}$ as the δ-neighborhood of $\theta^{t,*}_k$. By the continuity of the gradients and the non-negativity of the norm, it follows that the $\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2$ has a strictly positive lower bound within this neighborhood:

$$\inf_{\theta \in \mathcal{B}_\delta} \|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2 \triangleq \epsilon_{\text{min,task}} \ge 0 \tag{30}$$

Given that $\nabla_\theta \|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2 = 2\nabla^2_\theta L^t_{k,\text{rep}}(\theta)\nabla_\theta L^t_{k,\text{rep}}(\theta)$ and $\nabla_\theta \|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2 = 2\nabla^2_\theta L^t_{k,\text{task}}(\theta)\nabla_\theta L^t_{k,\text{task}}(\theta)$, applying the quotient rule to compute the gradient of the quotient mapping $\mathcal{Q}(\theta)$ yields:

$$\begin{aligned}\nabla_\theta \mathcal{Q}(\theta) =& \frac{2\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2 \nabla^2_\theta L^t_{k,\text{rep}}(\theta)\nabla_\theta L^t_{k,\text{rep}}(\theta)}{\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^4} \\ &- \frac{2\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2 \nabla^2_\theta L^t_{k,\text{task}}(\theta)\nabla_\theta L^t_{k,\text{task}}(\theta)}{\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^4}\end{aligned} \tag{31}$$

Taking the ℓ_2-norm of $\nabla_\theta \mathcal{Q}(\theta)$ and bounding the terms yields:

$$\begin{aligned}\|\nabla_\theta \mathcal{Q}(\theta)\|_2 \le& \frac{2\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2 \|\nabla^2_\theta L^t_{k,\text{rep}}(\theta)\|_2 \|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2}{\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^4} \\ &+ \frac{+2\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2 \|\nabla^2_\theta L^t_{k,\text{task}}(\theta)\|_2 \|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2}{\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^4}\end{aligned} \tag{32}$$

According to Assumption 1, we have $\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2 \le L_1$ and $\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2 \le L_2$. Furthermore, under Assumption 10, it follows that $\|\nabla^2_\theta L^t_{k,\text{task}}(\theta)\|_2 \le L$ and $\|\nabla^2_\theta L^t_{k,\text{rep}}(\theta)\|_2 \le L$. Consequently, the norm of $\nabla_\theta \mathcal{Q}(\theta)$ is bounded. Denoting this upper bound by $L_{\mathcal{Q}}$, we obtain:

$$\|\nabla_\theta \mathcal{Q}(\theta)\|_2 \le \frac{2LL_1L_2(L_1+L_2)}{(\epsilon_{\text{min,task}})^4} \triangleq L_{\mathcal{Q}}, \quad \forall \theta \in \mathcal{B}_\delta \tag{33}$$

That is, the quotient mapping $\mathcal{Q}(\theta)$ is $L_{\mathcal{Q}}$-Lipschitz continuous. Leveraging this quotient mapping, we define the composite objective function $\tilde{L}(\theta) = L^t_{k,\text{task}}(\theta) + \mathcal{Q}(\theta) L^t_{k,\text{rep}}(\theta)$. Computing its Hessian matrix $\nabla^2_\theta \tilde{L}(\theta)$ with respect to θ yields:

$$\begin{aligned}\nabla^2_\theta \tilde{L}(\theta) =& \nabla_\theta (\nabla_\theta L^t_{k,\text{task}}(\theta) + \nabla_\theta \mathcal{Q}(\theta) \cdot L^t_{k,\text{rep}}(\theta) + \mathcal{Q}(\theta) \cdot \nabla_\theta L^t_{k,\text{rep}}(\theta)) \\ =& \nabla^2_\theta L^t_{k,\text{task}}(\theta) + \left[\nabla_\theta \mathcal{Q}(\theta) \cdot (\nabla_\theta L^t_{k,\text{rep}}(\theta))^\top + L^t_{k,\text{rep}}(\theta) \cdot \nabla^2_\theta \mathcal{Q}(\theta)\right] \\ &+ \left[\nabla_\theta L^t_{k,\text{rep}}(\theta) \cdot (\nabla_\theta \mathcal{Q}(\theta))^\top + \mathcal{Q}(\theta) \cdot \nabla^2_\theta L^t_{k,\text{rep}}(\theta)\right]\end{aligned} \tag{34}$$

Since the replay loss weight at the saddle point is $\lambda^{t,*}_k = 1$, the total loss function evaluated at $(\theta, \lambda^{t,*}_k)$ simplifies to $L^t_{k,\text{total}}(\theta, \lambda^{t,*}_k) = L^t_{k,\text{task}}(\theta) + \lambda^{t,*}_k L^t_{k,\text{rep}}(\theta) = L^t_{k,\text{task}}(\theta) + L^t_{k,\text{rep}}(\theta)$, and its Hessian matrix is given by $\nabla^2_\theta L^t_{k,\text{total}}(\theta, \lambda^{t,*}_k) = \nabla^2_\theta L^t_{k,\text{task}}(\theta) + \nabla^2_\theta L^t_{k,\text{rep}}(\theta)$. Consequently, we have:

$$\begin{aligned}\nabla^2_\theta \tilde{L}(\theta) - \nabla^2_\theta L^t_{k,\text{total}}(\theta, \lambda^{t,*}_k) =& \nabla_\theta \mathcal{Q}(\theta) \cdot (\nabla_\theta L^t_{k,\text{rep}}(\theta))^\top + L^t_{k,\text{rep}}(\theta) \cdot \nabla^2_\theta \mathcal{Q}(\theta) \\ &+ \nabla_\theta L^t_{k,\text{rep}}(\theta) \cdot (\nabla_\theta \mathcal{Q}(\theta))^\top + (\mathcal{Q}(\theta) - \lambda^{t,*}_k) \cdot \nabla^2_\theta L^t_{k,\text{rep}}(\theta)\end{aligned} \tag{35}$$

Taking the ℓ_2-norm of the above expression. According to Assumption 1, we have $\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2 \le L_2$. Furthermore, under Assumption 10, it follows that $\|\nabla^2_\theta L^t_{k,\text{rep}}(\theta)\|_2 \le L$.

Consequently, we obtain:

$$\|\nabla_\theta^2 \tilde{L}(\theta) - \nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*})\|_2 \leq \sup_{\theta \in \Omega_\theta} (|\mathcal{Q}(\theta) - \lambda_k^{t,*}|) \cdot L + 2L_\mathcal{Q} L_2 + \sup_{\theta \in \Omega_\theta} (L_{k,\text{rep}}^t(\theta)) \cdot \|\nabla^2 \mathcal{Q}(\theta)\|_2 \tag{36}$$

Since $\lambda_k^{t,*} = \mathcal{Q}(\theta_k^{t,*})$, for any $\theta \in \mathcal{B}_\delta(\theta_k^{t,*})$, we have:

$$|\mathcal{Q}(\theta) - \lambda_k^{t,*}| = |\mathcal{Q}(\theta) - \mathcal{Q}(\theta_k^{t,*})| = \|\mathcal{Q}(\theta) - \mathcal{Q}(\theta_k^{t,*})\|_2 \leq L_\mathcal{Q} \|\theta - \theta_k^{t,*}\|_2 \leq L_\mathcal{Q} \delta \tag{37}$$

Consequently, we obtain $\sup_{\theta \in \mathcal{B}_\delta(\theta_k^{t,*})} |\mathcal{Q}(\theta) - \lambda_k^{t,*}| \leq L_\mathcal{Q}\delta$. Furthermore, under Assumption 5, it follows that $\sup_{\theta \in \Omega_\theta} L_{k,\text{rep}}^t(\theta) = 1$. Additionally, according to Assumption 1, the Hessian matrix mapping $\theta \mapsto \nabla_\theta^2 \mathcal{Q}(\theta)$ is continuous; combined with Assumption 2, the neighborhood $\mathcal{B}_\delta(\theta_k^{t,*}) \subset \Omega_\theta$ is a compact set. By applying the Weierstrass Extreme Value Theorem, there exists a finite supremum $H_\mathcal{Q} \triangleq \sup_{\theta \in \mathcal{B}_\delta} \|\nabla_\theta^2 \mathcal{Q}(\theta)\|_2 < \infty$. Thus, within the neighborhood of the saddle point, the spectral norm of the difference between the Hessian matrix of the composite objective $\tilde{L}(\theta)$ and the Hessian matrix at the saddle point is bounded:

$$\|\nabla_\theta^2 \tilde{L}(\theta) - \nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*})\|_2 \leq \delta L_\mathcal{Q} L_\text{lip} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q} < \mu, \forall \theta \in \mathcal{B}_\delta \tag{38}$$

Next, we prove the strong convexity of the composite objective $\tilde{L}(\theta)$ within the neighborhood of the saddle point. By decomposing the Hessian as $\nabla_\theta^2 \tilde{L}(\theta) = \nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*}) + (\nabla_\theta^2 \tilde{L}(\theta) - \nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*}))$, we define the upper bound function for the spectral norm of the error term as $\rho(\delta) \triangleq \sup_{\theta \in \mathcal{B}_\delta} \|\nabla_\theta^2 \tilde{L}(\theta) - \nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*})\|_2$. Since $\lambda_k^{t,*} = \mathcal{Q}(\theta_k^{t,*}) = 1$ and the relevant mappings are continuous, it follows that as $\theta \to \theta_k^{t,*}$, the terms $\mathcal{Q}(\theta) - \lambda_k^{t,*}$, $\nabla_\theta \mathcal{Q}(\theta)$, and $\nabla_\theta^2 \mathcal{Q}(\theta)$ all approach zero. Consequently, $\lim_{\delta \to 0^+} \rho(\delta) = 0$. Furthermore, based on the continuity of $\rho(\delta)$ and its limit at zero, for the strong convexity parameter $\mu > 0$ given in Assumption 11, we define the neighborhood radius:

$$\delta' = \rho^{-1}(\mu) \tag{39}$$

such that $0 \leq \rho(\delta) < \mu, \quad \forall \delta \in (0, \delta')$. According to Assumption 11, we have $\nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*}) = \nabla_\theta^2 L_{k,\text{total}}^t(\theta, 1) \succeq \mu \mathbf{I}, \quad \forall \theta \in \mathcal{B}_{\delta'}$, which is equivalent to its minimum eigenvalue satisfying $\lambda_{\min}(\nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*})) \geq \mu, \quad \forall \theta \in \mathcal{B}_{\delta'}$. Finally, by Weyl's perturbation theorem, we obtain:

$$\begin{aligned} \lambda_{\min}(\nabla_\theta^2 \tilde{L}(\theta)) \geq & \lambda_{\min}(\nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*})) - \|\nabla_\theta^2 \tilde{L}(\theta) - \nabla_\theta^2 L_{k,\text{total}}^t(\theta, \lambda_k^{t,*})\|_2 \\ \geq & \mu - \rho(\delta) > \mu - (\delta L_\mathcal{Q} L_\text{lip} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q}) \end{aligned} \tag{40}$$

Consequently, we have:

$$\nabla_\theta^2 \tilde{L}(\theta) \succeq (\mu - \rho(\delta))\mathbf{I} \succeq (\mu - (\delta L_\mathcal{Q} L_\text{lip} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q}))\mathbf{I} \succ \mathbf{0}, \quad \forall \theta \in \mathcal{B}_\delta \tag{41}$$

Consequently, the composite objective $\tilde{L}(\theta)$ is $(\mu - (\delta L_\mathcal{Q} L_\text{lip} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q}))$-strongly convex within the saddle point neighborhood $\mathcal{B}_\delta$. For any $\theta \in \mathcal{B}_\delta$, consider the parameterized line segment $\phi(\tau) = \theta_k^{t,*} + \tau(\theta - \theta_k^{t,*})$ for $\tau \in [0, 1]$. Since $\mathcal{B}_\delta$ is a convex set, it follows that $\phi(\tau) \in \mathcal{B}_\delta$. Therefore, the gradient difference $\nabla_\theta \tilde{L}(\theta) - \nabla_\theta \tilde{L}(\theta_k^{t,*})$ can be expressed as the integral of the Hessian matrix along this segment:

$$\nabla_\theta \tilde{L}(\theta) - \nabla_\theta \tilde{L}(\theta_k^{t,*}) = \int_0^1 \frac{d}{d\tau} \nabla_\theta \tilde{L}(\phi(\tau))\, d\tau = \int_0^1 \nabla_\theta^2 \tilde{L}(\phi(\tau))(\theta - \theta_k^{t,*})\, d\tau \tag{42}$$

Consequently, we have:

$$\begin{aligned} \langle \nabla_\theta \tilde{L}(\theta) - \nabla_\theta \tilde{L}(\theta_k^{t,*}), \theta - \theta_k^{t,*} \rangle &= \left\langle \int_0^1 \nabla_\theta^2 \tilde{L}(\phi(\tau))(\theta - \theta_k^{t,*})\, d\tau,\ \theta - \theta_k^{t,*} \right\rangle \\ &= \int_0^1 (\theta - \theta_k^{t,*})^\top \nabla_\theta^2 \tilde{L}(\phi(\tau))(\theta - \theta_k^{t,*})\, d\tau \\ &\leq (\mu - (\delta L_\mathcal{Q} L_\text{lip} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q})) \int_0^1 (\theta - \theta_k^{t,*})^\top (\theta - \theta_k^{t,*})\, d\tau \\ &= (\mu - (\delta L_\mathcal{Q} L_\text{lip} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q})) \|\theta - \theta_k^{t,*}\|_2^2 \end{aligned} \tag{43}$$

Next, we prove the local asymptotic stability of the optimization dynamics at the saddle point. Consider the following manipulation of the gradient of the composite objective function $\nabla_\theta \tilde{L}(\theta)$:

$$\begin{aligned}\nabla_\theta \tilde{L}(\theta) =&\nabla_\theta L^t_{k,\text{task}}(\theta) + \nabla_\theta \mathcal{Q}(\theta) \cdot L^t_{k,\text{rep}}(\theta) + \mathcal{Q}(\theta) \cdot \nabla_\theta L^t_{k,\text{rep}}(\theta) \\ =&[\nabla_\theta L^t_{k,\text{task}}(\theta) + \mathcal{Q}(\theta) \cdot \nabla_\theta L^t_{k,\text{rep}}(\theta)] + \nabla_\theta \mathcal{Q}(\theta) \cdot L^t_{k,\text{rep}}(\theta) \\ =& \left.\frac{\partial L^t_{k,\text{total}}(\theta, \lambda)}{\partial \theta}\right|_{\lambda=\mathcal{Q}(\theta)} + \nabla_\theta \mathcal{Q}(\theta) \cdot L^t_{k,\text{rep}}(\theta)\end{aligned} \tag{44}$$

Consequently, we obtain the expression for the gradient of the total loss function with respect to θ evaluated at $\lambda = \mathcal{Q}(\theta)$:

$$\nabla_\theta L^t_{k,\text{total}}(\theta, \mathcal{Q}(\theta)) \triangleq \left.\frac{\partial L^t_{k,\text{total}}(\theta, \lambda)}{\partial \theta}\right|_{\lambda=\mathcal{Q}(\theta)} = \nabla_\theta \tilde{L}(\theta) - \nabla_\theta \mathcal{Q}(\theta) \cdot L^t_{k,\text{rep}}(\theta) \tag{45}$$

Considering the $j+1$-th parameter update $\theta^{t,j+1}_k = \theta^{t,j}_k - \eta^{t,j}_k \nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))$, we have:

$$\begin{aligned}\|\theta^{t,j+1}_k - \theta^{t,*}_k\|_2^2 =&\|(\theta^{t,j}_k - \theta^{t,*}_k) - \nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2^2 \\ =&\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2 - 2\eta^{t,j}_k \langle \nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k)), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ &+ (\eta^{t,j}_k)^2 \|\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2^2\end{aligned} \tag{46}$$

For the inner product term $\langle \nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k)), \theta^{t,j}_k - \theta^{t,*}_k\rangle$, substituting $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k)) = \nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) L^t_{k,\text{rep}}(\theta^{t,j}_k)$ yields:

$$\begin{aligned}&\langle \nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k)), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ =&\langle \nabla_\theta \tilde{L}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle - \langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ =&\langle \nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle + \langle \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ &- \langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle\end{aligned} \tag{47}$$

Furthermore, utilizing the inequality $\langle \nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle \leq (\mu - (\delta L_{\mathcal{Q}} L_{\text{lip}} + 2L_{\mathcal{Q}} L_2 + H_{\mathcal{Q}}))\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2$, we obtain:

$$\begin{aligned}&\langle \nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k)), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ \leq&(\mu - (\delta L_{\mathcal{Q}} L_{\text{lip}} + 2L_{\mathcal{Q}} L_2 + H_{\mathcal{Q}}))\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2 + \langle \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ &- \langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle\end{aligned} \tag{48}$$

Consequently, we have:

$$\begin{aligned}&\|\theta^{t,j+1}_k - \theta^{t,*}_k\|_2^2 \\ \leq&(1 - 2\eta^{t,j}_k(\mu - (\delta L_{\mathcal{Q}} L_{\text{lip}} + 2L_{\mathcal{Q}} L_2 + H_{\mathcal{Q}})))\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2 - 2\eta^{t,j}_k \langle \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle \\ &+ 2\eta^{t,j}_k \langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle + (\eta^{t,j}_k)^2 \|\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2^2 \\ =&(1 - 2\eta^{t,j}_k(\mu - (\delta L_{\mathcal{Q}} L_{\text{lip}} + 2L_{\mathcal{Q}} L_2 + H_{\mathcal{Q}})))\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2 + (\eta^{t,j}_k)^2 \|\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2^2 \\ &+ 2\eta^{t,j}_k \langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle\end{aligned} \tag{49}$$

For the last term $\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k\rangle$, since the first-order necessary condition $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k, \mathcal{Q}(\theta^{t,*}_k)) = \mathbf{0}$ is satisfied at the saddle point, we have $\nabla_\theta \tilde{L}(\theta^{t,*}_k) =$

$\nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k, \mathcal{Q}(\theta^{t,*}_k)) + \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) L^t_{k,\text{rep}}(\theta^{t,*}_k) = \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) L^t_{k,\text{rep}}(\theta^{t,*}_k)$. Consequently, we obtain:

$$\begin{aligned}
&\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
=&\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
&+ \langle \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
=&\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle + \langle \mathbf{0}, \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
=&\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
=&\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
&+ \langle \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
=&\langle (\nabla_\theta \mathcal{Q}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k)) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
&+ \langle \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot (L^t_{k,\text{rep}}(\theta^{t,j}_k) - L^t_{k,\text{rep}}(\theta^{t,*}_k)), \theta^{t,j}_k - \theta^{t,*}_k \rangle
\end{aligned} \tag{50}$$

Furthermore, applying the norm inequality to the above inner product and utilizing $\|\nabla_\theta \mathcal{Q}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k)\|_2 \leq H_\mathcal{Q} \|\theta^{t,j}_k - \theta^{t,*}_k\|_2$ and $\|\nabla_\theta \mathcal{Q}(\theta^{t,*}_k)\|_2 \leq L_\mathcal{Q}$, along with the L_2-Lipschitz continuity which yields $\|L^t_{k,\text{rep}}(\theta^{t,j}_k) - L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2 \leq L_2 \|\theta^{t,j}_k - \theta^{t,*}_k\|_2$, we obtain:

$$\begin{aligned}
&\langle \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k), \theta^{t,j}_k - \theta^{t,*}_k \rangle \\
\leq&\|\nabla_\theta \mathcal{Q}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k)\|_2 \|L^t_{k,\text{rep}}(\theta^{t,j}_k)\|_2 \|\theta^{t,j}_k - \theta^{t,*}_k\|_2 \\
&+ \|\nabla_\theta \mathcal{Q}(\theta^{t,*}_k)\|_2 \|L^t_{k,\text{rep}}(\theta^{t,j}_k) - L^t_{k,\text{rep}}(\theta^{t,*}_k)\|_2 \|\theta^{t,j}_k - \theta^{t,*}_k\|_2 \\
\leq&(H_\mathcal{Q} \cdot \sup_{\theta \in \mathcal{B}_\delta} (L^t_{k,\text{rep}}(\theta)) + L_\mathcal{Q} L_2) \|\theta^{t,j}_k - \theta^{t,*}_k\|_2
\end{aligned} \tag{51}$$

Substituting this back into the original inequality, we obtain:

$$\begin{aligned}
&\|\theta^{t,j+1}_k - \theta^{t,*}_k\|_2^2 \\
\leq&(1 - 2\eta^{t,j}_k(\mu - (\delta L_\mathcal{Q} L_{\text{lip}} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q})))\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2 + (\eta^{t,j}_k)^2 \|\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2^2 \\
&+ 2\eta^{t,j}_k (H_\mathcal{Q} \cdot \sup_{\theta \in \mathcal{B}_\delta} (L^t_{k,\text{rep}}(\theta)) + L_\mathcal{Q} L_2) \|\theta^{t,j}_k - \theta^{t,*}_k\|_2 \\
=&(1 - 2\eta^{t,j}_k(\mu - (\delta L_\mathcal{Q} L_{\text{lip}} + 2L_\mathcal{Q} L_2 + H_\mathcal{Q}) - (H_\mathcal{Q} \cdot \sup_{\theta \in \mathcal{B}_\delta} (L^t_{k,\text{rep}}(\theta)) + L_\mathcal{Q} L_2)))\|\theta^{t,j}_k - \theta^{t,*}_k\|_2^2 \\
&+ (\eta^{t,j}_k)^2 \|\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2^2
\end{aligned} \tag{52}$$

Next, considering the term $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))$ and utilizing the first-order optimality condition at the saddle point $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k) = \mathbf{0}$, we obtain:

$$\begin{aligned}
&\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k)) \\
=&\nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k) \\
=&[\nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k)] + [\nabla_\theta \tilde{L}(\theta^{t,*}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k)] \\
&+ [\nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k)] \\
=&[\nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k)] + \nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k) \\
&+ [\nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k)] \\
=&[\nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k)] + [\nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k)]
\end{aligned} \tag{53}$$

Taking the norm of the above expression, we obtain:

$$\begin{aligned}
\|\nabla_\theta L^t_{k,\text{total}}(\theta^{t,j}_k, \mathcal{Q}(\theta^{t,j}_k))\|_2 =&\|\nabla_\theta \tilde{L}(\theta^{t,j}_k) - \nabla_\theta \tilde{L}(\theta^{t,*}_k)\|_2 \\
&+ \|\nabla_\theta \mathcal{Q}(\theta^{t,*}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,*}_k) - \nabla_\theta \mathcal{Q}(\theta^{t,j}_k) \cdot L^t_{k,\text{rep}}(\theta^{t,j}_k)\|_2
\end{aligned} \tag{54}$$

