FEDERATED UNLEARNING: A PERSPECTIVE OF STABILITY AND FAIRNESS

Jiaqi Shao[1,2], **Tao Lin**[3], **Xuanyu Cao**[1], **Bing Luo**[2] *
[1]The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]Duke Kunshan University, Suzhou, China
[3]Westlake University, Hangzhou, China

ABSTRACT

This paper explores the multifaceted consequences of federated unlearning (FU) with data heterogeneity. We introduce key metrics for FU assessment, concentrating on verification, global stability, and local fairness, and investigate the inherent trade-offs. Furthermore, we formulate the unlearning process with data heterogeneity through an optimization framework. Our key contribution lies in a comprehensive theoretical analysis of the trade-offs in FU and provides insights into data heterogeneity's impacts on FU. Leveraging these insights, we propose FU mechanisms to manage the trade-offs, guiding further development for FU mechanisms.

1 INTRODUCTION

With the advancement of user data regulations, such as GDPR (Regulation, 2018) and CCPA (Goldman, 2020), the concept of "*the right to be forgotten*" has gained prominence. It necessitates models' capability to forget or remove specific training data upon users' request, which is non-trivial since the models potentially memorize training data. Intuitively, the most straightforward approach is to *exactly retrain* the model from scratch without the data to be forgotten. However, this method is computationally expensive, especially for large-scale models. As a result, the machine unlearning (MU) paradigm is proposed to efficiently remove data influences from models (Cao & Yang, 2015). The effectiveness of unlearning, measured by *verification* approaches, requires the unlearning mechanism to closely replicate the results of exact retraining without heavy computational burden.

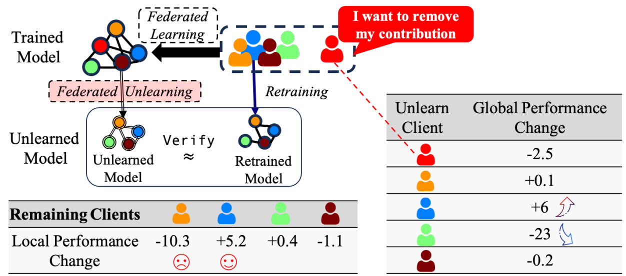
Federated learning (FL) has gained attention in academia and industry with increased data privacy concerns by allowing distributed clients to collaboratively train a model while keeping the data local (Kairouz et al., 2021). While MU offers strategies for traditional centralized machine learning context, federated unlearning (FU) introduces new challenges due to inherent data heterogeneity and privacy concerns in the federated context (Wang et al., 2023). Recent research of FU, such as Gao et al. (2022); Che et al. (2023); Pan et al. (2023a); Liu et al. (2023), mainly focused on verification and efficiency in FU, aligning with the main objectives of MU. However, the inherent data heterogeneity in federated systems introduces new challenges: (i) due to clients' diverse preferences for the global model, unlearning certain clients could result in *unequal impacts on individuals*; (ii) different clients contribute differently to the global model, thus unlearning specific clients can lead to *diverse impacts on model performance*.

On the challenges of FU under heterogeneous data. Figure 1 elaborates two key insights into the challenges posed by data heterogeneity in FU. (1) Local Fairness: As shown in Figure 1(a), FU can unequally impact remaining clients, where some clients benefit from unlearning, but others experience disadvantages. It illustrates a "*local fairness*" concern, pertaining to the uniform distribution of utility changes[1] among remaining clients after unlearning. (2) Global Stability: Unlearning different clients leads to different impacts on the global model's performance, as depicted in Figure 1(b). This highlights the *"global stability"* concern in FU, emphasizing the need to maintain consistent system performance.

FU undertakes the potential for problematic unbounded instability and unfairness. Without controlling the instability or unfairness, unlearning may potentially drive all remaining clients to leave the system, leading to catastrophic forgetting Liu et al. (2022a) and problematic behavior among selfish clients that exploit the resources of others, akin to free-rider attacks Fraboni et al. (2021). Thus, it is essential to consider the two *trade-off* inherent in FU, including the *FU verification vs. global stability* trade-off, as well as the *FU verification vs. local fairness* trade-off. These insights motivate us to ask:

*Corresponding author: Bing Luo, {bing.luo}@dukekunshan.edu.cn

[1]In this context, 'utility' refers to an individual client's experienced performance of the global model.