For the first term $\|\nabla_\theta \tilde{L}(\theta_k^{t,j}) - \nabla_\theta \tilde{L}(\theta_k^{t,*})\|_2$, applying the integral form of the mean value theorem yields:

$$\begin{aligned}
&\|\nabla_\theta \tilde{L}(\theta_k^{t,j}) - \nabla_\theta \tilde{L}(\theta_k^{t,*})\|_2 \\
\leq &\sup_{\xi \in \mathcal{B}_\delta} \|\nabla_\theta^2 \tilde{L}(\xi)\|_2 \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 \\
\leq &\sup_{\xi \in \mathcal{B}_\delta} (\|\nabla_\theta^2 L_{k,\text{total}}^t(\xi, \lambda_k^{t,*})\|_2 + \|\nabla_\theta^2 \tilde{L}(\xi) - \nabla_\theta^2 L_{k,\text{total}}^t(\xi, \lambda_k^{t,*})\|_2) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 \\
\leq &\sup_{\xi \in \mathcal{B}_\delta} (\|\nabla_\theta^2 L_{k,\text{task}}^t(\xi) + \mathcal{Q} \cdot (\theta_k^{t,*}) \nabla_\theta^2 L_{k,\text{rep}}^t(\xi)\|_2) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 \\
&+ \sup_{\xi \in \mathcal{B}_\delta} (\|\nabla_\theta^2 \tilde{L}(\xi) - \nabla_\theta^2 L_{k,\text{total}}^t(\xi, \lambda_k^{t,*})\|_2) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 \\
\leq &(L + \sup_{\xi \in \mathcal{B}_\delta} (\mathcal{Q}(\xi)) \cdot L) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + (\delta L_\mathcal{Q} L_{\text{lip}} + 2 L_\mathcal{Q} L_2 + H_\mathcal{Q}) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 \\
= &((1 + \sup_{\xi \in \mathcal{B}_\delta} (\mathcal{Q}(\xi)) + \delta L_\mathcal{Q}) L + 2 L_\mathcal{Q} L_2 + H_\mathcal{Q}) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2
\end{aligned} \tag{55}$$

For the second term $\|\nabla_\theta \mathcal{Q}(\theta_k^{t,*}) L_{k,\text{rep}}^t(\theta_k^{t,*}) - \nabla_\theta \mathcal{Q}(\theta_k^{t,j}) L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2$, we have:

$$\begin{aligned}
&\|\nabla_\theta \mathcal{Q}(\theta_k^{t,*}) \cdot L_{k,\text{rep}}^t(\theta_k^{t,*}) - \nabla_\theta \mathcal{Q}(\theta_k^{t,j}) \cdot L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2 \\
\leq &\|\nabla_\theta \mathcal{Q}(\theta_k^{t,*}) \cdot L_{k,\text{rep}}^t(\theta_k^{t,*}) - \nabla_\theta \mathcal{Q}(\theta_k^{t,*}) \cdot L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2 \\
&+ \|\nabla_\theta \mathcal{Q}(\theta_k^{t,*}) \cdot L_{k,\text{rep}}^t(\theta_k^{t,j}) - \nabla_\theta \mathcal{Q}(\theta_k^{t,j}) \cdot L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2 \\
\leq &\|\nabla_\theta \mathcal{Q}(\theta_k^{t,*})\|_2 \|L_{k,\text{rep}}^t(\theta_k^{t,*}) - L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2 + \|\nabla_\theta \mathcal{Q}(\theta_k^{t,*}) - \nabla_\theta \mathcal{Q}(\theta_k^{t,j})\|_2 \|L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2 \\
\leq &[L_2 L_\mathcal{Q} + H_\mathcal{Q} \cdot \sup_{\theta \in \mathcal{B}_\delta} (L_{k,\text{rep}}^t(\theta))] \|\theta_k^{t,j} - \theta_k^{t,*}\|_2
\end{aligned} \tag{56}$$

Consequently, we have:

$$\begin{aligned}
&\|\nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j}, \mathcal{Q}(\theta_k^{t,j}))\|_2 \\
\leq &((1 + \sup_{\xi \in \mathcal{B}_\delta} (\mathcal{Q}(\xi)) + \delta L_\mathcal{Q}) L + 3 L_\mathcal{Q} L_2 + H_\mathcal{Q}(1 + \sup_{\theta \in \mathcal{B}_\delta} (L_{k,\text{rep}}^t(\theta)))) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2
\end{aligned} \tag{57}$$

Substituting this back into the original expression and applying the bound $\sup_{\theta \in \mathcal{B}_\delta} L_{k,\text{rep}}^t(\theta) \leq 1$ given by Assumption 5, we obtain:

$$\begin{aligned}
&\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2 \\
\leq &(1 - 2\eta_k^{t,j}(\mu - (\delta L_\mathcal{Q} L_{\text{lip}} + 2 L_\mathcal{Q} L_2 + H_\mathcal{Q}) - (H_\mathcal{Q} \cdot \sup_{\theta \in \mathcal{B}_\delta} (L_{k,\text{rep}}^t(\theta)) + L_\mathcal{Q} L_2))) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 \\
&+ (\eta_k^{t,j})^2 ((1 + \sup_{\xi \in \mathcal{B}_\delta} (\mathcal{Q}(\xi)) + \delta L_\mathcal{Q}) L + 3 L_\mathcal{Q} L_2 + H_\mathcal{Q}(1 + \sup_{\theta \in \mathcal{B}_\delta} (L_{k,\text{rep}}^t(\theta))))^2 \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 \\
\leq &(1 - 2\eta_k^{t,j}(\mu - (\delta L_\mathcal{Q} L_{\text{lip}} + 3 L_\mathcal{Q} L_2 + 2 H_\mathcal{Q}))) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 \\
&+ (\eta_k^{t,j})^2 ((1 + \sup_{\xi \in \mathcal{B}_\delta} (\mathcal{Q}(\xi)) + \delta L_\mathcal{Q}) L + 3 L_\mathcal{Q} L_2 + 2 H_\mathcal{Q})^2 \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2
\end{aligned} \tag{58}$$

In the preceding proof, we have established that the composite objective $\tilde{L}(\theta) = L_{k,\text{task}}^t(\theta) + \mathcal{Q}(\theta) L_{k,\text{rep}}^t(\theta)$ is $(\mu - (\delta L_\mathcal{Q} L_{\text{lip}} + 2 L_\mathcal{Q} L_2 + H_\mathcal{Q}))$-strongly convex within the saddle point neighborhood $\mathcal{B}_\delta$. To simplify the notation, we define:

$$\begin{aligned}
\mu_1(\delta) = &\mu - (\delta L_\mathcal{Q} L_{\text{lip}} + 2 L_\mathcal{Q} L_2 + H_\mathcal{Q}) \\
L_{\text{smooth}}(\delta) = &(1 + \sup_{\xi \in \mathcal{B}_\delta} (\mathcal{Q}(\xi)) + \delta L_\mathcal{Q}) L + 3 L_\mathcal{Q} L_2 + 2 H_\mathcal{Q}
\end{aligned} \tag{59}$$

Consequently, the above inequality simplifies to $\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2 \leq (1 - 2\eta_k^{t,j}(\mu_1(\delta) - (L_\mathcal{Q} L_2 + H_\mathcal{Q})) + (\eta_k^{t,j})^2 (L_{\text{smooth}}(\delta))^2) \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2$. To ensure that it satisfies the single-step Lyapunov

linear contraction form $\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2 \leq (1 - \eta_k^{t,j}\mu_1(\delta))\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2$, a constraint on the learning rate is required. This yields:

$$-2\eta_k^{t,j} \cdot (\mu_1(\delta) - (L_{\mathcal{Q}}L_2 + H_{\mathcal{Q}})) + (\eta_k^{t,j})^2(L_{\text{smooth}}(\delta))^2 \leq -\eta_k^{t,j} \cdot \mu_1(\delta) \tag{60}$$

Consequently, we have:

$$2\eta_k^{t,j}(L_{\mathcal{Q}}L_2 + H_{\mathcal{Q}}) + (\eta_k^{t,j})^2(L_{\text{smooth}}(\delta))^2 \leq \eta_k^{t,j} \cdot \mu_1(\delta) \tag{61}$$

By allocating the target contraction equally to the first- and second-order terms, i.e., bounding each by $\eta_k^{t,j}\mu_1(\delta)/2$, we obtain:

$$\begin{cases} (\eta_k^{t,j})^2(L_{\text{smooth}}(\delta))^2 \leq \dfrac{\eta_k^{t,j}\mu_1(\delta)}{2} \\ 2\eta_k^{t,j}(L_{\mathcal{Q}}L_2 + H_{\mathcal{Q}}) \leq \dfrac{\eta_k^{t,j}\mu_1(\delta)}{2} \end{cases} \tag{62}$$

To satisfy the first inequality, it is required that $\eta_k^{t,j} \leq \frac{\mu_1(\delta)}{2(L_{\text{smooth}}(\delta))^2}$. For the second inequality, the condition $(L_{\mathcal{Q}}L_2 + H_{\mathcal{Q}}) \leq \frac{\mu_1(\delta)}{4}$ must hold, which leads to the conservative constraint on the learning rate: $\eta_k^{t,j} \leq \frac{\mu_1(\delta)}{4(L_{\mathcal{Q}}L_2 + H_{\mathcal{Q}})}$. Combining these conditions, the upper bound for the learning rate is obtained as:

$$\eta_k^{t,j} \leq \min\left(\frac{\mu_1(\delta)}{2(L_{\text{smooth}}(\delta))^2}, \frac{\mu_1(\delta)}{4(L_{\mathcal{Q}}L_2 + H_{\mathcal{Q}})}\right) \tag{63}$$

According to Assumption 3, the sum of squared learning rates converges, i.e., $\sum_{j=0}^{\infty}(\eta_k^{t,j})^2 < \infty$. This implies that $\lim_{j\to\infty} \eta_k^{t,j} = 0$, and thus there must exist $J_0 > 0$ such that the learning rate satisfies the aforementioned upper bound condition for all $j > J_0$. Consequently, within the saddle point neighborhood $\mathcal{B}_\delta$, the following contraction inequality holds:

$$\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2 \leq (1 - \eta_k^{t,j} \cdot \mu_1(\delta))\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 \tag{64}$$

Furthermore, for any $j \geq J_0$, applying the inequality $1 - \eta_k^{t,j}\mu_1(\delta) \leq \exp(-\eta_k^{t,j}\mu_1(\delta))$, we obtain:

$$\begin{aligned} \|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2 \leq & \|\theta_k^{t,J_0} - \theta_k^{t,*}\|_2^2 \prod_{m=J_0}^{j}(1 - \eta_k^{t,m} \cdot \mu_1(\delta)) \\ \leq & \|\theta_k^{t,J_0} - \theta_k^{t,*}\|_2^2 \cdot \exp(-\mu_1(\delta) \cdot \sum_{m=J_0}^{j} \eta_k^{t,m}) \end{aligned} \tag{65}$$

Since the initial error $\|\theta_k^{t,J_0} - \theta_k^{t,*}\|_2^2$ is a finite value, and according to Assumption 3 the series $\sum_{m=J_0}^{\infty} \eta_k^{t,m} = \infty$ diverges, the exponential decay term approaches zero as $j \to \infty$. Therefore, we have:

$$0 \leq \lim_{j\to\infty} \|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2 \leq \|\theta_k^{t,J_0} - \theta_k^{t,*}\|_2^2 \lim_{j\to\infty} \exp(-\mu_1(\delta) \cdot \sum_{m=J_0}^{j} \eta_k^{t,m}) = 0 \tag{66}$$

Consequently, we obtain $\lim_{j\to\infty} \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 = 0$, which implies:

$$\lim_{j\to\infty} \theta_k^{t,j} = \theta_k^{t,*} \tag{67}$$

By the continuity of $\mathcal{Q}(\theta)$, substituting the limit into the quotient mapping yields:

$$\lim_{j\to\infty} \mathcal{Q}(\theta_k^{t,j}) = \mathcal{Q}(\lim_{j\to\infty} \theta_k^{t,j}) = \mathcal{Q}(\theta_k^{t,*}) = \lambda_k^{t,*} \tag{68}$$

Consequently, within the δ-neighborhood of the saddle point $(\theta_k^{t,*}, \lambda_k^{t,*})$, the sequence of iterates asymptotically converges to this saddle point. This completes the proof of Lemma 1.

B.3 PROOF OF THEOREM 1

Theorem 1 (Local Saddle-Point Convergence of GRRS) *Under standard assumptions (see Appendix B.1), let $\theta_k^{t,0}$ be the initial model for client k in FL round t. After J local epochs via the Gradient-Responsive Replay Scheme, $\theta_k^{t,J}$ converges to the ideal saddle point $(\theta_k^{t,*}, \lambda_k^{t,*})$ within a truncation error bound:*

$$\mathbb{E}[L_{k,total}^t(\theta_k^{t,J}|\lambda_k^{t,J})] \leq L_{k,total}^t(\theta_k^{t,*}|\lambda_k^{t,*}) + \mathcal{E}_{k,bound}^t(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2) \tag{69}$$

where the ideal saddle point $(\theta_k^{t,*}, \lambda_k^{t,*})$ satisfies the minimax equilibrium property for any $\theta \in \Omega_\theta$ and $\lambda \in \Omega_\lambda$:

$$L_{k,total}^t(\theta_k^{t,*}|\lambda) \leq L_{k,total}^t(\theta_k^{t,*}|\lambda_k^{t,*}) \leq L_{k,total}^t(\theta|\lambda_k^{t,*}) \tag{70}$$

The truncation error upper bound $\mathcal{E}_{k,bound}^t\left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2\right)$ is a deterministic function depending on the local step-size decay sequence and the initial parameter residual, defined as:

$$\begin{aligned}
&\mathcal{E}_{k,bound}^t\left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2\right) \\
=&\frac{L}{2}\exp\left(-(\mu - 2L_2L_{\mathcal{Q}})\sum_{j=0}^{J-1}\eta_k^{t,j}\right)\mathbb{E}_\xi[\|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2] \\
&+\frac{L}{2}\sum_{j=0}^{J-1}\left(\eta_k^{t,j}\frac{{L_2}^2{\epsilon_{approx}}^2}{\mu} + (\eta_k^{t,j})^2G^2\right)\prod_{i=j+1}^{J-1}(1 - \eta_k^{t,i}(\mu - 2L_2L_{\mathcal{Q}})) \\
&+(L_{\mathcal{Q}}D_\theta + \epsilon_{approx})\cdot\sup_{\theta\in\Omega_\theta}L_{k,rep}^t(\theta)
\end{aligned} \tag{71}$$

where $\epsilon_{approx} \triangleq \sup_{\theta\in\Omega_\theta}\left|\frac{\|\nabla_h L_{k,rep}^t(\theta)\|_2^2}{\|\nabla_h L_{k,task}^t(\theta)\|_2^2} - \frac{\|\nabla_\theta L_{k,rep}^t(\theta)\|_2^2}{\|\nabla_\theta L_{k,task}^t(\theta)\|_2^2}\right|$ denotes the supremum of the surrogate error for the global replay loss weight, and $L_{\mathcal{Q}}$ is the Lipschitz constant of the gradient quotient mapping over the model parameters $\mathcal{Q}(\theta) = \|\nabla_\theta L_{k,task}^t(\theta)\|_2^2/\|\nabla_\theta L_{k,rep}^t(\theta)\|_2^2$, as specifically defined in the proof of Lemma 1.

Proof:

The existence and asymptotic stability of the ideal saddle point $(\theta_k^{t,*}, \lambda_k^{t,*})$ have been established in Lemma 1, and thus the proof is omitted here.

Consider the replay loss weight error term $\frac{\|\nabla_h L_{k,\text{rep}}^t(\theta)\|_2^2}{\|\nabla_h L_{k,\text{task}}^t(\theta)\|_2^2} - \frac{\|\nabla_\theta L_{k,\text{rep}}^t(\theta)\|_2^2}{\|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2^2}$. To simplify the notation, let $f_1(\theta) = \|\nabla_h L_{k,\text{rep}}^t(\theta)\|_2^2$, $g_1(\theta) = \|\nabla_h L_{k,\text{task}}^t(\theta)\|_2^2$, $f_2(\theta) = \|\nabla_\theta L_{k,\text{rep}}^t(\theta)\|_2^2$, and $g_2(\theta) = \|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2^2$. Then, we have:

$$\begin{aligned}
\left|\frac{\|\nabla_h L_{k,\text{rep}}^t(\theta)\|_2^2}{\|\nabla_h L_{k,\text{task}}^t(\theta)\|_2^2} - \frac{\|\nabla_\theta L_{k,\text{rep}}^t(\theta)\|_2^2}{\|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2^2}\right| &= \left|\frac{f_1(\theta)}{g_1(\theta)} - \frac{f_2(\theta)}{g_2(\theta)}\right| = \frac{|f_1(\theta)g_2(\theta) - f_2(\theta)g_1(\theta)|}{g_1(\theta)g_2(\theta)} \\
&= \frac{|f_1(\theta)g_2(\theta) - f_1(\theta)g_1(\theta) + f_1(\theta)g_1(\theta) - f_2(\theta)g_1(\theta)|}{g_1(\theta)g_2(\theta)} \\
&= \frac{|f_1(\theta)(g_2(\theta) - g_1(\theta)) + g_1(\theta)(f_1(\theta) - f_2(\theta))|}{g_1(\theta)g_2(\theta)} \\
&\leq \frac{|f_1(\theta)||g_1(\theta) - g_2(\theta)| + |g_1(\theta)||f_1(\theta) - f_2(\theta)|}{g_1(\theta)g_2(\theta)}
\end{aligned} \tag{72}$$

Under Assumption 1, we have $\|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2 \leq L_2$. According to Proposition 1 in Ash et al. (2020), it holds that $\|\nabla_h L_{k,\text{task}}^t(\theta)\|_2 \leq \|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2$ (i.e., $g_1(\theta) \leq f_1(\theta)$). Furthermore, under Assumptions 1 and 2, the mapping $\theta \mapsto \min(\|\nabla_h L_{k,\text{task}}^t(\theta)\|_2, \|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2)$ is a continuous positive function on Ω_θ. Therefore, we define:

$$\epsilon_{\min} \triangleq \inf_{\theta\in\Omega_\theta}\min(\|\nabla_h L_{k,\text{task}}^t(\theta)\|_2, \|\nabla_\theta L_{k,\text{task}}^t(\theta)\|_2) > 0 \tag{73}$$

Substituting these into the original expression yields:

$$\begin{aligned}\left|\frac{\|\nabla_h L^t_{k,\text{rep}}(\theta)\|_2^2}{\|\nabla_h L^t_{k,\text{task}}(\theta)\|_2^2}-\frac{\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2}{\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2}\right| \leq & \frac{|f_1(\theta)||g_1(\theta)-g_2(\theta)|+|g_1(\theta)||f_1(\theta)-f_2(\theta)|}{g_1(\theta)g_2(\theta)} \\ \leq & \frac{f_1(\theta)}{g_1(\theta)g_2(\theta)}(|g_1(\theta)-g_2(\theta)|+|f_1(\theta)-f_2(\theta)|) \\ \leq & \frac{{L_2}^2}{{\epsilon_{\min}}^4}(|g_1(\theta)-g_2(\theta)|+|f_1(\theta)-f_2(\theta)|) \\ = & \frac{{L_2}^2}{{\epsilon_{\min}}^4}(|\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2-\|\nabla_h L^t_{k,\text{task}}(\theta)\|_2^2|) \\ & +\frac{{L_2}^2}{{\epsilon_{\min}}^4}(|\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2-\|\nabla_h L^t_{k,\text{rep}}(\theta)\|_2^2|)\end{aligned} \tag{74}$$

Thus, the surrogate error is bounded. Furthermore, we define the supremum of the global replay loss weight surrogate error as ϵ_{approx}:

$$\epsilon_{\text{approx}} \triangleq \sup_{\theta\in\Omega_\theta}\left|\frac{\|\nabla_h L^t_{k,\text{rep}}(\theta)\|_2^2}{\|\nabla_h L^t_{k,\text{task}}(\theta)\|_2^2}-\frac{\|\nabla_\theta L^t_{k,\text{rep}}(\theta)\|_2^2}{\|\nabla_\theta L^t_{k,\text{task}}(\theta)\|_2^2}\right| \tag{75}$$

Consider the $(j+1)$-th model iteration. The model parameters before the update are denoted as $\theta_k^{t,j}$, and the replay loss weight is given by $\lambda_k^{t,j+1}=\|\nabla_h L^t_{k,\text{rep}}(\theta_k^{t,j})\|_2^2/\|\nabla_h L^t_{k,\text{task}}(\theta_k^{t,j})\|_2^2$. The gradient quotient mapping for the replay loss weight with respect to the model parameters is defined as $\mathcal{Q}(\theta_k^{t,j})=\|\nabla_\theta L^t_{k,\text{rep}}(\theta_k^{t,j})\|_2^2/\|\nabla_\theta L^t_{k,\text{task}}(\theta_k^{t,j})\|_2^2$, which has been shown to be $L_{\mathcal{Q}}$-Lipschitz continuous in Lemma 1. By the definition of ϵ_{approx}, it is evident that $|\lambda_k^{t,j+1}-\mathcal{Q}(\theta_k^{t,j})|\leq\epsilon_{\text{approx}}$. Consequently, regarding the ideal saddle point $(\theta_k^{t,*},\lambda_k^{t,*})$, we have:

$$\begin{aligned}|\lambda_k^{t,j+1}-\lambda_k^{t,*}| = & |\lambda_k^{t,j+1}-\mathcal{Q}(\theta_k^{t,j})+\mathcal{Q}(\theta_k^{t,j})-\lambda_k^{t,*}| \\ \leq & |\lambda_k^{t,j+1}-\mathcal{Q}(\theta_k^{t,j})|+|\mathcal{Q}(\theta_k^{t,j})-\mathcal{Q}(\theta_k^{t,*})| \\ \leq & \epsilon_{\text{approx}}+L_{\mathcal{Q}}\|\theta_k^{t,j}-\theta_k^{t,*}\|_2\end{aligned} \tag{76}$$

Consider a random mini-batch ξ, with the corresponding stochastic gradient given by $\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})$. Expanding the $(j+1)$-th model update yields:

$$\begin{aligned}&\|\theta_k^{t,j+1}-\theta_k^{t,*}\|_2^2 \\ =&\|\theta_k^{t,j}-\eta_k^{t,j}\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})-\theta_k^{t,*}\|_2^2 \\ =&\|\theta_k^{t,j}-\theta_k^{t,*}\|_2^2-2\langle\theta_k^{t,j}-\theta_k^{t,*},\eta_k^{t,j}\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})\rangle+\|\eta_k^{t,j}\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})\|_2^2 \\ =&\|\theta_k^{t,j}-\theta_k^{t,*}\|_2^2-2\eta_k^{t,j}\langle\theta_k^{t,j}-\theta_k^{t,*},\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})\rangle+(\eta_k^{t,j})^2\|\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})\|_2^2\end{aligned} \tag{77}$$

Taking the expectation of the above expression with respect to the random mini-batch ξ. By the unbiasedness of stochastic sampling, we have $\mathbb{E}_\xi[\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})]=\nabla_\theta L^t_{k,\text{total}}(\theta_k^{t,j}|\lambda_k^{t,j})$; furthermore, under Assumption 13, it holds that $\mathbb{E}_\xi[\|\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})\|_2^2]\leq G^2$. Combining these results yields:

$$\begin{aligned}\mathbb{E}_\xi[\|\theta_k^{t,j+1}-\theta_k^{t,*}\|_2^2] = & \|\theta_k^{t,j}-\theta_k^{t,*}\|_2^2-2\eta_k^{t,j}\langle\theta_k^{t,j}-\theta_k^{t,*},\mathbb{E}_\xi[\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})]\rangle \\ & +(\eta_k^{t,j})^2\mathbb{E}_\xi[\|\hat{\nabla}_\theta L^t_{k,\text{total}}(\theta_k^{t,j},\xi|\lambda_k^{t,j})\|_2^2] \\ \leq & \|\theta_k^{t,j}-\theta_k^{t,*}\|_2^2-2\eta_k^{t,j}\langle\theta_k^{t,j}-\theta_k^{t,*},\nabla_\theta L^t_{k,\text{total}}(\theta_k^{t,j}|\lambda_k^{t,j})\rangle+(\eta_k^{t,j})^2G^2\end{aligned} \tag{78}$$