(a) *Local Fairness*: Uneven Performance Changes on Local Test Dataset.

(b) *Global Stability*: Unlearning Different Clients Varys Global Model Performance.

Figure 1: **Federated Unlearning and Its Consequences.** (a) We consider an FL system of 5 clients with non-IID training data. One client requests to be unlearned. (b) We conduct 5 experiments, each featuring the unlearning of one specific client.

Q: *How can we assess the consequences of FU, and what are the theoretical and practical approaches to balancing the inherent trade-offs?*

To address the above question, we will construct a comprehensive theoretical framework for FU's consequences, which should offer a rigorous understanding of how *data heterogeneity* affects FU while balancing the trade-offs.

Contributions. We summarize our key contributions as follows:

1. **Quantitative Understanding of FU Metrics**: We introduce robust quantitative metrics for FU assessment, including FU *verification metric*, *global stability metric* and *local fairness metric* (in Section 4). These metrics provide comprehensive evaluations for trade-offs in FU.
2. **Theoretical Analysis of FU Trade-offs**: We present a theoretical analysis of the *trade-offs* in FU (Section 5). Under data heterogeneity, our results demonstrate challenges in balancing between FU verification and global stability, as well as between FU verification and local fairness.
3. **FU Mechanism and Theoretical Framework**: We propose a novel FU mechanism based on the theoretical framework, encompassing optimization strategies and penalty methods. We provide theoretical analysis and practical insights into balancing the tradeoffs, detailed in Appendix B. Furthermore, we empirically validate our FU mechanisms in non-convex settings, confirming theoretical insights by effectively balancing trade-offs in Section 6.

2 Related Work

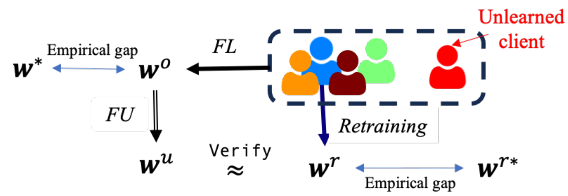
Mechine Unlearning & Federated Unlearning. Machine Unlearning (MU) targets efficient data removal from models (Guo et al., 2020; Wu et al., 2020b), while Federated Unlearning (FU) concerns "right to be forgotten" in FL with stringent privacy constraints. However, recent FU approaches often overlook FL's critical challenge of data heterogeneity (Liu et al., 2022b; Wu et al., 2022; Liu et al., 2020; Zhang et al., 2023). Our work bridges this gap by incorporating data heterogeneity into FU through an optimization framework, presenting theoretical insights on its implications.
Moreover, existing FU methods prioritize verifiable unlearning (Liu et al., 2021; 2022b), but the impacts or consequences of data heterogeneity on FU are underexplored. Our analysis extends beyond verification to examine the consequences, including fairness and stability in FU, underlining the importance of balancing trade-offs in FU.

Stability and Fairness in FL and FU. Stability in FL concerns consistent model performance to defend against external threats. (Yin et al., 2018; Fang et al., 2020; Li et al., 2021). In FU, it involves adjusting to internal changes due to unlearning. Our research extends to analyzing stability in the context of FU, particularly under data heterogeneity.

Fairness in FL ranges from proportional fairness (Wang et al., 2020; Yu et al., 2020) to safeguarding specific attributes (Gu et al., 2022) and ensuring uniform performance across clients (Li et al., 2019; Mohri et al., 2019). In FU, we explore the fairness implications of unlearning, specifically how it affects utility variance among remaining clients under data heterogeneity.