Consider the inner product term $\langle \theta_k^{t,j} - \theta_k^{t,*}, \nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j}|\lambda_k^{t,j})\rangle$. By definition, $L_{k,\text{total}}^t(\theta_k^{t,j}|\lambda_k^{t,j}) = L_{k,\text{task}}^t(\theta_k^{t,j}) + \lambda_k^{t,j} L_{k,\text{rep}}^t(\theta_k^{t,j})$. Since $\lambda_k^{t,j}$ is treated as a constant with respect to θ, we have:

$$\begin{aligned}
\nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j}|\lambda_k^{t,j}) =& \nabla_\theta L_{k,\text{task}}^t(\theta_k^{t,j}) + \lambda_k^{t,j} \cdot \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j}) \\
=& \nabla_\theta L_{k,\text{task}}^t(\theta_k^{t,j}) + (\lambda_k^{t,*} + (\lambda_k^{t,j} - \lambda_k^{t,*})) \cdot \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j}) \\
=& \nabla_\theta L_{k,\text{task}}^t(\theta_k^{t,j}) + \lambda_k^{t,*} \cdot \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j}) + (\lambda_k^{t,j} - \lambda_k^{t,*}) \cdot \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j}) \\
=& \nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j}|\lambda_k^{t,*}) + (\lambda_k^{t,j} - \lambda_k^{t,*}) \cdot \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j})
\end{aligned} \tag{79}$$

Substituting the above results into the previous expression. Under Assumption 11, we have $\langle \nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j}|\lambda_k^{t,*}), \theta_k^{t,j} - \theta_k^{t,*}\rangle \geq \mu\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2$; under Assumption 1, it holds that $\|\nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2 \leq L_2$. Applying the Cauchy-Schwarz inequality yields:

$$\begin{aligned}
&\mathbb{E}_\xi[\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2] \\
\leq& \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 - 2\eta_k^{t,j}\langle \nabla_\theta L_{k,\text{total}}^t(\theta_k^{t,j}|\lambda_k^{t,*}), \theta_k^{t,j} - \theta_k^{t,*}\rangle \\
&- 2\eta_k^{t,j}(\lambda_k^{t,j} - \lambda_k^{t,*})\langle \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j}), \theta_k^{t,j} - \theta_k^{t,*}\rangle + (\eta_k^{t,j})^2 G^2 \\
\leq& \|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 - 2\eta_k^{t,j}\mu\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 \\
&+ 2\eta_k^{t,j}|\lambda_k^{t,j} - \lambda_k^{t,*}|\langle \nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j}), \theta_k^{t,j} - \theta_k^{t,*}\rangle + (\eta_k^{t,j})^2 G^2 \\
\leq& (1 - 2\eta_k^{t,j}\mu)\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + 2\eta_k^{t,j}|\lambda_k^{t,j} - \lambda_k^{t,*}| \cdot \|\nabla_\theta L_{k,\text{rep}}^t(\theta_k^{t,j})\|_2\|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + (\eta_k^{t,j})^2 G^2 \\
\leq& (1 - 2\eta_k^{t,j}\mu)\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + 2\eta_k^{t,j} L_2|\lambda_k^{t,j} - \lambda_k^{t,*}| \cdot \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + (\eta_k^{t,j})^2 G^2
\end{aligned} \tag{80}$$

Furthermore, substituting $|\lambda_k^{t,j} - \lambda_k^{t,*}| \leq L_\mathcal{Q}\|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + \epsilon_{\text{approx}}$ yields:

$$\begin{aligned}
&\mathbb{E}_\xi[\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2] \\
\leq& (1 - 2\eta_k^{t,j}\mu)\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + 2\eta_k^{t,j} L_2|\lambda_k^{t,j} - \lambda_k^{t,*}| \cdot \|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + (\eta_k^{t,j})^2 G^2 \\
\leq& (1 - 2\eta_k^{t,j}\mu)\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + 2\eta_k^{t,j} L_2(L_\mathcal{Q}\|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + \epsilon_{\text{approx}})\|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + (\eta_k^{t,j})^2 G^2 \\
=& (1 - 2\eta_k^{t,j}\mu + 2\eta_k^{t,j} L_2 L_\mathcal{Q})\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + 2\eta_k^{t,j} L_2\epsilon_{\text{approx}}\|\theta_k^{t,j} - \theta_k^{t,*}\|_2 + (\eta_k^{t,j})^2 G^2
\end{aligned} \tag{81}$$

For the second term, by Young's inequality, we have $2L_2\epsilon_{\text{approx}}\|\theta_k^{t,j} - \theta_k^{t,*}\|_2 \leq \mu\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + \frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu}$. Substituting this into the above expression yields:

$$\begin{aligned}
\mathbb{E}_\xi[\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2] \leq& (1 - 2\eta_k^{t,j}\mu + 2\eta_k^{t,j} L_2 L_\mathcal{Q})\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + \eta_k^{t,j}\mu\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 \\
&+ \eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2 G^2 \\
=& (1 - \eta_k^{t,j}\mu + 2\eta_k^{t,j} L_2 L_\mathcal{Q})\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + \eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2 G^2 \\
=& (1 - \eta_k^{t,j}(\mu - 2L_2 L_\mathcal{Q}))\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2 + \eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2 G^2 \\
=& (1 - \eta_k^{t,j}(\mu - 2L_2 L_\mathcal{Q}))\mathbb{E}_\xi[\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2] + \eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2 G^2
\end{aligned} \tag{82}$$

Taking the expectation of both sides with respect to ξ yields:

$$\mathbb{E}_\xi[\|\theta_k^{t,j+1} - \theta_k^{t,*}\|_2^2] \leq (1 - \eta_k^{t,j}(\mu - 2L_2 L_\mathcal{Q}))\mathbb{E}_\xi[\|\theta_k^{t,j} - \theta_k^{t,*}\|_2^2] + \eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2 G^2 \tag{83}$$

Next, considering the entire J-step iteration process, we obtain:

$$
\begin{aligned}
&\mathbb{E}_{\xi}[\|\theta_k^{t,J} - \theta_k^{t,*}\|_2^2] \\
\leq &(1 - \eta_k^{t,J-1}(\mu - 2L_2L_{\mathcal{Q}}))\mathbb{E}_{\xi}[\|\theta_k^{t,J-1} - \theta_k^{t,*}\|_2^2] + \eta_k^{t,J-1}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,J-1})^2G^2 \\
\leq &(1 - \eta_k^{t,J-1}(\mu - 2L_2L_{\mathcal{Q}}))(1 - \eta_k^{t,J-2}(\mu - 2L_2L_{\mathcal{Q}}))\mathbb{E}_{\xi}[\|\theta_k^{t,J-2} - \theta_k^{t,*}\|_2^2] \\
&+ (1 - \eta_k^{t,J-1}(\mu - 2L_2L_{\mathcal{Q}}))\left(\eta_k^{t,J-2}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,J-2})^2G^2\right) \\
&+ \left(\eta_k^{t,J-1}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,J-1})^2G^2\right) \\
= &\left[\prod_{j=J-2}^{J-1}\left(1 - \eta_k^{t,j}(\mu - 2L_2L_{\mathcal{Q}})\right)\right]\mathbb{E}_{\xi}[\|\theta_k^{t,J-2} - \theta_k^{t,*}\|_2^2] \\
&+ \sum_{j=J-2}^{J-1}\left(\eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2G^2\right)\prod_{i=j+1}^{J-1}(1 - \eta_k^{t,i}(\mu - 2L_2L_{\mathcal{Q}})) \\
\leq &\dots \\
\leq &\left[\prod_{j=0}^{J-1}\left(1 - \eta_k^{t,j}(\mu - 2L_2L_{\mathcal{Q}})\right)\right]\mathbb{E}_{\xi}[\|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2] \\
&+ \sum_{j=0}^{J-1}\left(\eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2G^2\right)\prod_{i=j+1}^{J-1}(1 - \eta_k^{t,i}(\mu - 2L_2L_{\mathcal{Q}}))
\end{aligned}
\tag{84}
$$

Applying the inequality $1 - x \leq \exp(-x)$ to the product term in the above expression yields:

$$
\begin{aligned}
\mathbb{E}_{\xi}[\|\theta_k^{t,J} - \theta_k^{t,*}\|_2^2] \leq &\left[\prod_{j=0}^{J-1}\left(1 - \eta_k^{t,j}(\mu - 2L_2L_{\mathcal{Q}})\right)\right]\mathbb{E}_{\xi}[\|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2] \\
&+ \sum_{j=0}^{J-1}\left(\eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2G^2\right)\prod_{i=j+1}^{J-1}(1 - \eta_k^{t,i}(\mu - 2L_2L_{\mathcal{Q}})) \\
\leq &\exp\left(-(\mu - 2L_2L_{\mathcal{Q}})\sum_{j=0}^{J-1}\eta_k^{t,j}\right)\mathbb{E}_{\xi}[\|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2] \\
&+ \sum_{j=0}^{J-1}\left(\eta_k^{t,j}\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta_k^{t,j})^2G^2\right)\exp\left(-(\mu - 2L_2L_{\mathcal{Q}})\sum_{i=j+1}^{J-1}\eta_k^{t,i}\right)
\end{aligned}
\tag{85}
$$

Consequently, we obtain the optimization error bound after J steps of stochastic gradient descent. This bound mainly comprises two components: the residual contribution of the initial parameter residual after exponential decay, and the accumulated effect of the stochastic gradient noise and the replay weight approximation error over the iteration history.

Next, we expand the loss function after the J-th iteration, yielding:

$$
\begin{aligned}
L_{k,\text{total}}^t(\theta_k^{t,J}|\lambda_k^{t,J}) = &L_{k,\text{task}}^t(\theta_k^{t,J}) + \lambda_k^{t,J}\cdot L_{k,\text{rep}}^t(\theta_k^{t,J}) \\
= &L_{k,\text{task}}^t(\theta_k^{t,J}) + (\lambda_k^{t,*} + (\lambda_k^{t,J} - \lambda_k^{t,*}))\cdot L_{k,\text{rep}}^t(\theta_k^{t,J}) \\
= &L_{k,\text{task}}^t(\theta_k^{t,J}) + \lambda_k^{t,*}\cdot L_{k,\text{rep}}^t(\theta_k^{t,J}) + (\lambda_k^{t,J} - \lambda_k^{t,*})\cdot L_{k,\text{rep}}^t(\theta_k^{t,J}) \\
= &L_{k,\text{total}}^t(\theta_k^{t,J}|\lambda_k^{t,*}) + (\lambda_k^{t,J} - \lambda_k^{t,*})\cdot L_{k,\text{rep}}^t(\theta_k^{t,J})
\end{aligned}
\tag{86}
$$

For the first term $L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,*}_k)$, under the L-smoothness in Assumption 10, we apply the second-order Taylor upper bound at $\theta^{t,*}_k$. Substituting the stationary condition $\nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k \mid \lambda^{t,*}_k) = \mathbf{0}$ yields:

$$\begin{aligned} L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,*}_k) \leq & L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) + \langle \nabla_\theta L^t_{k,\text{total}}(\theta^{t,*}_k \mid \lambda^{t,*}_k), \theta^{t,J}_k - \theta^{t,*}_k \rangle + \frac{L}{2}\|\theta^{t,J}_k - \theta^{t,*}_k\|_2^2 \\ = & L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) + 0 + \frac{L}{2}\|\theta^{t,J}_k - \theta^{t,*}_k\|_2^2 \end{aligned} \tag{87}$$

Taking the expectation of both sides with respect to ξ yields:

$$\mathbb{E}_\xi\left[L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,*}_k)\right] \leq L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) + \frac{L}{2}\mathbb{E}_\xi\left[\|\theta^{t,J}_k - \theta^{t,*}_k\|_2^2\right] \tag{88}$$

For the second term $(\lambda^{t,J}_k - \lambda^{t,*}_k)L^t_{k,\text{rep}}(\theta^{t,J}_k)$, taking its absolute value and then the expectation with respect to ξ yields:

$$\begin{aligned} \mathbb{E}_\xi\left[(\lambda^{t,J}_k - \lambda^{t,*}_k)\cdot L^t_{k,\text{rep}}(\theta^{t,J}_k)\right] \leq & \mathbb{E}_\xi\left[|(\lambda^{t,J}_k - \lambda^{t,*}_k)|\cdot|L^t_{k,\text{rep}}(\theta^{t,J}_k)|\right] \\ \leq & \mathbb{E}_\xi\left[|(\lambda^{t,J}_k - \lambda^{t,*}_k)|\right]\cdot\left(\sup_{\theta\in\Omega_\theta} L^t_{k,\text{rep}}(\theta)\right) \end{aligned} \tag{89}$$

Under Assumption 2, Ω_θ is a compact set with diameter $D_\theta \triangleq \sup_{\theta_a,\theta_b\in\Omega_\theta}\|\theta_a - \theta_b\|_2 < \infty$. By the $L_\mathcal{Q}$-Lipschitz continuity of the quotient mapping $\mathcal{Q}(\theta)$, we have $|\lambda^{t,J}_k - \lambda^{t,*}_k| \leq L_\mathcal{Q}D_\theta + \epsilon_{\text{approx}}$. Substituting this into the previous expression yields:

$$\mathbb{E}_\xi\left[(\lambda^{t,J}_k - \lambda^{t,*}_k)\cdot L^t_{k,\text{rep}}(\theta^{t,J}_k)\right] \leq (L_\mathcal{Q}D_\theta + \epsilon_{\text{approx}})\cdot\sup_{\theta\in\Omega_\theta} L^t_{k,\text{rep}}(\theta) \tag{90}$$

Consequently, taking the expectation of the loss function $L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,J}_k)$ after the J-th iteration with respect to ξ yields:

$$\begin{aligned} & \mathbb{E}_\xi\left[L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,J}_k)\right] \\ = & \mathbb{E}_\xi\left[L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,*}_k)\right] + \mathbb{E}_\xi\left[(\lambda^{t,J}_k - \lambda^{t,*}_k)\cdot L^t_{k,\text{rep}}(\theta^{t,J}_k)\right] \\ \leq & L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) + \frac{L}{2}\mathbb{E}_\xi\left[\|\theta^{t,J}_k - \theta^{t,*}_k\|_2^2\right] + (L_\mathcal{Q}D_\theta + \epsilon_{\text{approx}})\cdot\sup_{\theta\in\Omega_\theta} L^t_{k,\text{rep}}(\theta) \\ \leq & L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) + \frac{L}{2}\exp\left(-(\mu - 2L_2L_\mathcal{Q})\sum_{j=0}^{J-1}\eta^{t,j}_k\right)\mathbb{E}_\xi[\|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2] \\ & + \frac{L}{2}\sum_{j=0}^{J-1}\left(\eta^{t,j}_k\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta^{t,j}_k)^2G^2\right)\prod_{i=j+1}^{J-1}(1 - \eta^{t,i}_k(\mu - 2L_2L_\mathcal{Q})) \\ & + (L_\mathcal{Q}D_\theta + \epsilon_{\text{approx}})\cdot\sup_{\theta\in\Omega_\theta} L^t_{k,\text{rep}}(\theta) \end{aligned} \tag{91}$$

Consequently, we define the error upper bound function $\mathcal{E}^t_{k,\text{bound}}\left(\{\eta^{t,j}_k\}_{j=0}^{J-1}, \|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2\right)$ as:

$$\begin{aligned} & \mathcal{E}^t_{k,\text{bound}}\left(\{\eta^{t,j}_k\}_{j=0}^{J-1}, \|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2\right) \\ = & \frac{L}{2}\exp\left(-(\mu - 2L_2L_\mathcal{Q})\sum_{j=0}^{J-1}\eta^{t,j}_k\right)\mathbb{E}_\xi[\|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2] \\ & + \frac{L}{2}\sum_{j=0}^{J-1}\left(\eta^{t,j}_k\frac{{L_2}^2{\epsilon_{\text{approx}}}^2}{\mu} + (\eta^{t,j}_k)^2G^2\right)\prod_{i=j+1}^{J-1}(1 - \eta^{t,i}_k(\mu - 2L_2L_\mathcal{Q})) \\ & + (L_\mathcal{Q}D_\theta + \epsilon_{\text{approx}})\cdot\sup_{\theta\in\Omega_\theta} L^t_{k,\text{rep}}(\theta) \end{aligned} \tag{92}$$

Finally, we obtain:

$$\mathbb{E}\left[L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,J}_k)\right] \leq L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) + \mathcal{E}^t_{k,\text{bound}}\left(\{\eta^{t,j}_k\}_{j=0}^{J-1}, \|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2\right) \tag{93}$$

This completes the proof of Theorem 1.

B.4 PROOF OF THEOREM 2

Theorem 2 (Spectral Efficiency of the Holistic Buffer Maintenance Strategy (HBMS)) *Given $M > 2C_{\max}$ and standard assumptions (see Appendix B.1), let $C_{\text{rand}} \triangleq 1$ be the spectral efficiency constant of class-wise random sampling, and let $C_\kappa \triangleq \inf_{\theta\in\Omega_\theta,\hat{g}\in\mathcal{G}} \exp(-\beta(\theta,\hat{g})\Delta\rho(\theta,\hat{g}))$ (depending only on M and $C_{\max}$) be that of the Holistic Buffer Maintenance Strategy, where $\Delta\rho(\theta,\hat{g}) \triangleq \rho_{\text{rand}} - \rho_{\text{HBMS}}$ is the spectral radius gap between the two methods. Then, $0 < C_\kappa < C_{\text{rand}}$ holds for all $k \in \mathcal{K}$ and $t \in [T]$.*

Proof:

To establish a baseline for spectral efficiency comparison, we define the normalized spectral efficiency constant of class-wise random sampling as $C_{\text{rand}} \triangleq 1$.

Given that the kernel decay rate scales as $\beta \propto 1/d$, and the dimension d is intrinsically tied to the model parameters θ, we define the kernel decay rate as a continuous function $\beta(\theta,\hat{g})$ over the model parameters θ and the normalized logits vector $\hat{g}$, thereby satisfying the prerequisites for the Continuous Mapping Theorem and unifying the analysis domain of kernel spectral geometry. By Assumption 2, the model parameter space Ω_θ is a compact set; by Assumption 4, the range of the normalized logits vector $\hat{g}$, denoted as $\mathcal{G} \subseteq \mathbb{R}^d$, is also compact. Consequently, their Cartesian product $\Omega_\theta \times \mathcal{G}$ remains compact under the product topology, and $\beta(\theta,\hat{g})$ is a continuous real-valued function defined on this compact set. Furthermore, by the Weierstrass Extreme Value Theorem, the image set $\text{Im}(\beta) \triangleq \{\beta(\theta,\hat{g}) \mid (\theta,\hat{g}) \in \Omega_\theta \times \mathcal{G}\}$ is a bounded closed interval in $\mathbb{R}^+$. Thus, there exist a strictly positive lower bound $\beta_{\min} > 0$ and a finite upper bound $\beta_{\max} < \infty$ such that:

$$\beta_{\min} \triangleq \min_{(\theta,\hat{g})\in\Omega_\theta\times\mathcal{G}} \beta(\theta,\hat{g}), \quad \beta_{\max} \triangleq \max_{(\theta,\hat{g})\in\Omega_\theta\times\mathcal{G}} \beta(\theta,\hat{g}). \tag{94}$$

That is, the kernel decay rate β satisfies:

$$\beta \in [\beta_{\min}, \beta_{\max}] \tag{95}$$

By Assumption 7, the number of classes satisfies $|\mathcal{C}_k^{\leq t}| \leq C_{\max}$. Under the conditions that $|\mathcal{C}_k^{\leq t}| \geq 2$ and each class contains at least 2 samples, the number of samples $m_{k,c}^t$ for class c in the buffer satisfies:

$$2 \leq m_{k,c}^t = \lfloor M/|\mathcal{C}_k^{\leq t}| \rfloor \leq M/2. \tag{96}$$

Since the range $\mathcal{G}$ of the normalized logits vector $\hat{g}$ is a compact set, we define $r_{\text{opt}} \triangleq \sup_{S\subset\mathcal{G},|S|=m_{k,c}^t} \min_{x_i\neq x_j\in S} \|\hat{g}(x_i) - \hat{g}(x_j)\|_2$ as the theoretical optimal max-min distance when placing $m_{k,c}^t$ points in $\mathcal{G}$. Let the d-dimensional Lebesgue measure (volume) of $\mathcal{G}$ be $\text{Vol}(\mathcal{G})$, and the volume of the unit ball in the d-dimensional Euclidean space be $V_d = \pi^{d/2}/\Gamma(d/2+1)$. On one hand, since the sum of the volumes of $m_{k,c}^t$ disjoint balls with radius $r_{\text{opt}}/2$ cannot exceed the volume of the container $\mathcal{G}$, we can derive the upper bound δ_{ub} for the theoretical optimal max-min distance r_{opt}. On the other hand, when placing points sequentially in the compact set $\mathcal{G}$ based on a greedy strategy, if the total volume of the repulsion balls corresponding to the first $m_{k,c}^t - 1$ points does not exceed the volume of $\mathcal{G}$, there must exist a point configuration such that the minimum pairwise distance satisfies the theoretical lower bound δ_{lb}. That is:

$$m_{k,c}^t V_d \left(\frac{\delta_{\text{ub}}}{2}\right)^d = \text{Vol}(\mathcal{G}) \quad \Rightarrow \quad \delta_{\text{ub}} = 2\left(\frac{\text{Vol}(\mathcal{G})}{m_{k,c}^t V_d}\right)^{1/d} \tag{97}$$

$$(m_{k,c}^t - 1) V_d \left(\frac{\delta_{\text{lb}}}{2}\right)^d = \text{Vol}(\mathcal{G}) \quad \Rightarrow \quad \delta_{\text{lb}} = 2\left(\frac{\text{Vol}(\mathcal{G})}{(m_{k,c}^t - 1) V_d}\right)^{1/d} \tag{98}$$

By definition, the theoretical optimal max-min distance r_{opt} satisfies $\delta_{\text{ub}} \geq r_{\text{opt}} \geq \delta_{\text{lb}}$.

Next, consider the Holistic Buffer Maintenance Strategy proposed in our paper:

In the first stage, for each known class c, the algorithm selects the $\min\{2|\mathcal{M}_{k,c}^t|, |X_c|\}$ samples with the highest IDV values from the candidate pool X_c to construct the buffer subset $\tilde{\mathcal{M}}_{k,c}^t$:

$$\text{IDV}(x|c_x) = -\log \bar{p}_{\mathcal{M}_k^{t-1}}(c_x|x) + \lambda_1 \cdot \min_{x_i \in \mathcal{M}_k^{t-1}} \|\hat{g}(x) - \hat{g}(x_i)\|_2^2 \tag{99}$$

Since the information term does not contribute to the optimization of the kernel condition number, maximizing the IDV value of a sample x is equivalent to maximizing the distance term $\min_{x_i \in \mathcal{M}_k^{t-1}} \|\hat{g}(x) - \hat{g}(x_i)\|_2^2$. According to the theoretical guarantee of the greedy max-min distance algorithm, in the metric space $(\mathcal{G}, \|\cdot\|_2)$, the minimum distance $\delta_{\min}^{\text{IDV}}$ obtained by this algorithm satisfies $\delta_{\min}^{\text{IDV}} \geq r_{\text{opt}}/2$. Consequently, we define the max-min distance approximation ratio as $\nu_{\text{pack}} \triangleq \delta_{\min}^{\text{IDV}}/r_{\text{opt}}$. It is evident that $\nu_{\text{pack}} \leq 1$, and it satisfies:

$$\nu_{\text{pack}} \triangleq \frac{\delta_{\min}^{\text{IDV}}}{r_{\text{opt}}} > \frac{r_{\text{opt}}/2}{r_{\text{opt}}} = \frac{1}{2} \tag{100}$$

That is, the max-min distance approximation ratio $\nu_{\text{pack}} \in [0.5, 1]$.

In the second stage, for each known class c, the algorithm selects the $|\mathcal{M}_{k,c}^t|$ samples with the highest CDV values from the candidate subset $\tilde{\mathcal{M}}_{k,c}^t$ to construct the final buffer subset $\mathcal{M}_{k,c}^t$ for class c:

$$\text{CDV}(x|c_x) = \log \frac{p(c_x|x)}{\bar{p}_{\mathcal{M}_k^{t-1} \cup \{x\}}(c_x \mid x)} + \lambda_2 \cdot \min_{x_i \in \mathcal{M}_k^{t-1}} \|\hat{g}(x) - \hat{g}(x_i)\|_2^2 \tag{101}$$

Obviously, the CDV selection operation can strictly and monotonically optimize the minimum pairwise distance of the point set $\min_{i \neq j,\, x_i, x_j \in \mathcal{M}_k^t} \|\hat{g}(x_i) - \hat{g}(x_j)\|_2$, thereby effectively retaining samples with high dispersion. Consequently, the minimum pairwise distance of the final point set in the buffer maintenance algorithm, defined as $\delta_{\min}^{\text{KS}} \triangleq \min_{i \neq j,\, x_i, x_j \in \mathcal{M}_k^t} \|\hat{g}(x_i) - \hat{g}(x_j)\|_2$, inevitably satisfies $\delta_{\min}^{\text{KS}} \geq \delta_{\min}^{\text{IDV}}$. That is, the minimum pairwise Euclidean distance $\delta_{\min}^{\text{KS}}$ obtained by the two-stage greedy selection algorithm satisfies:

$$\delta_{\min}^{\text{KS}} \geq \delta_{\min}^{\text{IDV}} \geq \nu_{\text{pack}} \cdot r_{\text{opt}} \geq \nu_{\text{pack}} \cdot \delta_{\text{lb}} = 2\nu_{\text{pack}} \left(\frac{\text{Vol}(\mathcal{G})}{(m_{k,c}^t - 1)V_d} \right)^{1/d} \tag{102}$$

where the max-min distance approximation ratio $\nu_{\text{pack}} \in [0.5, 1]$.

Next, consider the class-wise random sampling algorithm.

For class c, consider independently and uniformly sampling $m_{k,c}^t$ points in the d-dimensional compact set $\mathcal{G}$, and define the minimum Euclidean distance among all pairs of points as $\delta_{\min}^{\text{rand}} \triangleq \min_{1 \leq i < j \leq m_{k,c}^t} \|x_i - x_j\|_2$. For sufficiently small r, the probability that the distance between any pair of points is greater than r, i.e., the probability of the event $\{\delta_{\min}^{\text{rand}} > r\}$, is given by:

$$\mathbb{P}(\delta_{\min}^{\text{rand}} > r) = \prod_{i=1}^{m_{k,c}^t - 1} \mathbb{P}\left(\|x_{i+1} - x_j\|_2 > r,\ \forall j \leq i \,\middle|\, x_1, \ldots, x_i \right) \tag{103}$$

Let $B(x, r)$ denote the closed ball centered at x with radius r. The above equation implies that, given the positions of the first i points, the $(i+1)$-th point must fall within the region $\mathcal{G} \setminus \bigcup_{j=1}^i B(x_j, r)$. The volume of this feasible region is $\text{Vol}_{\text{feasible}}^{i+1} = \text{Vol}(\mathcal{G}) - \text{Vol}\left(\left(\bigcup_{j=1}^i B(x_j, r) \right) \cap \mathcal{G} \right)$. Furthermore, when $r \ll \text{diam}(\mathcal{G})$, the overlapping volumes among the balls and the boundary truncation effects can be neglected. Combined with the volume of the unit ball in the d-dimensional Euclidean space, $V_d = \pi^{d/2}/\Gamma(d/2+1)$, we obtain $\text{Vol}_{\text{feasible}}^{i+1} \approx \text{Vol}(\mathcal{G}) - \sum_{j=1}^i \text{Vol}(B(x_j, r)) = \text{Vol}(\mathcal{G}) - i \cdot V_d r^d$. Consequently, the conditional probability satisfies:

$$\mathbb{P}\left(\|x_{i+1} - x_j\|_2 > r,\ \forall j \leq i \,\middle|\, x_1, \ldots, x_i \right) \approx \frac{\text{Vol}(\mathcal{G}) - i \cdot V_d r^d}{\text{Vol}(\mathcal{G})} = 1 - \frac{i \cdot V_d r^d}{\text{Vol}(\mathcal{G})} \tag{104}$$

Substituting this into the original expression and applying the inequality $1 - x \leq \exp(-x)$, we obtain:

$$\mathbb{P}(\delta_{\min}^{\text{rand}} > r) \approx \prod_{i=1}^{m_{k,c}^t - 1} \left(1 - \frac{i \cdot V_d r^d}{\text{Vol}(\mathcal{G})}\right) \leq \prod_{i=1}^{m_{k,c}^t - 1} \exp\left(-\frac{i \cdot V_d r^d}{\text{Vol}(\mathcal{G})}\right) = \exp\left(-\frac{V_d r^d}{\text{Vol}(\mathcal{G})} \sum_{i=1}^{m_{k,c}^t - 1} i\right) \tag{105}$$

Substituting the arithmetic series sum formula $\sum_{i=1}^{m_{k,c}^t - 1} i = \frac{m_{k,c}^t (m_{k,c}^t - 1)}{2}$, we obtain:

$$\mathbb{P}(\delta_{\min}^{\text{rand}} > r) \leq \exp\left(-\frac{m_{k,c}^t (m_{k,c}^t - 1) V_d r^d}{2\text{Vol}(\mathcal{G})}\right) \tag{106}$$

Since the expectation of the non-negative random variable $\delta_{\min}^{\text{rand}}$ can be expressed as the integral of its survival function $\int_0^\infty \mathbb{P}(\delta_{\min}^{\text{rand}} > r)\mathrm{d}r$, we obtain:

$$\mathbb{E}[\delta_{\min}^{\text{rand}}] \leq \int_0^\infty \exp\left(-\frac{m_{k,c}^t (m_{k,c}^t - 1) V_d}{2\text{Vol}(\mathcal{G})} r^d\right) \mathrm{d}r \tag{107}$$

Let $u = \frac{m_{k,c}^t (m_{k,c}^t - 1) V_d}{2\text{Vol}(\mathcal{G})} r^d$. Then $r = \left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d} u^{1/d}$ and $\mathrm{d}r = \frac{1}{d}\left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d} u^{\frac{1}{d}-1}\mathrm{d}u$. Substituting these into the above expression, we obtain:

$$\begin{aligned}
\mathbb{E}[\delta_{\min}^{\text{rand}}] &\leq \int_0^\infty \exp(-u) \frac{1}{d} \left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d} u^{\frac{1}{d}-1}\mathrm{d}u \\
&= \left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d} \int_0^\infty \frac{1}{d} \exp(-u)\, u^{\frac{1}{d}-1}\mathrm{d}u
\end{aligned} \tag{108}$$

Given the Gamma function $\Gamma(z) \triangleq \int_0^\infty e^{-t} t^{z-1}\mathrm{d}t$ and its recurrence identity $\Gamma(z+1) = z\Gamma(z)$, substituting $z = 1/d$ yields:

$$\mathbb{E}[\delta_{\min}^{\text{rand}}] \leq \left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d} \frac{1}{d}\Gamma\left(\frac{1}{d}\right) = \Gamma\left(1 + \frac{1}{d}\right) \left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d} \tag{109}$$

From the above, we have derived $\delta_{\min}^{\text{KS}} \geq 2\nu_{\text{pack}} \left(\frac{\text{Vol}(\mathcal{G})}{(m_{k,c}^t - 1) V_d}\right)^{1/d}$. Taking the ratio of the minimum pairwise distances obtained by the two sampling schemes, we obtain:

$$\begin{aligned}
\frac{\delta_{\min}^{\text{KS}}}{\mathbb{E}[\delta_{\min}^{\text{rand}}]} &\geq \frac{2\nu_{\text{pack}} \left(\frac{\text{Vol}(\mathcal{G})}{(m_{k,c}^t - 1) V_d}\right)^{1/d}}{\Gamma\left(1 + \frac{1}{d}\right) \left(\frac{2\text{Vol}(\mathcal{G})}{m_{k,c}^t (m_{k,c}^t - 1) V_d}\right)^{1/d}} \\
&= \frac{2\nu_{\text{pack}}}{\Gamma\left(1 + \frac{1}{d}\right)} \cdot \left(\frac{\text{Vol}(\mathcal{G})}{(m_{k,c}^t - 1) V_d} \cdot \frac{m_{k,c}^t (m_{k,c}^t - 1) V_d}{2\text{Vol}(\mathcal{G})}\right)^{1/d} \\
&= \frac{2\nu_{\text{pack}}}{\Gamma\left(1 + \frac{1}{d}\right)} \cdot \left(\frac{m_{k,c}^t}{2}\right)^{1/d} = \frac{2^{(1-1/d)}\nu_{\text{pack}}}{\Gamma\left(1 + \frac{1}{d}\right)} \cdot \left(m_{k,c}^t\right)^{1/d}
\end{aligned} \tag{110}$$

Substituting $m_{k,c}^t \geq 2$ and $\nu_{\text{pack}} \in [0.5, 1]$, we obtain:

$$\begin{aligned}
\frac{\delta_{\min}^{\text{KS}}}{\mathbb{E}[\delta_{\min}^{\text{rand}}]} &\geq \frac{2^{(1-1/d)}\nu_{\text{pack}}}{\Gamma\left(1 + \frac{1}{d}\right)} \cdot \left(m_{k,c}^t\right)^{1/d} \geq \frac{2^{(1-1/d)}\nu_{\text{pack}}}{\Gamma\left(1 + \frac{1}{d}\right)} \cdot 2^{1/d} = \frac{2\nu_{\text{pack}}}{\Gamma\left(1 + \frac{1}{d}\right)} \\
&> \frac{2 \cdot 0.5}{\Gamma\left(1 + \frac{1}{d}\right)} = \frac{1}{\Gamma\left(1 + \frac{1}{d}\right)}
\end{aligned} \tag{111}$$

Since $d \geq 2$, we have $1 + 1/d \in (1, 1.5] \subset (1, 2)$. It is known that $\Gamma(x) < 1$ for all $x \in (1, 2)$ (i.e., $\Gamma(x) < \Gamma(1) = 1$), and $\Gamma(1.5) = \sqrt{\pi}/2$. Therefore, we have $\Gamma(1 + 1/d) < 1$. Substituting this into the above expression yields:

$$\frac{\delta_{\min}^{\mathrm{KS}}}{\mathbb{E}[\delta_{\min}^{\mathrm{rand}}]} > \frac{1}{\Gamma\left(1 + \frac{1}{d}\right)} > \frac{1}{1} = 1 \tag{112}$$

Furthermore, consider the Holistic Buffer Maintenance Strategy proposed in our paper. For class c, this strategy constructs an $m_{k,c}^t \times m_{k,c}^t$ Gaussian kernel matrix $\mathbf{K}^{\mathrm{KS}}$, whose entries are defined as follows: for any pair of samples (x_i, x_j) (where $i, j \in \{1, \ldots, m_{k,c}^t\}$ and $i \neq j$), $K(x_i, x_j) = \exp\left(-\beta \|\hat{g}(x_i) - \hat{g}(x_j)\|_2^2\right)$. Since the pairwise distance satisfies $\|\hat{g}(x_i) - \hat{g}(x_j)\|_2 \geq \delta_{\min}^{\mathrm{KS}}$, which implies that the squared distance $\|\hat{g}(x_i) - \hat{g}(x_j)\|_2^2 \geq (\delta_{\min}^{\mathrm{KS}})^2$, any off-diagonal entry is bounded above by $K(x_i, x_j) \leq \exp\left(-\beta(\delta_{\min}^{\mathrm{KS}})^2\right)$. Summing this upper bound over all off-diagonal entries yields:

$$\sum_{j=1, j\neq i}^{m_{k,c}^t} |K(x_i, x_j)| \leq \sum_{j=1, j\neq i}^{m_{k,c}^t} \exp\left(-\beta\left(\delta_{\min}^{\mathrm{KS}}\right)^2\right) = (m_{k,c}^t - 1)\exp\left(-\beta\left(\delta_{\min}^{\mathrm{KS}}\right)^2\right) \tag{113}$$

The kernel matrix $\mathbf{K}^{\mathrm{KS}}$ is a real symmetric matrix, and by the definition of the normalized Gaussian kernel, all its diagonal entries satisfy $K(x_i, x_i) = 1$. According to the Gershgorin Circle Theorem, any eigenvalue λ of $\mathbf{K}^{\mathrm{KS}}$ must lie within a closed disk centered at the diagonal entry 1 with a radius equal to the sum of the absolute values of the off-diagonal entries in the corresponding row, $R_i = \sum_{j\neq i} |K(x_i, x_j)|$. Defining the row-sum upper bound of the spectral radius as $\rho_{\mathrm{KS}}^{\mathrm{up}} \triangleq \max_i R_i$, which is the maximum of the row sums of the absolute values of the off-diagonal entries, we obtain:

$$\rho_{\mathrm{KS}}^{\mathrm{up}} = \max_i \sum_{j=1, j\neq i}^{m_{k,c}^t} |K(x_i, x_j)| \leq (m_{k,c}^t - 1)\exp\left(-\beta\left(\delta_{\min}^{\mathrm{KS}}\right)^2\right) \tag{114}$$

Since all eigenvalues of $\mathbf{K}^{\mathrm{KS}}$ lie within the closed interval $[1 - \rho_{\mathrm{KS}}^{\mathrm{up}},\ 1 + \rho_{\mathrm{KS}}^{\mathrm{up}}]$, its spectral condition number $\kappa_{\mathrm{KS}} \triangleq \lambda_{\max}(\mathbf{K}^{\mathrm{KS}})/\lambda_{\min}(\mathbf{K}^{\mathrm{KS}})$ satisfies:

$$\kappa_{\mathrm{KS}} \leq \frac{1 + \rho_{\mathrm{KS}}^{\mathrm{up}}}{1 - \rho_{\mathrm{KS}}^{\mathrm{up}}} \tag{115}$$

Similarly, for the class-wise random sampling method, adopting the kernel function $K(x_i, x_j) = \exp\left(-\beta \|\hat{g}(x_i) - \hat{g}(x_j)\|_2^2\right)$, we define the proxy lower bound for the expected spectral radius of its Gaussian kernel matrix $\mathbf{K}^{\mathrm{rand}}$ as $\rho_{\mathrm{rand}}^{\mathrm{low}}$:

$$\rho_{\mathrm{rand}}^{\mathrm{low}} \triangleq (m_{k,c}^t - 1)\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right). \tag{116}$$

Since $\rho_{\mathrm{rand}}^{\mathrm{low}}$ is a lower bound of ρ_{rand}, the spectral condition number $\kappa_{\mathrm{rand}} \triangleq \lambda_{\max}(\mathbf{K}^{\mathrm{rand}})/\lambda_{\min}(\mathbf{K}^{\mathrm{rand}})$ of $\mathbf{K}^{\mathrm{rand}}$ satisfies:

$$\kappa_{\mathrm{rand}} \geq \frac{1 + \rho_{\mathrm{rand}}^{\mathrm{low}}}{1 - \rho_{\mathrm{rand}}^{\mathrm{low}}} \tag{117}$$

Thus, a conservative upper bound on the ratio of the spectral condition numbers $\kappa_{\mathrm{KS}}/\kappa_{\mathrm{rand}}$ is given by:

$$\frac{\kappa_{\mathrm{KS}}}{\kappa_{\mathrm{rand}}} \leq \frac{1 + \rho_{\mathrm{KS}}^{\mathrm{up}}}{1 - \rho_{\mathrm{KS}}^{\mathrm{up}}} \cdot \frac{1 - \rho_{\mathrm{rand}}^{\mathrm{low}}}{1 + \rho_{\mathrm{rand}}^{\mathrm{low}}} \tag{118}$$

Define the function $f(\rho) \triangleq \ln \frac{1+\rho}{1-\rho}$ with the domain $\rho \in [0, 1)$. Since $f'(\rho) = \frac{2}{(1-\rho)^2} > 0$, $f(\rho)$ is strictly increasing and continuously differentiable on its domain. Taking the natural logarithm

of both sides of the above inequality, and letting ξ lie between $\rho_{\mathrm{KS}}^{\mathrm{up}}$ and $\rho_{\mathrm{rand}}^{\mathrm{low}}$, we obtain by the Lagrange Mean Value Theorem:

$$\ln\left(\frac{\kappa_{\mathrm{KS}}}{\kappa_{\mathrm{rand}}}\right) \leq f(\rho_{\mathrm{KS}}^{\mathrm{up}}) - f(\rho_{\mathrm{rand}}^{\mathrm{low}}) = f'(\xi)\left(\rho_{\mathrm{KS}}^{\mathrm{up}} - \rho_{\mathrm{rand}}^{\mathrm{low}}\right) \tag{119}$$

Defining the boundary difference $\Delta\rho^{\mathrm{bound}} \triangleq \rho_{\mathrm{rand}}^{\mathrm{low}} - \rho_{\mathrm{KS}}^{\mathrm{up}}$, we obtain:

$$\begin{aligned}
\Delta\rho^{\mathrm{bound}} \geq & (m_{k,c}^t - 1)\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right) - (m_{k,c}^t - 1)\exp\left(-\beta(\delta_{\min}^{\mathrm{KS}})^2\right) \\
= & (m_{k,c}^t - 1)\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)\left(1 - \frac{\exp\left(-\beta(\delta_{\min}^{\mathrm{KS}})^2\right)}{\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)}\right) \\
= & (m_{k,c}^t - 1)\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)\left(1 - \exp\left(-\beta[(\delta_{\min}^{\mathrm{KS}})^2 - (\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2]\right)\right). \\
= & (m_{k,c}^t - 1)\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)\left(1 - \exp\left(-\beta\left[\left(\frac{(\delta_{\min}^{\mathrm{KS}})}{\mathbb{E}[\delta_{\min}^{\mathrm{rand}}]}\right)^2 - 1\right](\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)\right) \\
> & (m_{k,c}^t - 1)\exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)\left(1 - \exp\left(-\beta(\mathbb{E}[\delta_{\min}^{\mathrm{rand}}])^2\right)\right)
\end{aligned} \tag{120}$$

Note that $\Delta\rho^{\mathrm{bound}}(\beta, m_{k,c}^t, \delta)$ is a multivariate continuous elementary function with respect to $\beta, m_{k,c}^t$, and δ. Since the range $\mathcal{G} \subseteq \mathbb{R}^d$ of the normalized logits vector $\hat{g}$ is a compact set, there exists a maximum distance in the feature space (i.e., the diameter) $\overline{\delta} = \mathrm{diam}(\mathcal{G}) < \infty$. Given that $m_{k,c}^t \leq M/2$, we have $\delta_{\min}^{\mathrm{KS}} \geq 2\nu_{\mathrm{pack}}\left(\frac{\mathrm{Vol}(\mathcal{G})}{(M/2)V_d}\right)^{1/d} > \nu_{\mathrm{pack}}\left(\frac{\mathrm{Vol}(\mathcal{G})}{MV_d}\right)^{1/d}$. Defining the conservative lower bound of the minimum pairwise distance as $\underline{\delta} \triangleq \nu_{\mathrm{pack}}\left(\frac{\mathrm{Vol}(\mathcal{G})}{MV_d}\right)^{1/d}$, the parameter domain $\mathcal{V}^{\mathrm{bound}}$ of $\Delta\rho^{\mathrm{bound}}(\beta, m_{k,c}^t, \delta)$ is given by:

$$\mathcal{V}^{\mathrm{bound}} \triangleq [\beta_{\min}, \beta_{\max}] \times [2, M/2] \times [\underline{\delta}, \overline{\delta}] \tag{121}$$

That is, $\mathcal{V}^{\mathrm{bound}}$ forms a compact hyperrectangle in $\mathbb{R}^3$. By the Weierstrass Extreme Value Theorem, the continuous function $\Delta\rho^{\mathrm{bound}}$ attains its minimum $\Delta\rho_{\min}^{\mathrm{bound}}$ on this compact set, which satisfies:

$$\Delta\rho^{\mathrm{bound}} \geq \Delta\rho_{\min}^{\mathrm{bound}} \triangleq \inf_{\mathcal{V}^{\mathrm{bound}}} \Delta\rho^{\mathrm{bound}} > 0 \tag{122}$$

Thus, we have $\Delta\rho^{\mathrm{bound}} > 0$. Combined with $f'(\xi) > 0$, we can derive the upper bound on the ratio of the spectral condition numbers:

$$\frac{\kappa_{\mathrm{KS}}}{\kappa_{\mathrm{rand}}} \leq \exp\left(f(\rho_{\mathrm{KS}}^{\mathrm{up}}) - f(\rho_{\mathrm{rand}}^{\mathrm{low}})\right) = \exp\left(f'(\xi)\left(\rho_{\mathrm{KS}}^{\mathrm{up}} - \rho_{\mathrm{rand}}^{\mathrm{low}}\right)\right) = \exp\left(-f'(\xi)\Delta\rho^{\mathrm{bound}}\right) < 1 \tag{123}$$

Furthermore, for notational consistency, let $\rho_{\mathrm{HBMS}} \triangleq \rho_{\mathrm{KS}}$. Based on the previous derivations, we have $\rho_{\mathrm{HBMS}} \leq \rho_{\mathrm{KS}}^{\mathrm{up}}$ and $\rho_{\mathrm{rand}} \geq \rho_{\mathrm{rand}}^{\mathrm{low}}$. Defining the true spectral radius difference as $\Delta\rho(\theta, \hat{g}) \triangleq \rho_{\mathrm{rand}} - \rho_{\mathrm{HBMS}}$, we obtain $\Delta\rho(\theta, \hat{g}) \geq \rho_{\mathrm{rand}}^{\mathrm{low}} - \rho_{\mathrm{KS}}^{\mathrm{up}} = \Delta\rho^{\mathrm{bound}}$. Combined with the previous conclusion that $\Delta\rho^{\mathrm{bound}} \geq \Delta\rho_{\min}^{\mathrm{bound}} > 0$, it follows that the global infimum of the true spectral radius difference, $\Delta\rho_{\min} \triangleq \inf_{\theta\in\Omega_\theta, \hat{g}\in\mathcal{G}} \Delta\rho(\theta, \hat{g})$, satisfies:

$$\Delta\rho(\theta, \hat{g}) \geq \rho_{\mathrm{rand}}^{\mathrm{low}} - \rho_{\mathrm{KS}}^{\mathrm{up}} = \Delta\rho^{\mathrm{bound}} \geq \Delta\rho_{\min}^{\mathrm{bound}} > 0 \tag{124}$$

Consequently, the composite function $\beta(\theta, \hat{g})\Delta\rho(\theta, \hat{g})$ has a strictly positive lower bound on the compact set $\Omega_\theta \times \mathcal{G}$. Defining the global spectral efficiency constant of the Holistic Buffer Maintenance Strategy as $C_\kappa \triangleq \inf_{\theta\in\Omega_\theta, \hat{g}\in\mathcal{G}} \exp\left(-\beta(\theta, \hat{g})\Delta\rho(\theta, \hat{g})\right)$, it satisfies:

$$C_\kappa \leq \exp\left(-\beta_{\min}\Delta\rho_{\min}\right) < 1 = C_{\mathrm{rand}}. \tag{125}$$

This result applies to the RKHS generalization error term, demonstrating that our method can reduce the generalization error upper bound induced by feature space complexity by a deterministic factor. This completes the proof of Theorem 2.