3 Preliminaries: Federated Learning (FL) & Unlearning (FU)

Federated Learning: Suppose there is a client set $\mathcal{N}$ ($|\mathcal{N}| = N$), contributing to FL training. Each client $i \in \mathcal{N}$ has a local training dataset with size n_i, and the data is non-IID across different clients. The optimal FL model is defined below:



FL Models	FU Models
$\mathbf{w}^*$: Optimal Global Model $\mathbf{w}^o$: Empirically Trained Model	$\mathbf{w}^{r*}$: Optimal Unlearned Model $\mathbf{w}^r$: Retrained Model $\mathbf{w}^u$: Unlearned Model

Figure 2: Key Model Notations ($\boldsymbol{w}^u$ represents the unlearned model from the FU mechanism; $\boldsymbol{w}^r$ is from retraining with remaining clients, the special case of $\boldsymbol{w}^u$).

Definition 3.1 (Optimal FL Model)**.** The optimal solution of the FL global model, denoted as $\boldsymbol{w}^*$, can be expressed by the following optimization problem:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}}[F(\boldsymbol{w}) := \textstyle\sum_{i\in\mathcal{N}} p_i f_i(\boldsymbol{w})]$$

Here, f_i denotes the local objective function of client i with the aggregation weight $p_i = \frac{n_i}{\sum_{k\in\mathcal{N}} n_k}$.

The *empirically trained* model $\boldsymbol{w}^o$ is obtained by training all clients $i \in \mathcal{N}$, serving as the empirical approximation of the theoretical optimum $\boldsymbol{w}^*$ (summarized in Figure 2).

Federated Unlearning: FU mechanisms aim to obtain an unlearned model $\boldsymbol{w}^u$ that removes the influence of client set $\mathcal{J}$ who requests to be forgotten from the original trained model $\boldsymbol{w}^o$. This paper focuses on the FU mechanism that begins with $\boldsymbol{w}^o$ and iteratively updates it through the participation of the remaining clients ($i \notin \mathcal{J}$). The model is refined over T rounds akin to FL, converging to the unlearned model $\boldsymbol{w}^u$. The optimal unlearned model could be defined as:

Definition 3.2 (Optimal FU Model)**.** In the context of FU, the optimal unlearned model is defined as the minimizer for the global objective $F_{-\mathcal{J}}$ of all remaining clients $i \notin \mathcal{J}$. It is expressed as:

$$\boldsymbol{w}^{r*} = \arg\min_{\boldsymbol{w}}[F_{-\mathcal{J}}(\boldsymbol{w}) := \textstyle\sum_{i\notin\mathcal{J}} p'_i f_i(\boldsymbol{w})],$$

Here, p'_i is the normalized aggregation weight during unlearning, *i.e.*, $p'_i = \frac{p_i}{1-P_{\mathcal{J}}}$ with $P_{\mathcal{J}} = \sum_{j\in\mathcal{J}} p_j$. The *exact retrained* model $\boldsymbol{w}^r$ is obtained after retraining all remaining clients[2], serving as the empirical approximation of the theoretical optimum $\boldsymbol{w}^{r*}$ (summarized in Figure 2).

4 FU Metrics

This section introduces quantitative metrics for evaluating FU mechanisms. In Section 4.1, we elaborate on the verification metric to *verify the effectiveness* of the FU process. Section 4.2 assesses FU's impact on the system's global stability, quantifying how FU *alters the global performance* of the model. Additionally, Section 4.3 evaluates FU's impact on local fairness, capturing how FU *unequally impacts individuals* in remaining clients. These metrics lay the foundation for our comprehensive framework that captures the inherent trade-offs in FU.

4.1 FU Verification

Verification is critical to evaluate how much the unlearning mechanism effectively removes the data (Yang & Zhao, 2023). Previous studies on unlearning verification often utilize a weak reference like $\boldsymbol{w}^r$ from retraining for comparison (Halimi et al., 2022; Che et al., 2023), but this reference is often inconsistent for variability and randomness in the retraining process. Thus, we employ the optimal unlearned model $\boldsymbol{w}^{r*}$ as a theoretical benchmark for verification, allowing for consistent and replicable evaluations in FU. We define a verification metric by the performance gap between the unlearned model and $\boldsymbol{w}^{r*}$:

Definition 4.1 (FU Verification Metric, V)**.** Consider an FU mechanism $\mathcal{M}$ designed to remove specific clients' influence, resulting in the *unlearned model* $\boldsymbol{w}^u$. The unlearning verification metric $V(\boldsymbol{w}^u)$ quantifies the effectiveness of $\mathcal{M}$ and is defined as:

$$V(\boldsymbol{w}^u) = \mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^u)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}), \tag{1}$$

where $F_{-\mathcal{J}}(\cdot)$ measures performance over the *remaining clients* after unlearning clients set $\mathcal{J}$.