B.5 Proof of Lemma 2

Lemma 2 (Recursive Regret Bound for Local Training in FedKACE) *Under standard assumptions (see Appendix B.1) and a buffer capacity $M \geq 2C_{\max}$, consider the t-th FL round. Let n_k^t be the local data size for client k at FL round t. Let $f_{k,\text{cum}}^{t,*}$ be the optimal model under the cumulative distribution $p_k^{\leq t}$, parameterized by $\theta_{k,\text{cum}}^{t,*} \triangleq \arg\min_{\theta\in\Omega_\theta} \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_\theta)]$. Meanwhile, define the regret of model θ on client k at FL round t as $\text{Regret}_k^t(\theta) \triangleq \mathbb{E}_{p_k^{\leq t}}\left[\mathcal{R}(f_\theta) - \mathcal{R}(f_{k,\text{cum}}^{t,*})\right]$, and the local training regret functional $\text{Regret}_k^{\text{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \triangleq \text{Regret}_k^t(\theta_k^{t,J})$, where $\theta_k^{t,0}$ and $\{\eta_k^{t,j}\}_{j=0}^{J-1}$ are the initial parameter and learning rate sequence, respectively. This functional satisfies the following recursive upper bound:*

$$\begin{aligned}\text{Regret}_k^{\text{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \leq{}& \left(1-\omega_k^t\right)\text{Regret}_k^{\text{local}}(t-1;\theta_k^{t-1,0},\{\eta_k^{t-1,j}\}_{j=0}^{J-1}) \\ &+ \omega_k^t \cdot U_k(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\ &+ \left(1-\omega_k^t\right)\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}}) - \mathcal{R}(f_{\theta_k^{t-1,J}})\right]\end{aligned} \tag{126}$$

where $\omega_k^t \triangleq n_k^t / \sum_{\tau=1}^t n_k^\tau$ is the weight of p_k^t in $p_k^{\leq t}$, and the optimization error term $U_k(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1})$ is defined as:

$$\begin{aligned}&U_k(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\ \triangleq{}&\mathcal{E}_{k,\textit{bound}}^t\left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0}-\theta_k^{t,*}\|_2^2\right) + \left[\lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\text{lip}}D_\theta\right] \\ &+ \lambda_{\max}C_\kappa \cdot \left(2 + \left(8L_{\text{lip}}B + 4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}}\right)\end{aligned} \tag{127}$$

where $D_\theta \triangleq \sup_{\theta_a,\theta_b\in\Omega_\theta}\|\theta_a-\theta_b\|_2$ is the parameter space diameter; B is the norm radius bound of hypothesis class $\mathcal{F}_\theta$ in RKHS $\mathcal{H}_K$; $\lambda_{\max}$ is the maximum replay loss weight; and the optimization error bound $\mathcal{E}_{k,\textit{bound}}^t\left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0}-\theta_k^{t,*}\|_2^2\right)$ and kernel spectral efficiency constant C_κ are defined in Theorem 1 and Theorem 2, respectively. For the initial FL round $t=1$, the regret functional is upper-bounded by:

$$\begin{aligned}\text{Regret}_k^{\text{local}}(1;\theta_k^{1,0},\{\eta_k^{1,j}\}_{j=0}^{J-1}) \leq{}&\frac{L}{2}\exp\left(-\mu\sum_{j=0}^{J-1}\eta_k^{1,j}\right)\|\theta_k^{1,0}-\theta_{1,\text{cum}}^{1,*}\|_2^2 \\ &+ \frac{L}{2}\sum_{j=0}^{J-1}(\eta_k^{1,j})^2G^2\prod_{i=j+1}^{J-1}(1-\mu\eta_k^{1,i}) \\ &+ \frac{4L_{\text{Lip}}B}{\sqrt{n_k^1}} + \sqrt{\frac{8\log(2/\delta)}{n_k^1}}\end{aligned} \tag{128}$$

Proof:

Let n_k^τ be the local data size and p_k^τ be the corresponding empirical distribution received by client k at round τ. The cumulative data distribution over the first t rounds is defined as the data-size-weighted average of the empirical distributions, i.e., $p_k^{\leq t} \triangleq \frac{1}{\sum_{\tau=1}^t n_k^\tau}\sum_{\tau=1}^t n_k^\tau p_k^\tau$. Furthermore, let $\omega_k^t \triangleq n_k^t/\sum_{\tau=1}^t n_k^\tau$ denote the weight of the t-th FL round's distribution p_k^t, which can be recursively expressed as:

$$\begin{aligned}p_k^{\leq t} ={}&\frac{1}{\sum_{\tau=1}^t n_k^\tau}\sum_{\tau=1}^{t-1}n_k^\tau p_k^\tau + \frac{n_k^\tau}{\sum_{\tau=1}^t n_k^\tau}p_k^t = \frac{\sum_{\tau=1}^{t-1}n_k^\tau}{\sum_{\tau=1}^t n_k^\tau}\left(\frac{1}{\sum_{\tau=1}^{t-1}n_k^\tau}\sum_{\tau=1}^{t-1}n_k^\tau p_k^\tau\right) + \frac{n_k^\tau}{\sum_{\tau=1}^t n_k^\tau}p_k^t \\ ={}&\frac{\sum_{\tau=1}^{t-1}n_k^\tau}{\sum_{\tau=1}^t n_k^\tau}p_k^{\leq t-1} + \frac{n_k^\tau}{\sum_{\tau=1}^t n_k^\tau}p_k^t = (1-\omega_k^t)p_k^{\leq t-1} + \omega_k^t p_k^t\end{aligned} \tag{129}$$

Clearly, $p_k^{\leq t}$ is a convex combination of $p_k^{\leq t-1}$ and p_k^t. Substituting this into the definition of regret yields:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
=&\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] = \mathbb{E}_{(1-\omega_k^t)p_k^{\leq t-1}+\omega_k^t p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
=&(1-\omega_k^t)\mathbb{E}_{p_k^{\leq t-1}}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]+\omega_k^t\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]
\end{aligned} \tag{130}$$

For the first term $\mathbb{E}_{p_k^{\leq t-1}}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$, we expand as follows:

$$\begin{aligned}
&\mathbb{E}_{p_k^{\leq t-1}}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
=&\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{\theta_k^{t-1,J}})+\mathcal{R}(f_{\theta_k^{t-1,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t-1,*})+\mathcal{R}(f_{k,\mathrm{cum}}^{t-1,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})\right] \\
=&\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{\theta_k^{t-1,J}})\right]+\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t-1,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t-1,*})\right] \\
&+\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\mathrm{cum}}^{t-1,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})\right] \\
=&\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{\theta_k^{t-1,J}})\right]+\mathrm{Regret}_k^{\mathrm{local}}(t-1;\theta_k^{t-1,0},\{\eta_k^{t-1,j}\}_{j=0}^{J-1}) \\
&+\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\mathrm{cum}}^{t-1,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})\right]
\end{aligned} \tag{131}$$

Let $\Delta_{k,\mathrm{shift}}(t) \triangleq \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\mathrm{cum}}^{t-1,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})\right]$ denote the distribution shift term. Substituting this into the above inequality yields the recursive bound for the local regret:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
=&\left(1-\omega_k^t\right)\mathrm{Regret}_k^{\mathrm{local}}(t-1;\theta_k^{t-1,0},\{\eta_k^{t-1,j}\}_{j=0}^{J-1}) \\
&+\left(1-\omega_k^t\right)\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{\theta_k^{t-1,J}})\right]+\left(1-\omega_k^t\right)\Delta_{k,\mathrm{shift}}(t) \\
&+\omega_k^t\cdot\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]
\end{aligned} \tag{132}$$

Next, consider the term $\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$. Since the data D_k^t are drawn i.i.d. from the distribution p_k^t, we have $\mathbb{E}_{p_k^t}[\mathcal{R}(f_\theta)] = \mathbb{E}_{p_k^t}[L_{k,\mathrm{task}}^t(\theta)]$, which yields:

$$\begin{aligned}
&\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
=&\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})]-\mathbb{E}_{p_k^t}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
=&\mathbb{E}_{p_k^t}[L_{k,\mathrm{task}}^t(\theta_k^{t,J})]-\mathbb{E}_{p_k^t}[L_{k,\mathrm{task}}^t(\theta_k^{t,*})]+\mathbb{E}_{p_k^t}[L_{k,\mathrm{task}}^t(\theta_k^{t,*})]-\mathbb{E}_{p_k^t}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
=&\left(L_{k,\mathrm{task}}^t(\theta_k^{t,J})-L_{k,\mathrm{task}}^t(\theta_k^{t,*})\right)+\left(\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,*}})]-\mathbb{E}_{p_k^t}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right)
\end{aligned} \tag{133}$$

Further, at the J-th epoch, the total loss is given by $L_{k,\mathrm{total}}^t(\theta) = L_{k,\mathrm{task}}^t(\theta)+\lambda_k^{t,J}L_{k,\mathrm{rep}}^t(\theta)$, which yields $L_{k,\mathrm{task}}^t(\theta) = L_{k,\mathrm{total}}^t(\theta)-\lambda_k^{t,J}L_{k,\mathrm{rep}}^t(\theta)$. Additionally, the replay loss can be expressed as $L_{k,\mathrm{rep}}^t(\theta) = \hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_\theta)$. Noting that $\theta_k^{t,J}$ and $\theta_k^{t,*}$ are fixed parameters, we obtain:

$$\begin{aligned}
&\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
=&\left(L_{k,\mathrm{total}}^t(\theta_k^{t,J})-L_{k,\mathrm{total}}^t(\theta_k^{t,*})\right)-\lambda_k^{t,J}\left(L_{k,\mathrm{rep}}^t(\theta_k^{t,J})-L_{k,\mathrm{rep}}^t(\theta_k^{t,*})\right) \\
&+\left(\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,*}})]-\mathbb{E}_{p_k^t}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right) \\
=&\left(L_{k,\mathrm{total}}^t(\theta_k^{t,J})-L_{k,\mathrm{total}}^t(\theta_k^{t,*})\right)-\lambda_k^{t,J}\left(\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,J}})-\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,*}})\right) \\
&+\left(\mathbb{E}_{p_k^t}[\mathcal{R}(f_{\theta_k^{t,*}})]-\mathbb{E}_{p_k^t}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right)
\end{aligned} \tag{134}$$

For the first term $L^t_{k,\text{total}}(\theta^{t,J}_k) - L^t_{k,\text{total}}(\theta^{t,*}_k)$, applying Theorem 1 yields:

$$\begin{aligned} L^t_{k,\text{total}}(\theta^{t,J}_k) - L^t_{k,\text{total}}(\theta^{t,*}_k) &= \mathbb{E}\left[L^t_{k,\text{total}}(\theta^{t,J}_k|\lambda^{t,J}_k)\right] - L^t_{k,\text{total}}(\theta^{t,*}_k|\lambda^{t,*}_k) \\ &\leq \mathcal{E}^t_{k,\text{bound}}\left(\{\eta^{t,j}_k\}_{j=0}^{J-1}, \|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2\right) \end{aligned} \tag{135}$$

For the second term $-\lambda^{t,J}_k\left(\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k})\right)$, let $\lambda_{\max} \triangleq \sup_{\lambda\in\Omega_\lambda}\lambda < \infty$ denote the upper bound of the global replay weight. By taking absolute values and applying a conservative bound, we obtain:

$$\begin{aligned} &-\lambda^{t,J}_k\left(\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k})\right) \\ =&-\lambda^{t,J}_k\left(\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k}) + \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right) \\ &-\lambda^{t,J}_k\left(\mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k}) - \hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k})\right) \\ =&-\lambda^{t,J}_k\left(\left[\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k})\right] - \left[\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right]\right) \\ &-\lambda^{t,J}_k\left(\left[\mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right]\right) \\ \leq&\lambda_{\max}\left[\left|\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k})\right| + \left|\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right|\right] \\ &+\lambda_{\max}\left[\left|\mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right|\right] \end{aligned} \tag{136}$$

For the third term $\left|\mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right|$, let $D_\theta \triangleq \sup_{\theta_a,\theta_b\in\Omega_\theta}\|\theta_a - \theta_b\|_2 < \infty$ denote the diameter of the parameter space Ω_θ. By combining Assumption 10 and Assumption 13, we obtain the following bound:

$$\begin{aligned} \left|\mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right| &\leq \left|\left\langle\nabla_\theta\mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k}), \theta^{t,J}_k - \theta^{t,*}_k\right\rangle + \frac{L}{2}\|\theta^{t,J}_k - \theta^{t,*}_k\|_2^2\right| \\ &\leq \left(G\|\theta^{t,J}_k - \theta^{t,*}_k\|_2 + \frac{L}{2}\|\theta^{t,J}_k - \theta^{t,*}_k\|_2^2\right) \\ &\leq \left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) \end{aligned} \tag{137}$$

Thus, we obtain:

$$\begin{aligned} &-\lambda^{t,J}_k\left(\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k})\right) \\ \leq&\lambda_{\max}\left[\left|\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k})\right| + \left|\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right|\right] \\ &+\lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) \end{aligned} \tag{138}$$

For the third term $\mathbb{E}_{p^t_k}[\mathcal{R}(f_{\theta^{t,*}_k})] - \mathbb{E}_{p^t_k}[\mathcal{R}(f^{t,*}_{k,\text{cum}})]$, let $\theta^{t,*}_{k,\text{cum}}$ denote the parameter of the cumulative optimal model $f^{t,*}_{k,\text{cum}}$. Under Assumption 2 and Assumption 5, we obtain:

$$\begin{aligned} \mathbb{E}_{p^t_k}[\mathcal{R}(f_{\theta^{t,*}_k})] - \mathbb{E}_{p^t_k}[\mathcal{R}(f^{t,*}_{k,\text{cum}})] =& L^t_{k,\text{task}}(\theta^{t,*}_k) - L^t_{k,\text{task}}(\theta^{t,*}_{k,\text{cum}}) \\ \leq& L_{\text{lip}}\|\theta^{t,*}_k - \theta^{t,*}_{k,\text{cum}}\|_2 \leq L_{\text{lip}}D_\theta \end{aligned} \tag{139}$$

Substituting the above results yields the following upper bound for $\mathbb{E}_{p^t_k}[\mathcal{R}(f_{\theta^{t,J}_k}) - \mathcal{R}(f^{t,*}_{k,\text{cum}})]$:

$$\begin{aligned} &\mathbb{E}_{p^t_k}[\mathcal{R}(f_{\theta^{t,J}_k}) - \mathcal{R}(f^{t,*}_{k,\text{cum}})] \\ \leq&\mathcal{E}^t_{k,\text{bound}}\left(\{\eta^{t,j}_k\}_{j=0}^{J-1}, \|\theta^{t,0}_k - \theta^{t,*}_k\|_2^2\right) + \lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\text{lip}}D_\theta \\ &+\lambda_{\max}\left[\left|\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,J}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,J}_k})\right| + \left|\hat{\mathcal{R}}_{\mathcal{M}^t_k}(f_{\theta^{t,*}_k}) - \mathcal{R}_{p^{\leq t-1}_k}(f_{\theta^{t,*}_k})\right|\right] \end{aligned} \tag{140}$$

Next, consider the term $\Delta_{k,\text{shift}}(t) = \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\text{cum}}^{t-1,*}) - \mathcal{R}(f_{k,\text{cum}}^{t,*})\right]$. Since $f_{k,\text{cum}}^{t-1,*}$ is the optimal model over the historical distribution $p_k^{\leq t-1}$, we have $\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\text{cum}}^{t,*})\right] \geq \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\text{cum}}^{t-1,*})\right]$, which yields:

$$\Delta_{k,\text{shift}}(t) = \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\text{cum}}^{t-1,*}) - \mathcal{R}(f_{k,\text{cum}}^{t,*})\right] \leq \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\text{cum}}^{t-1,*})\right] - \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{k,\text{cum}}^{t-1,*})\right] = 0 \tag{141}$$

Substituting the above results into the recursive bound for the local regret yields:

$$\begin{aligned}
&\text{Regret}_k^{\text{local}}(t; \theta_k^{t,0}, \{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
&\leq \left(1-\omega_k^t\right) \text{Regret}_k^{\text{local}}(t-1; \theta_k^{t-1,0}, \{\eta_k^{t-1,j}\}_{j=0}^{J-1}) + \left(1-\omega_k^t\right) \mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}}) - \mathcal{R}(f_{\theta_k^{t-1,J}})\right] \\
&\quad + \omega_k^t \cdot \mathcal{E}_{k,\text{bound}}^t\left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2\right) + \omega_k^t \cdot \left[\lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\text{lip}} D_\theta\right] \\
&\quad + \omega_k^t \cdot \lambda_{\max}\left[\left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,J}}) - \mathcal{R}_{p_k^{\leq t-1}}(f_{\theta_k^{t,J}})\right| + \left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,*}}) - \mathcal{R}_{p_k^{\leq t-1}}(f_{\theta_k^{t,*}})\right|\right]
\end{aligned} \tag{142}$$

Next, we bound the term $\left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,J}}) - \mathcal{R}_{p_k^{\leq t-1}}(f_{\theta_k^{t,J}})\right| + \left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,*}}) - \mathcal{R}_{p_k^{\leq t-1}}(f_{\theta_k^{t,*}})\right|$. First, consider the function $\left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_\theta) - \mathcal{R}_{p_k^{\leq t-1}}(f_\theta)\right|$. Note that $\mathcal{M}_k^t$ is not sampled i.i.d. from $p_k^{\leq t-1}$, preventing the direct application of standard i.i.d. analysis. To address this, we introduce an auxiliary buffer $\mathcal{M}_{k,\text{rand}}^t$ of capacity M, obtained by class-wise random sampling from $p_k^{\leq t-1}$, where the sampling within each class is asymptotically equivalent to i.i.d. sampling. Defining its empirical risk as $\hat{\mathcal{R}}_{\mathcal{M}_{k,\text{rand}}^t}(f_\theta)$ and leveraging the buffer kernel spectral efficiency advantage constant C_κ from Theorem 2, we obtain:

$$\left|\hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_\theta) - \mathcal{R}_{p_k^{\leq t-1}}(f_\theta)\right| \leq C_\kappa \cdot \left|\mathcal{R}_{p_k^{\leq t-1}}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}_{k,\text{rand}}^t}(f_\theta)\right| \tag{143}$$

For the buffer $\mathcal{M}_{k,\text{rand}}^t$, consider the sub-buffer $\mathcal{M}_{k,c,\text{rand}}^t$ corresponding to class $c \in \mathcal{C}_k^{\leq t-1}$. By Assumption 6, the sample allocation satisfies $|\mathcal{M}_{k,c,\text{rand}}^t| = \Theta(M/|\mathcal{C}_k^{\leq t-1}|)$. Let $\mathcal{H}_K$ be the reproducing kernel Hilbert space induced by the kernel function $K(\cdot,\cdot)$, with the hypothesis function class norm bounded by B. Denote $\hat{R}_{|\mathcal{M}_{k,c,\text{rand}}^t|}(\cdot)$ as the empirical Rademacher complexity. According to Theorem 8 in Bartlett & Mendelson (2002), for any model $f_\theta \in \mathcal{F}_\theta \triangleq \{f \in \mathcal{H}_K : \|f\|_{\mathcal{H}_K} \leq B\}$ and loss function $\ell \in \tilde{\ell}$, the following inequality holds with probability at least $1-\delta$:

$$\mathcal{R}_{p_k^{\leq t-1}(\cdot|c)}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}_{k,c,\text{rand}}^t}(f_\theta) \leq \hat{R}_{M/|\mathcal{C}_k^{\leq t-1}|}(\tilde{\ell} \circ \mathcal{F}_\theta) + \sqrt{\frac{8\log(2/\delta)}{M/|\mathcal{C}_k^{\leq t-1}|}} \tag{144}$$

Furthermore, combining Assumption 5 with Theorem 12 in Bartlett & Mendelson (2002), we have:

$$\hat{R}_{M/|\mathcal{C}_k^{\leq t-1}|}(\tilde{\ell} \circ \mathcal{F}_\theta) \leq 2L_{\text{lip}} \cdot \hat{R}_{M/|\mathcal{C}_k^{\leq t-1}|}(\mathcal{F}_\theta) \tag{145}$$

Thus, we obtain:

$$\begin{aligned}
\left|\mathcal{R}_{p_k^{\leq t-1}(\cdot|c)}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}_{k,c,\text{rand}}^t}(f_\theta)\right| \leq& \hat{R}_{M/|\mathcal{C}_k^{\leq t-1}|}(\tilde{\ell} \circ \mathcal{F}_\theta) + \sqrt{\frac{8\log(2/\delta)}{M/|\mathcal{C}_k^{\leq t-1}|}} \\
\leq& 2L_{\text{lip}} \cdot \hat{R}_{M/|\mathcal{C}_k^{\leq t-1}|}(\mathcal{F}_\theta) + \sqrt{\frac{8\log(2/\delta)}{M/|\mathcal{C}_k^{\leq t-1}|}}
\end{aligned} \tag{146}$$

Further, consider the kernel matrix $\mathbf{K}_c \in \mathbb{R}^{(M/|\mathcal{C}_k^{\leq t-1}|)\times(M/|\mathcal{C}_k^{\leq t-1}|)}$ constructed from the sample set $\{x_{i,c}\}_{i=1}^{M/|\mathcal{C}_k^{\leq t-1}|}$, whose trace is defined as $\text{tr}(\mathbf{K}_c) \triangleq \sum_{i=1}^{M/|\mathcal{C}_k^{\leq t-1}|} K(x_{i,c}, x_{i,c})$. According to

Lemma 22 in Bartlett & Mendelson (2002), the empirical Rademacher complexity of the RKHS function class $\mathcal{F}_\theta$ satisfies:

$$\hat{R}_{M/|\mathcal{C}_k^{\leq t-1}|}(\mathcal{F}_\theta) \leq \frac{2B}{M/|\mathcal{C}_k^{\leq t-1}|}\sqrt{\mathrm{tr}(\mathbf{K}_c)} = \frac{2B}{M/|\mathcal{C}_k^{\leq t-1}|}\sqrt{\sum_{i=1}^{M/|\mathcal{C}_k^{\leq t-1}|} K(x_{i,c}, x_{i,c})} \tag{147}$$

Under the standard kernel normalization condition, $K(x,x) \leq 1$ holds for any input x, which implies $\sum_{i=1}^{M/|\mathcal{C}_k^{\leq t-1}|} K(x_{i,c}, x_{i,c}) \leq M/|\mathcal{C}_k^{\leq t-1}|$. Substituting this into the above inequality yields:

$$\begin{aligned}\left|\mathcal{R}_{p_k^{\leq t-1}(\cdot|c)}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right| \leq & 2L_{\mathrm{lip}} \cdot \frac{2B}{M/|\mathcal{C}_k^{\leq t-1}|}\sqrt{\sum_{i=1}^{M/|\mathcal{C}_k^{\leq t-1}|} K(x_{i,c}, x_{i,c})} + \sqrt{\frac{8\log(2/\delta)}{M/|\mathcal{C}_k^{\leq t-1}|}} \\ \leq & \frac{4L_{\mathrm{lip}}B}{M/|\mathcal{C}_k^{\leq t-1}|}\sqrt{M/|\mathcal{C}_k^{\leq t-1}|} + \sqrt{\frac{8\log(2/\delta)}{M/|\mathcal{C}_k^{\leq t-1}|}} \\ \leq & \left(4L_{\mathrm{lip}}B + \sqrt{8\log(2/\delta)}\right)\sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}}\end{aligned} \tag{148}$$

Consequently, we obtain the upper bound on the risk deviation of the buffer for a single class. Since the global risk is the weighted average of the conditional risks over classes with prior probabilities, let $\pi_k^{\leq t-1}(c)$ denote the prior weight of class c, satisfying $\sum_{c\in\mathcal{C}_k^{\leq t-1}} \pi_k^{\leq t-1}(c) = 1$. The global true risk and empirical risk can be respectively expressed as:

$$\mathcal{R}_{p_k^{\leq t-1}}(f_\theta) = \sum_{c\in\mathcal{C}_k^{\leq t-1}} \pi_k^{\leq t-1}(c)\mathcal{R}_{p_k^{\leq t-1}(\cdot|c)}(f_\theta), \quad \hat{\mathcal{R}}_{\mathcal{M}^t_{k,\mathrm{rand}}}(f_\theta) = \sum_{c\in\mathcal{C}_k^{\leq t-1}} \frac{1}{|\mathcal{C}_k^{\leq t-1}|}\hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta) \tag{149}$$

Applying the triangle inequality yields:

$$\begin{aligned}&\left|\mathcal{R}_{p_k^{\leq t-1}}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}^t_{k,\mathrm{rand}}}(f_\theta)\right| \\ \leq & \left|\sum_{c\in\mathcal{C}_k^{\leq t-1}} \pi_k^{\leq t-1}(c)\left(\mathcal{R}_{p_k^{\leq t-1}(\cdot|c)}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right)\right| \\ & + \left|\sum_{c\in\mathcal{C}_k^{\leq t-1}}\left(\frac{1}{|\mathcal{C}_k^{\leq t-1}|} - \pi_k^{\leq t-1}(c)\right)\hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right| \\ \leq & \sum_{c\in\mathcal{C}_k^{\leq t-1}} \pi_k^{\leq t-1}(c)\left|\mathcal{R}_{p_k^{\leq t-1}(\cdot|c)}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right| \\ & + \left|\sum_{c\in\mathcal{C}_k^{\leq t-1}}\left(\frac{1}{|\mathcal{C}_k^{\leq t-1}|} - \pi_k^{\leq t-1}(c)\right)\hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right| \\ \leq & \left(4L_{\mathrm{lip}}B + \sqrt{8\log(2/\delta)}\right)\sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} + \left|\sum_{c\in\mathcal{C}_k^{\leq t-1}}\left(\frac{1}{|\mathcal{C}_k^{\leq t-1}|} - \pi_k^{\leq t-1}(c)\right)\hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right|\end{aligned} \tag{150}$$

For the term $\left|\sum_{c\in\mathcal{C}_k^{\leq t-1}}\left(\frac{1}{|\mathcal{C}_k^{\leq t-1}|} - \pi_k^{\leq t-1}(c)\right)\hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta)\right|$, noting that $\hat{\mathcal{R}}_{\mathcal{M}^t_{k,c,\mathrm{rand}}}(f_\theta) \leq 1$ due to the boundedness of the loss function in Assumption 5, applying the variational characterization

of the total variation distance over the finite class set $\mathcal{C}_k^{\leq t-1}$ yields:

$$\left| \sum_{c \in \mathcal{C}_k^{\leq t-1}} \left(\frac{1}{|\mathcal{C}_k^{\leq t-1}|} - \pi_k^{\leq t-1}(c) \right) \hat{\mathcal{R}}_{\mathcal{M}_{k,c,\text{rand}}^t}(f_\theta) \right| \leq D_{\text{TV}} \left(\frac{1}{|\mathcal{C}_k^{\leq t-1}|} \middle\| \pi_k^{\leq t-1}(c) \right) \leq 1 \tag{151}$$

Substituting this back into the above inequality yields:

$$\left| \mathcal{R}_{p_k^{\leq t-1}}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}_{k,\text{rand}}^t}(f_\theta) \right| \leq 1 + \left(4L_{\text{lip}} B + \sqrt{8 \log(2/\delta)} \right) \sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} \tag{152}$$

Consequently, we obtain the upper bound for $\left| \hat{\mathcal{R}}_{\mathcal{M}_{k,\text{rand}}^t}(f_\theta) - \mathcal{R}_{p_k^{\leq t-1}}(f_\theta) \right|$. Incorporating the global spectral efficiency advantage constant C_κ from Theorem 2 yields:

$$\begin{aligned} \left| \hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_\theta) - \mathcal{R}_{p_k^{\leq t-1}}(f_\theta) \right| \leq & C_\kappa \cdot \left| \mathcal{R}_{p_k^{\leq t-1}}(f_\theta) - \hat{\mathcal{R}}_{\mathcal{M}_{k,\text{rand}}^t}(f_\theta) \right| \\ \leq & C_\kappa \cdot \left(1 + \left(4L_{\text{lip}} B + \sqrt{8 \log(2/\delta)} \right) \sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} \right) \end{aligned} \tag{153}$$

Finally, we obtain:

$$\begin{aligned} & \left| \hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,J}}) - \mathcal{R}_{p_k^{\leq t-1}}(f_{\theta_k^{t,J}}) \right| + \left| \hat{\mathcal{R}}_{\mathcal{M}_k^t}(f_{\theta_k^{t,*}}) - \mathcal{R}_{p_k^{\leq t-1}}(f_{\theta_k^{t,*}}) \right| \\ \leq & 2C_\kappa \cdot \left(1 + \left(4L_{\text{lip}} B + \sqrt{8 \log(2/\delta)} \right) \sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} \right) \\ = & C_\kappa \cdot \left(2 + \left(8L_{\text{lip}} B + 4\sqrt{2 \log(2/\delta)} \right) \sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} \right) \end{aligned} \tag{154}$$

Substituting this into the recursive bound for the local regret yields:

$$\begin{aligned} & \text{Regret}_k^{\text{local}}(t; \theta_k^{t,0}, \{\eta_k^{t,j}\}_{j=0}^{J-1}) \\ \leq & \left(1 - \omega_k^t\right) \text{Regret}_k^{\text{local}}(t-1; \theta_k^{t-1,0}, \{\eta_k^{t-1,j}\}_{j=0}^{J-1}) + \left(1 - \omega_k^t\right) \mathbb{E}_{p_k^{\leq t-1}} \left[\mathcal{R}(f_{\theta_k^{t,J}}) - \mathcal{R}(f_{\theta_k^{t-1,J}}) \right] \\ & + \omega_k^t \cdot \mathcal{E}_{k,\text{bound}}^t \left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2 \right) + \omega_k^t \cdot \left[\lambda_{\max} \left(G D_\theta + \frac{L}{2} {D_\theta}^2 \right) + L_{\text{lip}} D_\theta \right] \\ & + \omega_k^t \cdot \lambda_{\max} C_\kappa \cdot \left(2 + \left(8L_{\text{lip}} B + 4\sqrt{2 \log(2/\delta)} \right) \sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} \right) \end{aligned} \tag{155}$$

To simplify the notation, we define the composite error function $U_k(t; \theta_k^{t,0}, \{\eta_k^{t,j}\}_{j=0}^{J-1})$ as follows:

$$\begin{aligned} & U_k(t; \theta_k^{t,0}, \{\eta_k^{t,j}\}_{j=0}^{J-1}) \\ \triangleq & \mathcal{E}_{k,\text{bound}}^t \left(\{\eta_k^{t,j}\}_{j=0}^{J-1}, \|\theta_k^{t,0} - \theta_k^{t,*}\|_2^2 \right) + \left[\lambda_{\max} \left(G D_\theta + \frac{L}{2} {D_\theta}^2 \right) + L_{\text{lip}} D_\theta \right] \\ & + \lambda_{\max} C_\kappa \cdot \left(2 + \left(8L_{\text{lip}} B + 4\sqrt{2 \log(2/\delta)} \right) \sqrt{\frac{|\mathcal{C}_k^{\leq t-1}|}{M}} \right) \end{aligned} \tag{156}$$

where $D_\theta \triangleq \sup_{\theta_a, \theta_b \in \Omega_\theta} \|\theta_a - \theta_b\|_2$ is the diameter of the parameter space Ω_θ; B is the upper bound on the norm radius of the hypothesis class $\mathcal{F}_\theta$ in the RKHS; and $\lambda_{\max}$ is the maximum replay

loss weight. Consequently, the recursive bound for the regret can be expressed as:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
\le&\left(1-\omega_k^t\right)\mathrm{Regret}_k^{\mathrm{local}}(t-1;\theta_k^{t-1,0},\{\eta_k^{t-1,j}\}_{j=0}^{J-1})+\omega_k^t\cdot U_k(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
&+\left(1-\omega_k^t\right)\mathbb{E}_{p_k^{\le t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}})-\mathcal{R}(f_{\theta_k^{t-1,J}})\right]
\end{aligned} \tag{157}$$

For the initial local regret $\mathrm{Regret}_k^{\mathrm{local}}(1;\theta_k^{1,0},\{\eta_k^{1,j}\}_{j=0}^{J-1})$ at $t=1$, since the buffer is not yet initialized, the local training reduces to standard SGD optimization under the task loss $L_{k,\text{task}}^1$ only. Under Assumptions 3 and 10-13, the optimization error is bounded by the SGD convergence bound given in Theorem 1 of Bach & Moulines (2011); under Assumptions 4 and 5, the generalization gap between the empirical and expected risks is controlled by the i.i.d. RKHS generalization bound jointly established by Theorems 8, 12, and 22 in Bach & Moulines (2011). Combining these two bounds and applying the L-smoothness in Assumption 10 yields:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{local}}(1;\theta_k^{1,0},\{\eta_k^{1,j}\}_{j=0}^{J-1}) \\
\le&\frac{L}{2}\exp\left(-\mu\sum_{j=0}^{J-1}\eta_k^{1,j}\right)\|\theta_k^{1,0}-\theta_{1,\mathrm{cum}}^{1,*}\|_2^2+\frac{L}{2}\sum_{j=0}^{J-1}(\eta_k^{1,j})^2G^2\prod_{i=j+1}^{J-1}(1-\mu\eta_k^{1,i}) \\
&+\frac{4L_{\mathrm{Lip}}B}{\sqrt{n_k^1}}+\sqrt{\frac{8\log(2/\delta)}{n_k^1}}
\end{aligned} \tag{158}$$

This completes the proof of Lemma 2.

B.6 PROOF OF LEMMA 3

Lemma 3 (Single-step regret bound of the FedKACE global model) *Under standard assumptions (see Appendix B.1), let each client in the t-th FL round initialize with the global model θ_g^{t-1}, yielding the aggregated global model $\theta_g^t = \frac{1}{K}\sum_{k'=1}^K \theta_{k'}^{t,J}$ after J local steps. Meanwhile, define the regret of model θ on client k at FL round t as $\mathrm{Regret}_k^t(\theta) \triangleq \mathbb{E}_{p_k^{\leq t}}\left[\mathcal{R}(f_\theta) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})\right]$, and the global model's regret functional on client k as $\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \triangleq \mathrm{Regret}_k^t(\theta_g^t)$, where $\mathcal{I}_t \triangleq \left\{\{\eta_{k'}^{t,j}\}_{j=0}^{J-1} \mid \forall k' \in \mathcal{K}\right\}$ is the set of local learning rate sequences at FL round t. This functional satisfies the following explicit single-step upper bound:*

$$\begin{aligned}\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \leq & \frac{1}{K}\sum_{k'=1}^K \mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) \\ & + \frac{1}{K}\sum_{k'=1,k'\neq k}^K \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*}) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\ & + \frac{2(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu) - \frac{\mu}{2K}\sum_{k'=1}^K \|\theta_{k'}^{t,J} - \theta_g^t\|_2^2\end{aligned} \tag{159}$$

where:

$$\Delta^{\mathrm{TV}}(t,\nu) \triangleq \frac{1}{(t+1)^\nu}\left(2^\nu + \frac{3}{\nu}\left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right]\right) + \frac{3}{\nu^2} \tag{160}$$

is the global upper bound of the total variation (TV) distance between the cumulative distributions of any pair of clients $\|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}}, \forall k,k' \in \mathcal{K}$ at FL round t; $\mathrm{Regret}_{k'}^{\mathrm{local}}(t;\cdot)$ is the local model regret functional defined in Lemma 1; and $\nu = n_{\min}/n_{\max} \leq 1$, with $n_{max} \triangleq \max_{k'\in\mathcal{K},\tau\leq T}(n_{k'}^\tau)$ and $n_{min} \triangleq \min_{k'\in\mathcal{K},\tau\leq T}(n_{k'}^\tau)$ being the maximum and minimum possible per-round data sizes across all K clients over all T FL rounds, respectively.

Proof:

Since the global model $\theta_g^t = \frac{1}{K}\sum_{k'=1}^K \theta_{k'}^{t,J}$ is a convex combination of the local models, according to Assumption 11 and Jensen's inequality for strongly convex functions, the corresponding risk function satisfies:

$$\begin{aligned}\mathcal{R}(f_{\theta_g^t}) =& \mathcal{R}(f_{\frac{1}{K}\sum_{k'=1}^K \theta_{k'}^{t,J}}) \\ \leq & \sum_{k'=1}^K \frac{1}{K}\mathcal{R}(f_{\theta_{k'}^{t,J}}) - \frac{\mu}{2}\sum_{k'=1}^K \frac{1}{K}\|\theta_{k'}^{t,J} - \frac{1}{K}\sum_{k''=1}^K \theta_{k''}^{t,J}\|_2^2 \\ =& \frac{1}{K}\sum_{k'=1}^K \mathcal{R}(f_{\theta_{k'}^{t,J}}) - \frac{\mu}{2K}\sum_{k'=1}^K \|\theta_{k'}^{t,J} - \theta_g^t\|_2^2\end{aligned} \tag{161}$$

Thus, we obtain:

$$\begin{aligned}\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t})] \leq & \frac{1}{K}\sum_{k'=1}^K \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] - \frac{\mu}{2K}\sum_{k'=1}^K \mathbb{E}_{p_k^{\leq t}}\left[\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2\right] \\ = & \frac{1}{K}\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t})] + \frac{1}{K}\sum_{k'=1,k'\neq k}^K \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] - \frac{\mu}{2K}\sum_{k'=1}^K \|\theta_{k'}^{t,J} - \theta_g^t\|_2^2\end{aligned} \tag{162}$$

For the term $\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})]$ in the first part, Assumption 5 implies that the risk is bounded, i.e., $\mathcal{R}(f_{\theta_{k'}^{t,J}}) \leq 1$. Based on this, we proceed with the following derivation:

$$\begin{aligned}
\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] =& \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] + \left(\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] - \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})]\right) \\
\leq& \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] + \left|\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] - \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})]\right| \\
\leq& \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] + \|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} \cdot \sup\left|\mathcal{R}(f_{\theta_{k'}^{t,J}})\right| \\
=& \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] + \|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}}
\end{aligned} \tag{163}$$

Substituting the risk term of the global model on client k, we obtain:

$$\begin{aligned}
\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t})] \leq& \frac{1}{K}\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t})] + \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\left(\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] + \|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}}\right) \\
&- \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2 \\
=& \frac{1}{K}\sum_{k'=1}^{K}\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] + \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2
\end{aligned} \tag{164}$$

Furthermore, for the expectation term $\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})]$, we have:

$$\begin{aligned}
\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] =& \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{\theta_{k'}^{t,J}})] - \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] + \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] \\
=& \mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) + \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]
\end{aligned} \tag{165}$$

Substituting this into the above equation yields:

$$\begin{aligned}
\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t})] \leq& \frac{1}{K}\sum_{k'=1}^{K}\left(\mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) + \mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]\right) \\
&+ \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2 \\
=& \frac{1}{K}\sum_{k'=1}^{K}\mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) + \frac{1}{K}\sum_{k'=1}^{K}\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] \\
&+ \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2
\end{aligned} \tag{166}$$

Subtracting $\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$ from both sides yields:

$$\begin{aligned}
&\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
\leq &\frac{1}{K}\sum_{k'=1}^{K} \mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) + \frac{1}{K}\sum_{k'=1}^{K}\left(\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right) \\
&+ \frac{1}{K}\sum_{k'=1,k'\neq k}^{K} \|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2 \\
= &\frac{1}{K}\sum_{k'=1}^{K} \mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) + \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\left(\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right) \\
&+ \frac{1}{K}\sum_{k'=1,k'\neq k}^{K} \|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2
\end{aligned} \tag{167}$$

Substituting the definition of the regret of the global model on client k, $\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \triangleq \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{\theta_g^t}) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$, yields:

$$\begin{aligned}
\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \leq &\frac{1}{K}\sum_{k'=1}^{K} \mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) \\
&+ \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\left(\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right) \\
&+ \frac{1}{K}\sum_{k'=1,k'\neq k}^{K} \|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J} - \theta_g^t\|_2^2
\end{aligned} \tag{168}$$

For the term $\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$, we can rewrite it as follows:

$$\begin{aligned}
&\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
= &\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] + \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
= &\left(\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]\right) + \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*}) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]
\end{aligned} \tag{169}$$

For the term $\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]$, following a similar argument as before, we have:

$$\begin{aligned}
\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] \leq &\left|\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]\right| \\
\leq &\|p_{k'}^{\leq t} - p_k^{\leq t}\|_{\mathrm{TV}} \cdot \sup|\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})| \\
\leq &\|p_{k'}^{\leq t} - p_k^{\leq t}\|_{\mathrm{TV}}
\end{aligned} \tag{170}$$

Thus, we obtain:

$$\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})] - \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \leq \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*}) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] + \|p_{k'}^{\leq t} - p_k^{\leq t}\|_{\mathrm{TV}} \tag{171}$$

Substituting the expression for the regret of the global model on client k yields:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t)\\
\leq&\frac{1}{K}\sum_{k'=1}^{K}\mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1})\\
&+\frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\left(\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]+\|p_{k'}^{\leq t}-p_k^{\leq t}\|_{\mathrm{TV}}\right)\\
&+\frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\|p_k^{\leq t}-p_{k'}^{\leq t}\|_{\mathrm{TV}}-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2\\
=&\frac{1}{K}\sum_{k'=1}^{K}\mathrm{Regret}_{k'}^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_{k'}^{t,j}\}_{j=0}^{J-1})+\frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\\
&+\frac{2}{K}\sum_{k'=1,k'\neq k}^{K}\|p_k^{\leq t}-p_{k'}^{\leq t}\|_{\mathrm{TV}}-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2
\end{aligned}\tag{172}$$

For the total variation (TV) distance term between the cumulative distributions, $\|p_k^{\leq t}-p_{k'}^{\leq t}\|_{\mathrm{TV}}$, given that different clients share the same conditional probability density function $p(x|c)$, we define the cumulative sample proportion of class c for client k up to FL round t as $\pi_k^{\leq t}(c)\triangleq\left\{\sum_{\tau=1}^{t}n_k^\tau\mathbb{P}_{p_k^\tau}(Y=c)\right\}/\left\{\sum_{\tau=1}^{t}n_k^\tau\right\}$. Clearly, this proportion forms a probability distribution over the global class set $\mathcal{C}$, satisfying $\sum_{c\in\mathcal{C}}\pi_k^{\leq t}(c)=1$ and $\pi_k^{\leq t}(c)\geq 0$. Consequently, the cumulative distributions of clients k and k' up to FL round t can be decomposed into a mixture of classes:

$$p_k^{\leq t}(x)=\sum_{c\in\mathcal{C}}\pi_k^{\leq t}(c)p(x|c);\quad p_{k'}^{\leq t}(x)=\sum_{c\in\mathcal{C}}\pi_{k'}^{\leq t}(c)p(x|c)\tag{173}$$

By the definition of the total variation distance, we obtain:

$$\begin{aligned}
\|p_k^{\leq t}-p_{k'}^{\leq t}\|_{\mathrm{TV}}=&\frac{1}{2}\int_{\mathcal{X}}\left|\sum_{c\in\mathcal{C}}\left(\pi_k^{\leq t}(c)-\pi_{k'}^{\leq t}(c)\right)p(x|c)\right|\mathrm{d}x\\
\leq&\frac{1}{2}\sum_{c\in\mathcal{C}}\left|\int_{\mathcal{X}}\left(\pi_k^{\leq t}(c)-\pi_{k'}^{\leq t}(c)\right)p(x|c)\mathrm{d}x\right|\\
\leq&\frac{1}{2}\sum_{c\in\mathcal{C}}\left|\pi_k^{\leq t}(c)-\pi_{k'}^{\leq t}(c)\right|\int_{\mathcal{X}}p(x|c)\mathrm{d}x\\
=&\frac{1}{2}\sum_{c\in\mathcal{C}}\left|\pi_k^{\leq t}(c)-\pi_{k'}^{\leq t}(c)\right|
\end{aligned}\tag{174}$$

For the term $\left|\pi_k^{\leq t}(c)-\pi_{k'}^{\leq t}(c)\right|$ corresponding to class c, similar to the proof of Lemma 2, let n_k^τ be the amount of local data received by client k at the τ-th FL round, and define the mixing weight of the t-th FL round data distribution p_k^t as $\omega_k^t\triangleq n_k^t/\sum_{\tau=1}^{t}n_k^\tau$, which satisfies $\sum_{\tau=1}^{t}\omega_k^\tau=1$. Translating the cumulative distribution recurrence $p_k^{\leq t}(x)=(1-\omega_k^t)p_k^{\leq t-1}(x)+\omega_k^t p_k^t(x)$ to class c, we obtain the following recurrence:

$$\pi_k^{\leq t}(c)=(1-\omega_k^t)\pi_k^{\leq t-1}(c)+\omega_k^t\pi_k^t(c);\qquad\pi_{k'}^{\leq t}(c)=(1-\omega_{k'}^t)\pi_{k'}^{\leq t-1}(c)+\omega_{k'}^t\pi_{k'}^t(c)\tag{175}$$

In particular, if client k does not receive any data samples of class c in the t-th round, its sample proportion $\pi_k^t(c) = 0$, in which case the above exactly holds. Consequently, we obtain:

$$\begin{aligned}
&\pi_{\widetilde{k}}^{\leq t}(c) - \pi_{\widetilde{k'}}^{\leq t}(c) \\
=&(1-\omega_k^t)\pi_{\widetilde{k}}^{\leq t-1}(c) + \omega_k^t\pi_k^t(c) - (1-\omega_{k'}^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) - \omega_{k'}^t\pi_{k'}^t(c) \\
=&(1-\omega_k^t)\pi_{\widetilde{k}}^{\leq t-1}(c) - (1-\omega_k^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) + (1-\omega_k^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) - (1-\omega_{k'}^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) \\
&+ \omega_k^t\pi_k^t(c) - \omega_{k'}^t\pi_{k'}^t(c) \\
=&(1-\omega_k^t)(\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k}}^{\leq t-1}(c)) + (\omega_{k'}^t - \omega_k^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) + \omega_k^t\pi_k^t(c) - \omega_{k'}^t\pi_{k'}^t(c) \\
=&(1-\omega_k^t)(\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k}}^{\leq t-1}(c)) + (\omega_{k'}^t - \omega_k^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) \\
&+ \omega_k^t\pi_k^t(c) - \omega_k^t\pi_{k'}^t(c) + \omega_k^t\pi_{k'}^t(c) - \omega_{k'}^t\pi_{k'}^t(c) \\
=&(1-\omega_k^t)(\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k}}^{\leq t-1}(c)) + (\omega_{k'}^t - \omega_k^t)\pi_{\widetilde{k'}}^{\leq t-1}(c) \\
&+ \omega_k^t\left(\pi_k^t(c) - \pi_{k'}^t(c)\right) + (\omega_k^t - \omega_{k'}^t)\pi_{k'}^t(c)
\end{aligned} \tag{176}$$

Taking the absolute value on both sides yields:

$$\begin{aligned}
\left|\pi_{\widetilde{k}}^{\leq t}(c) - \pi_{\widetilde{k'}}^{\leq t}(c)\right| \leq&(1-\omega_k^t)\left|\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k'}}^{\leq t-1}(c)\right| + \left|\omega_{k'}^t - \omega_k^t\right|\pi_{\widetilde{k'}}^{\leq t-1}(c) \\
&+ \omega_k^t\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| + \left|\omega_k^t - \omega_{k'}^t\right|\pi_{k'}^t(c) \\
=&(1-\omega_k^t)\left|\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k'}}^{\leq t-1}(c)\right| + \left|\omega_{k'}^t - \omega_k^t\right|(\pi_{\widetilde{k'}}^{\leq t-1}(c) + \pi_{k'}^t(c)) \\
&+ \omega_k^t\left|\pi_k^t(c) - \pi_{k'}^t(c)\right|
\end{aligned} \tag{177}$$

Summing both sides over all $c \in \mathcal{C}$ yields:

$$\begin{aligned}
&\sum_{c\in\mathcal{C}}\left|\pi_{\widetilde{k}}^{\leq t}(c) - \pi_{\widetilde{k'}}^{\leq t}(c)\right| \\
\leq&(1-\omega_k^t)\sum_{c\in\mathcal{C}}\left|\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k'}}^{\leq t-1}(c)\right| + \left|\omega_{k'}^t - \omega_k^t\right|\sum_{c\in\mathcal{C}}\left(\pi_{\widetilde{k'}}^{\leq t-1}(c) + \pi_{k'}^t(c)\right) \\
&+ \omega_k^t\sum_{c\in\mathcal{C}}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| \\
=&(1-\omega_k^t)\sum_{c\in\mathcal{C}}\left|\pi_{\widetilde{k}}^{\leq t-1}(c) - \pi_{\widetilde{k'}}^{\leq t-1}(c)\right| + 2\left|\omega_{k'}^t - \omega_k^t\right| + \omega_k^t\sum_{c\in\mathcal{C}}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right|
\end{aligned} \tag{178}$$

This yields a recurrence relation for $\sum_{c\in\mathcal{C}}\left|\pi_{\widetilde{k}}^{\leq t}(c) - \pi_{\widetilde{k'}}^{\leq t}(c)\right|$. For the third term $\sum_{c\in\mathcal{C}}|\pi_k^t(c) - \pi_{k'}^t(c)|$ above, we partition the global class set $\mathcal{C}$ into four mutually exclusive subsets based on the classes $\mathcal{C}_k^t$ and $\mathcal{C}_{k'}^t$ observed by clients k and k' in FL round t: $\mathcal{C}_{k\cap k'}^t \triangleq \mathcal{C}_k^t \cap \mathcal{C}_{k'}^t$, $\mathcal{C}_{k,\text{only}}^t \triangleq \mathcal{C}_k^t \setminus \mathcal{C}_{k'}^t$, $\mathcal{C}_{k',\text{only}}^t \triangleq \mathcal{C}_{k'}^t \setminus \mathcal{C}_k^t$, and $\mathcal{C}_{\text{others}}^t \triangleq \mathcal{C} \setminus (\mathcal{C}_k^t \cup \mathcal{C}_{k'}^t)$. Consequently, this summation can be decomposed as:

$$\begin{aligned}
\sum_{c\in\mathcal{C}}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| =& \sum_{c\in\mathcal{C}_{k\cap k'}^t}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| + \sum_{c\in\mathcal{C}_{k,\text{only}}^t}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| \\
&+ \sum_{c\in\mathcal{C}_{k',\text{only}}^t}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| + \sum_{c\in\mathcal{C}_{\text{others}}^t}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| \\
=& \sum_{c\in\mathcal{C}_{k\cap k'}^t}\left|\pi_k^t(c) - \pi_{k'}^t(c)\right| + \sum_{c\in\mathcal{C}_{k,\text{only}}^t}\pi_k^t(c) + \sum_{c\in\mathcal{C}_{k',\text{only}}^t}\pi_{k'}^t(c) + 0
\end{aligned} \tag{179}$$

Substituting the identities $\sum_{c\in\mathcal{C}^t_{k,\text{only}}}\pi^t_k(c) = 1-\sum_{c\in\mathcal{C}^t_{k\cap k'}}\pi^t_k(c)$ and $\sum_{c\in\mathcal{C}^t_{k',\text{only}}}\pi^t_{k'}(c) = 1-\sum_{c\in\mathcal{C}^t_{k\cap k'}}\pi^t_{k'}(c)$ into the above equation yields:

$$\begin{aligned}\sum_{c\in\mathcal{C}}\left|\pi^t_k(c)-\pi^t_{k'}(c)\right| =& \sum_{c\in\mathcal{C}^t_{k\cap k'}}\left|\pi^t_k(c)-\pi^t_{k'}(c)\right| + \left(1-\sum_{c\in\mathcal{C}^t_{k\cap k'}}\pi^t_k(c)\right)+\left(1-\sum_{c\in\mathcal{C}^t_{k\cap k'}}\pi^t_{k'}(c)\right)\\ =&2+\sum_{c\in\mathcal{C}^t_{k\cap k'}}\left(\left|\pi^t_k(c)-\pi^t_{k'}(c)\right|-\pi^t_k(c)-\pi^t_{k'}(c)\right)\\ =&2-\sum_{c\in\mathcal{C}^t_{k\cap k'}}\left(\pi^t_k(c)+\pi^t_{k'}(c)-\left|\pi^t_k(c)-\pi^t_{k'}(c)\right|\right)\\ =&2-\sum_{c\in\mathcal{C}^t_{k\cap k'}}2\min\left(\pi^t_k(c),\pi^t_{k'}(c)\right)=2\left(1-\sum_{c\in\mathcal{C}^t_{k\cap k'}}\min\left(\pi^t_k(c),\pi^t_{k'}(c)\right)\right)\end{aligned} \tag{180}$$

By Assumption 7 and the normalization of probability measures, the infimum of the probability mass for any single class is $1/C_{\max}$, which is achieved when the probability mass is uniformly distributed over the $C_{\max}$ global classes. Therefore, for any class $c\in\mathcal{C}^t_{k\cap k'}$, it holds that $\min\left(\pi^t_k(c),\pi^t_{k'}(c)\right)\geq 1/C_{\max}$. Consequently, we obtain:

$$\sum_{c\in\mathcal{C}}\left|\pi^t_k(c)-\pi^t_{k'}(c)\right|\leq 2\left(1-\left|\mathcal{C}^t_{k\cap k'}\right|\cdot\frac{1}{C_{\max}}\right)=2\left(1-\frac{|\mathcal{C}^t_k\cap\mathcal{C}^t_{k'}|}{C_{\max}}\right) \tag{181}$$

Substituting this back into the original inequality yields:

$$\begin{aligned}\sum_{c\in\mathcal{C}}\left|\pi^{\leq t}_k(c)-\pi^{\leq t}_{k'}(c)\right| \leq&(1-\omega^t_k)\sum_{c\in\mathcal{C}}\left|\pi^{\leq t-1}_k(c)-\pi^{\leq t-1}_{k'}(c)\right|\\ &+2\left|\omega^t_{k'}-\omega^t_k\right|+2\omega^t_k\left(1-\frac{|\mathcal{C}^t_k\cap\mathcal{C}^t_{k'}|}{C_{\max}}\right)\end{aligned} \tag{182}$$

Let $n_{\max}\triangleq\max_{k'\in\mathcal{K},\tau\leq T}(n^\tau_{k'})$ and $n_{\min}\triangleq\min_{k'\in\mathcal{K},\tau\leq T}(n^\tau_{k'})$ be the maximum and minimum per-round data sizes across all K clients over the T FL rounds, respectively, and let $\nu=n_{\min}/n_{\max}\leq 1$. We then obtain the following bounds: $\omega^t_k\leq\frac{n_{\max}}{t\cdot n_{\min}}=\frac{1}{\nu t}$, $|\omega^t_{k'}-\omega^t_k|\leq\omega^t_{k'}+\omega^t_k\leq\frac{2}{\nu t}$, and $1-\omega^t_k\leq 1-\frac{n_{\min}}{t\cdot n_{\max}}=1-\frac{\nu}{t}$. Substituting these bounds back into the previous expression yields:

$$\begin{aligned}\sum_{c\in\mathcal{C}}\left|\pi^{\leq t}_k(c)-\pi^{\leq t}_{k'}(c)\right| \leq&(1-\frac{\nu}{t})\sum_{c\in\mathcal{C}}\left|\pi^{\leq t-1}_k(c)-\pi^{\leq t-1}_{k'}(c)\right|+\frac{4}{\nu t}+\frac{2}{\nu t}\left(1-\frac{|\mathcal{C}^t_k\cap\mathcal{C}^t_{k'}|}{C_{\max}}\right)\\ =&(1-\frac{\nu}{t})\sum_{c\in\mathcal{C}}\left|\pi^{\leq t-1}_k(c)-\pi^{\leq t-1}_{k'}(c)\right|+\frac{2}{\nu t}\left(3-\frac{|\mathcal{C}^t_k\cap\mathcal{C}^t_{k'}|}{C_{\max}}\right)\end{aligned} \tag{183}$$

Let $\Delta_{k,k'}^t \triangleq \sum_{c\in\mathcal{C}} \left|\pi_k^{\leq t}(c) - \pi_{k'}^{\leq t}(c)\right|$ and $\rho_{k,k'}^t \triangleq \left|\mathcal{C}_k^t \cap \mathcal{C}_{k'}^t\right| / C_{\max}$. Recursively expanding the above recurrence yields:

$$\begin{aligned}
\Delta_{k,k'}^t \leq & (1-\frac{\nu}{t})\Delta_{k,k'}^{t-1} + \frac{2}{\nu t}\left(3-\rho_{k,k'}^t\right) \\
\leq & \left(1-\frac{\nu}{t}\right)\left(1-\frac{\nu}{(t-1)}\right)\Delta_{k,k'}^{t-2} + \left(1-\frac{\nu}{t}\right)\frac{2}{\nu(t-1)}\left(3-\rho_{k,k'}^{t-1}\right) + \frac{2}{\nu t}\left(3-\rho_{k,k'}^t\right) \\
\leq & \dots \\
\leq & \left[\prod_{\tau=t-m+1}^{t}\left(1-\frac{\nu}{\tau}\right)\right]\Delta_{k,k'}^{t-m} + \sum_{\tau=t-m+1}^{t}\left[\prod_{i=\tau+1}^{t}\left(1-\frac{\nu}{i}\right)\right]\left(\frac{2}{\nu\tau}\left(3-\rho_{k,k'}^{\tau}\right)\right) \\
\leq & \dots \\
\leq & \left(\prod_{\tau=2}^{t}\left(1-\frac{\nu}{\tau}\right)\right)\Delta_{k,k'}^{1} + \sum_{\tau=2}^{t}\left[\prod_{i=\tau+1}^{t}\left(1-\frac{\nu}{i}\right)\right]\left(\frac{2}{\nu\tau}\left(3-\rho_{k,k'}^{\tau}\right)\right)
\end{aligned} \tag{184}$$

For the product term $\prod_{i=\tau+1}^{t}\left(1-\frac{\nu}{i}\right)$, applying the inequality $1-x \leq \exp(-x)$ and the integral bound $\frac{1}{i} \geq \int_i^{i+1}\frac{1}{x}\mathrm{d}x$ for all $i>0$ yields:

$$\begin{aligned}
\prod_{i=\tau+1}^{t}\left(1-\frac{\nu}{i}\right) \leq & \prod_{i=\tau+1}^{t}\exp\left(-\frac{\nu}{i}\right) = \exp\left(-\nu\sum_{i=\tau+1}^{t}\frac{1}{i}\right) \\
\leq & \exp\left(-\nu\sum_{i=\tau+1}^{t}\int_i^{i+1}\frac{1}{x}\mathrm{d}x\right) = \exp\left(-\nu\int_{\tau+1}^{t+1}\frac{1}{x}\mathrm{d}x\right) \\
= & \exp\left(-\nu\ln\left(\frac{t+1}{\tau+1}\right)\right) = \left(\frac{t+1}{\tau+1}\right)^{-\nu} = \left(\frac{\tau+1}{t+1}\right)^{\nu}
\end{aligned} \tag{185}$$

Substituting the recursive expansion of $\Delta_{k,k'}^t$ and utilizing $\rho_{k,k'}^{\tau} \geq 0$ yields:

$$\begin{aligned}
\Delta_{k,k'}^t \leq & \left(\frac{2}{t+1}\right)^{\nu}\Delta_{k,k'}^1 + \sum_{\tau=2}^{t}\left(\frac{\tau+1}{t+1}\right)^{\nu}\left(\frac{2}{\nu\tau}\left(3-\rho_{k,k'}^{\tau}\right)\right) \\
\leq & \left(\frac{2}{t+1}\right)^{\nu}\Delta_{k,k'}^1 + \frac{6}{\nu}\sum_{\tau=2}^{t}\left(\frac{\tau+1}{t+1}\right)^{\nu}\frac{1}{\tau}
\end{aligned} \tag{186}$$

For the summation term $\sum_{\tau=2}^{t}\left(\frac{\tau+1}{t+1}\right)^{\nu}\frac{1}{\tau}$, consider the function $g(x) = \frac{(x+1)^{\nu}}{(t+1)^{\nu}}\cdot\frac{1}{x}$. Utilizing $\nu \leq 1$ and the generalized Bernoulli's inequality, we obtain:

$$\begin{aligned}
g(x) = & \frac{1}{(t+1)^{\nu}}\cdot\frac{(x+1)^{\nu}}{x} = \frac{1}{(t+1)^{\nu}}x^{\nu-1}\left(1+\frac{1}{x}\right)^{\nu} \\
\leq & \frac{1}{(t+1)^{\nu}}x^{\nu-1}\left(1+\frac{\nu}{x}\right) = \frac{1}{(t+1)^{\nu}}\left[x^{\nu-1}+\nu x^{\nu-2}\right]
\end{aligned} \tag{187}$$

Since $\nu \leq 1$, the function $x^{\nu-1} + \nu x^{\nu-2}$ is strictly decreasing on the interval $[2, t]$. Utilizing this monotonicity for integral bounding, the summation satisfies:

$$\begin{aligned}
\sum_{\tau=2}^{t} g(\tau) \leq & g(2) + \int_2^t \frac{1}{(t+1)^\nu} \left[x^{\nu-1} + \nu x^{\nu-2}\right] \mathrm{d}x \\
= & g(2) + \frac{1}{(t+1)^\nu} \int_2^t x^{\nu-1} \mathrm{d}x + \frac{\nu}{(t+1)^\nu} \int_2^t x^{\nu-2} \mathrm{d}x \\
= & g(2) + \frac{1}{(t+1)^\nu} \left[\frac{x^\nu}{\nu}\right]_2^t + \frac{\nu}{(t+1)^\nu} \left[\frac{x^{\nu-1}}{\nu-1}\right]_2^t \\
= & g(2) + \frac{1}{(t+1)^\nu} \cdot \frac{t^\nu - 2^\nu}{\nu} + \frac{\nu}{(t+1)^\nu} \cdot \frac{t^{\nu-1} - 2^{\nu-1}}{\nu - 1} \\
\leq & g(2) + \frac{t^\nu}{(t+1)^\nu} \cdot \frac{1}{\nu} + \frac{\nu 2^{\nu-1}}{(1-\nu)(t+1)^\nu} \\
\leq & g(2) + \frac{1}{\nu} + \frac{\nu 2^{\nu-1}}{(1-\nu)(t+1)^\nu}
\end{aligned} \tag{188}$$

Furthermore, substituting $g(2) = (3/2)^\nu / 2(t+1)^\nu$ yields:

$$\begin{aligned}
\sum_{\tau=2}^{t} g(\tau) \leq & \frac{1}{\nu} + \frac{\nu 2^{\nu-1}}{(1-\nu)(t+1)^\nu} + \frac{(3/2)^\nu}{2(t+1)^\nu} \\
= & \frac{1}{\nu} + \frac{1}{(t+1)^\nu} \left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right]
\end{aligned} \tag{189}$$

Substituting the recursive expansion of $\Delta_{k,k'}^t$ yields:

$$\begin{aligned}
\Delta_{k,k'}^t \leq & \left(\frac{2}{t+1}\right)^\nu \Delta_{k,k'}^1 + \frac{6}{\nu} \left(\frac{1}{\nu} + \frac{1}{(t+1)^\nu} \left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right]\right) \\
= & \left(\frac{2}{t+1}\right)^\nu \Delta_{k,k'}^1 + \frac{6}{\nu(t+1)^\nu} \left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right] + \frac{6}{\nu^2}
\end{aligned} \tag{190}$$

Thus, we obtain the upper bound on the total variation distance between the cumulative data distributions of clients k and k' at FL round t:

$$\begin{aligned}
\|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}} \leq & \frac{1}{2} \sum_{c \in \mathcal{C}} \left|\pi_k^{\leq t}(c) - \pi_{k'}^{\leq t}(c)\right| \\
\leq & \left(\frac{2}{t+1}\right)^\nu + \frac{3}{\nu(t+1)^\nu} \left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right] + \frac{3}{\nu^2} \\
= & \frac{1}{(t+1)^\nu} \left(2^\nu + \frac{3}{\nu} \left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right]\right) + \frac{3}{\nu^2}
\end{aligned} \tag{191}$$

Thus, we define the global upper bound on the cross-client TV distance at FL round t as:

$$\Delta^{\mathrm{TV}}(t, \nu) \triangleq \frac{1}{(t+1)^\nu} \left(2^\nu + \frac{3}{\nu} \left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right]\right) + \frac{3}{\nu^2} \tag{192}$$

Substituting the expression for the regret of the global model on client k yields:

$$\begin{aligned}
\mathrm{Regret}_k^{\mathrm{global}}(t; \theta_g^{t-1}, \mathcal{I}_t) \leq & \frac{1}{K} \sum_{k'=1}^{K} \mathrm{Regret}_{k'}^{\mathrm{local}}(t; \theta_g^{t-1}, \{\eta_{k'}^{t,j}\}_{j=0}^{J-1}) \\
& + \frac{1}{K} \sum_{k'=1, k' \neq k}^{K} \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*}) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\
& + \frac{2(K-1)}{K} \Delta^{\mathrm{TV}}(t, \nu) - \frac{\mu}{2K} \sum_{k'=1}^{K} \|\theta_{k'}^{t,J} - \theta_g^t\|_2^2
\end{aligned} \tag{193}$$

where $\nu = n_{\min}/n_{\max} \leq 1$, and $n_{\max} \triangleq \max_{k' \in \mathcal{K}, \tau \leq T}(n_{k'}^{\tau})$ and $n_{\min} \triangleq \min_{k' \in \mathcal{K}, \tau \leq T}(n_{k'}^{\tau})$ denote the maximum and minimum per-round data sizes across all clients over the T FL rounds, respectively. This completes the proof of Lemma 3.

B.7 PROOF OF THEOREM 3

Theorem 3 (Regret Gap between Local and Global Models in FedKACE) *Under our problem formulation and standard assumptions (see Appendix B.1), let the optimal model for client k under the local cumulative data distribution $p_k^{\leq t}$ be $f_{k,\mathrm{cum}}^{t,*} \triangleq \arg\min_{\theta\in\Omega_\theta} \mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_\theta)]$, and define $\mathrm{Regret}_k^t(\theta) \triangleq \mathbb{E}_{p_k^{\leq t}}\left[\mathcal{R}(f_\theta) - \mathcal{R}(f_{k,\mathrm{cum}}^{t,*})\right]$ as the regret of model θ on client k at FL round t, the regret gap between the local model $\theta_k^{t,J}$ and the global model θ_g^t satisfies:*

$$\mathop{\mathbb{E}}_{k\in\mathcal{K}}\left[\mathrm{Regret}_k^t(\theta_g^t) - \mathrm{Regret}_k^t(\theta_k^{t,J})\right] \leq \frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu) - \frac{\mu}{2K}\sum_{k=1}^{K}\|\theta_k^{t,J} - \theta_g^t\|_2^2 \tag{194}$$

where:

$$\Delta^{\mathrm{TV}}(t,\nu) \triangleq \frac{1}{(t+1)^\nu}\left(2^\nu + \frac{3}{\nu}\left[\frac{(3/2)^\nu}{2} + \frac{\nu 2^{\nu-1}}{1-\nu}\right]\right) + \frac{3}{\nu^2} \tag{195}$$

is the global upper bound of the total variation (TV) distance between the cumulative distributions of any pair of clients $\|p_k^{\leq t} - p_{k'}^{\leq t}\|_{\mathrm{TV}}, \forall k, k' \in \mathcal{K}$ at FL round t, the derivation of which is provided in Lemma 3 (Supplementary Material B.6); $\nu = n_{\min}/n_{\max}$ is the ratio of the minimum to the maximum per-round data size, defined as $n_{\min} \triangleq \min_{k\in\mathcal{K},\tau\leq T}(n_k^\tau)$ and $n_{\max} \triangleq \max_{k\in\mathcal{K},\tau\leq T}(n_k^\tau)$. For small t, $\Delta^{\mathrm{TV}}(t,\nu)$ dominates, indicating that the local model $\theta_k^{t,J}$ is superior; for large t, the negative model discrepancy term dominates, favoring the global model θ_g^t.

Proof:

Following the proofs of Lemmas 2 and 3, we introduce functionals to uniformly represent the local and global model regrets. Specifically, the local regret functional is defined as $\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \triangleq \mathrm{Regret}_k^t(\theta_k^{t,J})$, where $\theta_k^{t,0}$ and $\{\eta_k^{t,j}\}_{j=0}^{J-1}$ are the initial model and the sequence of local learning rates at FL round t, respectively; meanwhile, the global regret functional is defined as $\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \triangleq \mathrm{Regret}_k^t(\theta_g^t)$, where θ_g^{t-1} and $\mathcal{I}_t \triangleq \left\{\{\eta_{k'}^{t,j}\}_{j=0}^{J-1} \mid \forall k' \in \mathcal{K}\right\}$ are the global model at FL round $t-1$ and the set of learning rate sequences for all clients at FL round t, respectively.

First, by utilizing the recursive bound on the local model regret from Lemma 2 and unrolling it to $\tau = 2$, we obtain:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
\leq&\left(1-\omega_k^t\right)\mathrm{Regret}_k^{\mathrm{local}}(t-1;\theta_k^{t-1,0},\{\eta_k^{t-1,j}\}_{j=0}^{J-1}) + \omega_k^t\cdot U_k(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
&+\left(1-\omega_k^t\right)\mathbb{E}_{p_k^{\leq t-1}}\left[\mathcal{R}(f_{\theta_k^{t,J}}) - \mathcal{R}(f_{\theta_k^{t-1,J}})\right] \\
\leq&\cdots \\
\leq&\left[\prod_{\tau=2}^{t}(1-\omega_k^\tau)\right]\mathbb{E}\left[\mathrm{Regret}_k^{\mathrm{local}}\left(1;\theta_k^{1,0},\{\eta_k^{1,j}\}_{j=0}^{J-1}\right)\right] \\
&+\sum_{\tau=2}^{t}\left[\omega_k^\tau\prod_{s=\tau+1}^{t}(1-\omega_k^s)\right]U_k\left(\tau;\theta_k^{\tau,0},\{\eta_k^{\tau,j}\}_{j=0}^{J-1}\right) \\
&+\sum_{\tau=2}^{t}\left[(1-\omega_k^\tau)\prod_{s=\tau+1}^{t}(1-\omega_k^s)\right]\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}}) - \mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]
\end{aligned} \tag{196}$$

Next, let n_k^τ denote the amount of new data received by client k at FL round τ, and define the cumulative data size over the first τ FL rounds as $N_k^\tau \triangleq \sum_{i=1}^{\tau} n_k^i$. Accordingly, we have $\omega_k^\tau =$

n_k^τ/N_k^τ and $1-\omega_k^\tau = N_k^{\tau-1}/N_k^\tau$. Substituting these relations, we obtain:

$$\begin{aligned}
\prod_{\tau=2}^{t}(1-\omega_k^\tau) &= \frac{N_k^1}{N_k^2}\cdot\frac{N_k^2}{N_k^3}\cdots\frac{N_k^{t-1}}{N_k^t} = \frac{N_k^1}{N_k^t} = \frac{n_k^1}{N_k^t} \\
\omega_k^\tau\prod_{s=\tau+1}^{t}(1-\omega_k^s) &= \frac{n_k^\tau}{N_k^\tau}\cdot\frac{N_k^\tau}{N_k^t} = \frac{n_k^\tau}{N_k^t} \\
(1-\omega_k^\tau)\prod_{s=\tau+1}^{t}(1-\omega_k^s) &= \frac{N_k^{\tau-1}}{N_k^\tau}\cdot\frac{N_k^\tau}{N_k^t} = \frac{N_k^{\tau-1}}{N_k^t}
\end{aligned} \tag{197}$$

Substituting this into the expression for the local model regret, we obtain:

$$\begin{aligned}
\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_k^{t,0},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \leq & \frac{n_k^1}{N_k^t}\mathbb{E}\left[\mathrm{Regret}_k^{\mathrm{local}}\left(1;\theta_k^{1,0},\{\eta_k^{1,j}\}_{j=0}^{J-1}\right)\right] \\
& + \sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}U_k\left(\tau;\theta_k^{\tau,0},\{\eta_k^{\tau,j}\}_{j=0}^{J-1}\right) \\
& + \sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}}) - \mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]
\end{aligned} \tag{198}$$

Following the algorithm procedure, the initial model for the local training of client k at round $\tau \geq 1$ of FL is given by $\theta_k^{\tau,0} = \theta_g^{\tau-1} = \frac{1}{K}\sum_{k'=1}^{K}\theta_{k'}^{\tau-1,J}$. Let $\theta_k^{\tau,*}$ denote the ideal saddle-point model for the local training of client k at round τ. Based on these definitions, we obtain:

$$\begin{aligned}
U_k(\tau;\theta_g^{\tau-1},\{\eta_k^{\tau,j}\}_{j=0}^{J-1}) \triangleq & \mathcal{E}_{k,\text{bound}}^t\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1}, \|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right) \\
& + \left[\lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\mathrm{lip}}D_\theta\right] \\
& + \lambda_{\max}C_\kappa\cdot\left(2+\left(8L_{\mathrm{lip}}B + 4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{|\mathcal{C}_k^{\leq\tau-1}|}{M}}\right)
\end{aligned} \tag{199}$$

Considering this summation term in the local model regret expression and applying the bound $|\mathcal{C}_k^{\leq\tau-1}| \leq C_{\max}$, we obtain:

$$\begin{aligned}
& \sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}U_k(\tau;\theta_g^{\tau-1},\{\eta_k^{\tau,j}\}_{j=0}^{J-1}) \\
\leq & \sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}\mathcal{E}_{k,\text{bound}}^t\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1}, \|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right) \\
& + \sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}\left[\lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\mathrm{lip}}D_\theta\right] \\
& + \sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}\lambda_{\max}C_\kappa\cdot\left(2+\left(8L_{\mathrm{lip}}B + 4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{|\mathcal{C}_k^{\leq\tau-1}|}{M}}\right) \\
\leq & \sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}\mathcal{E}_{k,\text{bound}}^t\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1}, \|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right) \\
& + \lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\mathrm{lip}}D_\theta \\
& + \lambda_{\max}C_\kappa\cdot\left(2+\left(8L_{\mathrm{lip}}B + 4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{C_{\max}}{M}}\right)
\end{aligned} \tag{200}$$

Substituting this into the expression for the local model regret, we obtain:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \\
\leq&\frac{n_k^1}{N_k^t}\mathrm{Regret}_k^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_k^{1,j}\}_{j=0}^{J-1}\right) \\
&+\sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t}\mathcal{E}_{k,\mathrm{bound}}^t\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right) \\
&+\sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}})-\mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]+\lambda_{\max}\left(GD_\theta+\frac{L}{2}{D_\theta}^2\right)+L_{\mathrm{lip}}D_\theta \\
&+\lambda_{\max}C_\kappa\cdot\left(2+\left(8L_{\mathrm{lip}}B+4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{C_{\max}}{M}}\right)
\end{aligned} \tag{201}$$

Next, we bound the summation term $\sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}})-\mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]$. By Assumption 5 and Lemma 1 in Kuznetsov & Mohri (2017), we obtain:

$$\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}})-\mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]\leq\sup_{f\in\mathcal{F}}\left|\mathcal{R}_{p_k^{\leq\tau}}(f)-\mathcal{R}_{p_k^{\leq\tau-1}}(f)\right|\leq\|p_k^{\leq\tau}-p_k^{\leq\tau-1}\|_{\mathrm{TV}} \tag{202}$$

Substituting the expanded form of the cumulative data distribution $p_k^{\leq\tau}=\sum_{i=1}^{\tau}\frac{n_k^i}{N_k^\tau}p_k^i=\frac{n_k^\tau}{N_k^\tau}p_k^\tau+\frac{N_k^{\tau-1}}{N_k^\tau}p_k^{\leq\tau-1}$, and applying the convexity and triangle inequality of the Total Variation (TV) distance, we obtain:

$$\begin{aligned}
\left\|p_k^{\leq\tau}-p_k^{\leq\tau-1}\right\|_{\mathrm{TV}}=&\left\|\frac{n_k^\tau}{N_k^\tau}p_k^\tau+\frac{N_k^{\tau-1}}{N_k^\tau}p_k^{\leq\tau-1}-p_k^{\leq\tau-1}\right\|_{\mathrm{TV}}=\left\|\frac{n_k^\tau}{N_k^\tau}\left(p_k^\tau-p_k^{\leq\tau-1}\right)\right\|_{\mathrm{TV}} \\
=&\frac{n_k^\tau}{N_k^\tau}\left\|p_k^\tau-\sum_{j=1}^{\tau-1}\frac{n_k^j}{N_k^{\tau-1}}p_k^j\right\|_{\mathrm{TV}}=\frac{n_k^\tau}{N_k^\tau}\left\|\sum_{j=1}^{\tau-1}\frac{n_k^j}{N_k^{\tau-1}}\left(p_k^\tau-p_k^j\right)\right\|_{\mathrm{TV}} \\
\leq&\frac{n_k^\tau}{N_k^\tau N_k^{\tau-1}}\sum_{j=1}^{\tau-1}n_k^j\left\|p_k^\tau-p_k^j\right\|_{\mathrm{TV}} \\
\leq&\frac{n_k^\tau}{N_k^\tau N_k^{\tau-1}}\sum_{j=1}^{\tau-1}n_k^j\left(\sum_{i=j+1}^{\tau}\|p_k^i-p_k^{i-1}\|_{\mathrm{TV}}\right)
\end{aligned} \tag{203}$$

Furthermore, according to Assumption 8, we have $D_{\mathrm{TV}}(p_k^i\|p_k^{i-1})=\|p_k^i-p_k^{i-1}\|_{\mathrm{TV}}\leq\frac{C_{\max}}{i^\alpha}$. Substituting this into the above inequality, we obtain:

$$\left\|p_k^{\leq\tau}-p_k^{\leq\tau-1}\right\|_{\mathrm{TV}}\leq\frac{n_k^\tau}{N_k^\tau N_k^{\tau-1}}\sum_{j=1}^{\tau-1}n_k^j\left(\sum_{i=j+1}^{\tau}\frac{C_{\max}}{i^\alpha}\right) \tag{204}$$

It follows that:

$$\begin{aligned}\sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}})-\mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]\leq&\sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\frac{n_k^\tau}{N_k^\tau N_k^{\tau-1}}\sum_{j=1}^{\tau-1}n_k^j\left(\sum_{i=j+1}^{\tau}\frac{C_{\max}}{i^\alpha}\right)\\=&\sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t N_k^\tau}\left(\sum_{j=1}^{\tau-1}n_k^j\sum_{i=j+1}^{\tau}\frac{C_{\max}}{i^\alpha}\right)\\=&\sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^t N_k^\tau}\left(C_{\max}\sum_{i=2}^{\tau}\frac{1}{i^\alpha}\sum_{j=1}^{i-1}n_k^j\right)\\=&\frac{C_{\max}}{N_k^t}\sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^\tau}\sum_{i=2}^{\tau}\frac{N_k^{i-1}}{i^\alpha}\end{aligned}\tag{205}$$

Next, analogously to the proof of Lemma 2, we define the maximum and minimum single-round data sizes across all clients over all T rounds of FL as $n_{\max}\triangleq\max_{k'\in\mathcal{K},\tau\leq T}(n_{k'}^\tau)$ and $n_{\min}\triangleq\min_{k'\in\mathcal{K},\tau\leq T}(n_{k'}^\tau)$, respectively. Letting $\nu=n_{\min}/n_{\max}\leq 1$, we obtain:

$$\begin{aligned}\sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}})-\mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]\leq&\frac{C_{\max}}{N_k^t}\sum_{\tau=2}^{t}\frac{n_k^\tau}{N_k^\tau}\sum_{i=2}^{\tau}\frac{N_k^{i-1}}{i^\alpha}\\\leq&\frac{C_{\max}}{t\cdot n_{\min}}\sum_{\tau=2}^{t}\frac{n_{\max}}{\tau\cdot n_{\min}}\sum_{i=2}^{\tau}\frac{(i-1)\cdot n_{\max}}{i^\alpha}\\\leq&\frac{C_{\max}\cdot n_{\max}}{t\cdot n_{\min}}\sum_{\tau=2}^{t}\frac{n_{\max}}{\tau\cdot n_{\min}}\sum_{i=2}^{\tau}\frac{i}{i^\alpha}\\=&\frac{C_{\max}}{\nu^2}\frac{1}{t}\sum_{\tau=2}^{t}\frac{1}{\tau}\sum_{i=2}^{\tau}i^{1-\alpha}\end{aligned}\tag{206}$$

For the summation term $\frac{1}{t}\sum_{\tau=2}^{t}\frac{1}{\tau}\sum_{i=2}^{\tau}i^{1-\alpha}$, when $\alpha>0$ and $t\geq 2$, we obtain the following piecewise upper bound $\mathcal{A}(t,\alpha)$:

$$\frac{1}{t}\sum_{\tau=2}^{t}\frac{1}{\tau}\sum_{i=2}^{\tau}i^{1-\alpha}\leq\mathcal{A}(t,\alpha)\triangleq\begin{cases}\dfrac{t^{1-\alpha}}{(2-\alpha)^2}+\dfrac{t^{-\alpha}}{1-\alpha}+\dfrac{1}{t}, & 0<\alpha<1\\ 1, & \alpha=1\\ \dfrac{t^{1-\alpha}}{(2-\alpha)^2}+\dfrac{1}{(\alpha-1)t}+\dfrac{1}{t}, & 1<\alpha<2\\ \dfrac{(\ln t)^2+2\ln t}{2t}, & \alpha=2\\ \dfrac{\alpha-1}{(\alpha-2)t}(\ln t+1), & \alpha>2\end{cases}\tag{207}$$

It is easy to see that $\mathcal{A}(t,\alpha)$ is monotonically decreasing in α, monotonically increasing in t when $\alpha\in(0,1)$, and monotonically decreasing in t when $\alpha>1$. Substituting this into the original expression, we obtain:

$$\sum_{\tau=2}^{t}\frac{N_k^{\tau-1}}{N_k^t}\mathbb{E}_{p_k^{\leq\tau-1}}\left[\mathcal{R}(f_{\theta_k^{\tau,J}})-\mathcal{R}(f_{\theta_k^{\tau-1,J}})\right]\leq\frac{C_{\max}}{\nu^2}\mathcal{A}(t,\alpha)\tag{208}$$

Substituting this into the expression for the local model regret, we obtain:

$$\begin{aligned}\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \leq & \frac{1}{\nu t}\mathrm{Regret}_k^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_k^{1,j}\}_{j=0}^{J-1}\right) \\ & + \frac{1}{\nu t}\sum_{\tau=2}^{t}\mathcal{E}_{k,\text{bound}}^t\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right) \\ & + \frac{C_{\max}}{\nu^2}\mathcal{A}(t,\alpha) + \lambda_{\max}\left(GD_\theta + \frac{L}{2}{D_\theta}^2\right) + L_{\mathrm{lip}}D_\theta \\ & + \lambda_{\max}C_\kappa\cdot\left(2+\left(8L_{\mathrm{lip}}B+4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{C_{\max}}{M}}\right)\end{aligned} \tag{209}$$

To simplify the notation, we define the distribution shift functional $\Delta_{\mathrm{shift}}(t,C_{\max},\nu,\alpha)$ and the buffer optimization error functional $\Xi_{\mathrm{buffer}}(C_\kappa,C_{\max},M)$ as follows:

$$\Delta_{\mathrm{shift}}(t,C_{\max},\nu,\alpha) \triangleq \frac{C_{\max}}{\nu^2}\cdot\begin{cases}\dfrac{t^{1-\alpha}}{(2-\alpha)^2}+\dfrac{t^{-\alpha}}{1-\alpha}+\dfrac{1}{t}, & 0<\alpha<1\\ 1, & \alpha=1\\ \dfrac{t^{1-\alpha}}{(2-\alpha)^2}+\dfrac{1}{(\alpha-1)t}+\dfrac{1}{t}, & 1<\alpha<2\\ \dfrac{(\ln t)^2+2\ln t}{2t}, & \alpha=2\\ \dfrac{\alpha-1}{(\alpha-2)t}(\ln t+1), & \alpha>2\end{cases} \tag{210}$$

$$\begin{aligned}\Xi_{\mathrm{buffer}}(C_\kappa,C_{\max},M) \triangleq & L_{\mathrm{lip}}D_\theta + \lambda_{\max}\left(GD_\theta+\frac{L}{2}{D_\theta}^2\right) \\ & + \lambda_{\max}C_\kappa\cdot\left(2+\left(8L_{\mathrm{lip}}B+4\sqrt{2\log(2/\delta)}\right)\sqrt{\frac{C_{\max}}{M}}\right)\end{aligned} \tag{211}$$

Consequently, the expression for the local model regret can be compactly written as:

$$\begin{aligned}\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_k^{t,j}\}_{j=0}^{J-1}) \leq & \frac{1}{\nu t}\mathrm{Regret}_k^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_k^{1,j}\}_{j=0}^{J-1}\right) \\ & + \frac{1}{\nu t}\sum_{\tau=2}^{t}\mathcal{E}_{k,\text{bound}}^t\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right) \\ & + \Delta_{\mathrm{shift}}(t,C_{\max},\nu,\alpha) + \Xi_{\mathrm{buffer}}(C_\kappa,C_{\max},M)\end{aligned} \tag{212}$$

Substituting this into the expression for the global model regret given in Lemma 3, we obtain:

$$\begin{aligned}\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \leq & \frac{1}{K}\sum_{k'=1}^{K}\frac{1}{\nu t}\mathrm{Regret}_{k'}^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_{k'}^{1,j}\}_{j=0}^{J-1}\right) \\ & + \frac{1}{K}\sum_{k'=1}^{K}\frac{1}{\nu t}\sum_{\tau=2}^{t}\mathcal{E}_{k',\text{bound}}^t\left(\{\eta_{k'}^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_{k'}^{\tau,*}\|_2^2\right) \\ & + \Delta_{\mathrm{shift}}(t,C_{\max},\nu,\alpha) + \Xi_{\mathrm{buffer}}(C_\kappa,C_{\max},M) \\ & + \frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})] \\ & + \frac{2(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu) - \frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2\end{aligned} \tag{213}$$

Consider the summation term $\frac{1}{K}\sum_{k'=1,k'\neq k}^{K}\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$. For the summand $\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]$, analogously to the previous proof, we have:

$$\begin{aligned}
&\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})-\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\\
=&\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]-\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\\
=&\left[\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]-\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]\right]+\left[\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k',\mathrm{cum}}^{t,*})]-\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right]\\
&+\left[\mathbb{E}_{p_{k'}^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]-\mathbb{E}_{p_k^{\leq t}}[\mathcal{R}(f_{k,\mathrm{cum}}^{t,*})]\right]\\
\leq&\|p_k^{\leq t}-p_{k'}^{\leq t}\|_{\mathrm{TV}}+0+\|p_{k'}^{\leq t}-p_k^{\leq t}\|_{\mathrm{TV}}\\
\leq&2\cdot\Delta^{\mathrm{TV}}(t,\nu)
\end{aligned}\tag{214}$$

Substituting this back into the expression for the global model regret, we obtain:

$$\begin{aligned}
&\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t)\\
\leq&\frac{1}{K}\sum_{k'=1}^{K}\frac{1}{\nu t}\left[\mathrm{Regret}_{k'}^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_{k'}^{1,j}\}_{j=0}^{J-1}\right)+\sum_{\tau=2}^{t}\mathcal{E}_{k',\mathrm{bound}}^{t}\left(\{\eta_{k'}^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_{k'}^{\tau,*}\|_2^2\right)\right]\\
&+\Delta_{\mathrm{shift}}(t,C_{\max},\nu,\alpha)+\Xi_{\mathrm{buffer}}(C_\kappa,C_{\max},M)\\
&+\frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu)-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2
\end{aligned}\tag{215}$$

Next, we discuss the evolution of the local and global regrets with respect to t. We define the regret difference function $\Delta_{k,\mathrm{regret}}(t)$ as follows:

$$\begin{aligned}
&\Delta_{k,\mathrm{regret}}(t)\\
=&\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t)-\mathrm{Regret}_k^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_k^{t,j}\}_{j=0}^{J-1})\\
=&\frac{1}{K}\sum_{k'=1}^{K}\frac{1}{\nu t}\left[\mathrm{Regret}_{k'}^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_{k'}^{1,j}\}_{j=0}^{J-1}\right)+\sum_{\tau=2}^{t}\mathcal{E}_{k',\mathrm{bound}}^{t}\left(\{\eta_{k'}^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_{k'}^{\tau,*}\|_2^2\right)\right]\\
&-\frac{1}{\nu t}\left[\mathrm{Regret}_k^{\mathrm{local}}\left(1;\theta_g^0,\{\eta_k^{1,j}\}_{j=0}^{J-1}\right)+\sum_{\tau=2}^{t}\mathcal{E}_{k,\mathrm{bound}}^{t}\left(\{\eta_k^{\tau,j}\}_{j=0}^{J-1},\|\theta_g^{\tau-1}-\theta_k^{\tau,*}\|_2^2\right)\right]\\
&+\frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu)-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2\\
\triangleq&\Delta_{k,\mathrm{opt}}(\mathcal{K},\nu,t)+\frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu)-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2
\end{aligned}\tag{216}$$

where $\Delta_{k,\mathrm{opt}}(\mathcal{K},\nu,t)$ captures the fluctuations induced by stochastic optimization. It is easy to proof that $\mathbb{E}_{k\in\mathcal{K}}\left[\Delta_{k,\mathrm{opt}}(\mathcal{K},\nu,t)\right]=0$, which yields:

$$\begin{aligned}
\mathbb{E}_{k\in\mathcal{K}}\left[\Delta_{k,\mathrm{regret}}(t)\right]&=\mathbb{E}_{k\in\mathcal{K}}\left[\Delta_{k,\mathrm{opt}}(\mathcal{K},\nu,t)\right]+\frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu)-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2\\
&=\frac{4(K-1)}{K}\Delta^{\mathrm{TV}}(t,\nu)-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2
\end{aligned}\tag{217}$$

where:

$$\Delta^{\mathrm{TV}}(t,\nu)\triangleq\frac{1}{(t+1)^\nu}\left(2^\nu+\frac{3}{\nu}\left[\frac{(3/2)^\nu}{2}+\frac{\nu 2^{\nu-1}}{1-\nu}\right]\right)+\frac{3}{\nu^2}\tag{218}$$

When t is small, $\mathbb{E}_{k\in\mathcal{K}}\left[\Delta_{k,\mathrm{regret}}(t)\right]$ is dominated by $\Delta^{\mathrm{TV}}(t,\nu)$, meaning that the negative impact caused by the cross-client data distribution discrepancy far outweighs the parameter divergence penalty brought by model aggregation. This yields $\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \geq \mathrm{Regret}_k^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_k^{t,j}\}_{j=0}^{J-1})$, indicating that the local model $\theta_k^{t,J}$ outperforms the global model θ_g^t on client k.

When t is large, $\Delta^{\mathrm{TV}}(t,\nu)$ decays significantly and converges to the constant $3/\nu^2$. At this stage, due to the inconsistent buffer data distributions across clients, the local models inevitably overfit to their respective local buffers, resulting in an irremovable steady-state model divergence (client drift). Consequently, the parameter divergence penalty from model aggregation, $-\frac{\mu}{2K}\sum_{k'=1}^{K}\|\theta_{k'}^{t,J}-\theta_g^t\|_2^2$, dominates in magnitude. This yields $\mathrm{Regret}_k^{\mathrm{global}}(t;\theta_g^{t-1},\mathcal{I}_t) \leq \mathrm{Regret}_k^{\mathrm{local}}(t;\theta_g^{t-1},\{\eta_k^{t,j}\}_{j=0}^{J-1})$, indicating that the global model θ_g^t outperforms the local model $\theta_k^{t,J}$ on client k.

This completes the proof of Theorem 3.

C LIMITATIONS

Given buffer size M and D new samples per FL round, HBMS requires $O(M(M+D))$ computation. While acceptable for small M, this becomes prohibitive as M grows. For large M, where random sampling already provides sufficient representativeness and the marginal benefit of explicit optimization diminishes, we suggest reducing the complexity to $O(r(M+D))$ by randomly sampling $r < M$ landmark samples from the buffer, trading theoretical guarantees for efficiency.

Our method and theoretical analysis adopts uniform FedAvg aggregation to maintain analytical tractability, without investigating alternative aggregation strategies such as weighted or personalized aggregation. While this simplification precludes exploring potentially superior aggregation schemes, the adopted uniform aggregation remains a near-optimal choice under the constraint of zero additional communication overhead: more sophisticated aggregation strategies typically necessitate extra client–server interactions or auxiliary information exchange.

D AI ASSISTANCE

We used Large Language Models (LLMs) to assist with language editing and polishing of this manuscript.