[2]In this paper, the retraining employs FedAvg (McMahan et al., 2017).

4.2 Global Stability

Effective unlearning often requires modification to the trained model, which can lead to performance variations of the global model. In FU, maintaining stable performance is particularly challenging under data heterogeneity, as unlearning critical clients can significantly compromise model performance. To measure the extent of performance stability in FU, we propose the global stability metric as follows:

Definition 4.2 (Global Stability Metric, S)**.** Given an FU mechanism $\mathcal{M}$ and its resulting unlearned model $\boldsymbol{w}^u$. The metric $S(\boldsymbol{w}^u)$ evaluates the global stability of $\mathcal{M}$, measuring the performance gap between the unlearned model $\boldsymbol{w}^u$ and the optimal original FL model $\boldsymbol{w}^*$:

$$S(\boldsymbol{w}^u) = \mathbb{E}\left[F(\boldsymbol{w}^u)\right] - F(\boldsymbol{w}^*). \tag{2}$$

This metric S evaluates the stability of the FU process, facilitating understanding of theoretical analysis in Section 5.1.

4.3 Local Fairness

In FL, fairness can be associated with the consistency of model performance across different clients. Specifically, a model is considered fairer if its performance has a smaller variance across clients (Li et al., 2020a). For FU, effectively unlearning certain clients can unequally impact remaining clients because they have diverse preferences for the global model (data heterogeneity). As demonstrated in Section 1, some clients experience significant utility degradation after unlearning, potentially prompting their departure and further degrading system performance. To measure this FU impact, we propose the local fairness metric as follows:

Definition 4.3 (Local Fairness Metric, Q)**.** Given an FU mechanism $\mathcal{M}$ and its resulting unlearned model $\boldsymbol{w}^u$. The metric $Q(\boldsymbol{w}^u)$ evaluates local fairness of $\mathcal{M}$, assessing the unequal impact of FU on remaining clients:

$$Q(\boldsymbol{w}^u) = \textstyle\sum_{i \notin \mathcal{J}} p_i' \left|\Delta f_i(\boldsymbol{w}^u) - \overline{\Delta f}\right|, \tag{3}$$

where $\Delta f_i(\boldsymbol{w}^u) = \mathbb{E}\left[f_i(\boldsymbol{w}^u)\right] - f_i(\boldsymbol{w}^*)$ represents the utility change for remaining client $i \notin \mathcal{J}$ due to FU. Moreover, $\overline{\Delta f} = \sum_{i \notin \mathcal{J}} p_i' \Delta f_i(\boldsymbol{w}^u)$ is the weighted average of local utility changes among remaining clients, serving as a benchmark for assessing deviation in the impacts of unlearning.

The metric Q is inspired by the mean absolute deviation (MAD) in measuring fairness of FL (Ezzeldin et al., 2023). The metric Q captures FU's impact on utility changes experienced by remaining clients and further facilitates theoretical analysis of fairness implication in Section 5.2.

5 Theoretical Analysis on Trade-offs in FU

This section provides a theoretical analysis of the trade-offs in FU, particularly focusing on the balance between FU verification and stability, as well as FU verification and fairness, as outlined in Section 5.1 and 5.2, respectively. Our analysis critically examines the challenges posed by data heterogeneity in FU. To begin, we formally state the assumptions required for the theoretical analysis.

Assumption 5.1 (Data Heterogeneity in FL)**.** Given a subset of remaining clients $\mathcal{S}$, the data heterogeneity among remaining clients can be quantified as follows:

$$\mathbb{E}_{i \in S} \left\|\nabla f_i(\boldsymbol{w}) - \nabla F_{-\mathcal{J}}(\boldsymbol{w})\right\|^2 \leq \zeta_{\mathcal{S}}^2 + \beta_{\mathcal{S}}^2 \|\nabla F_{-\mathcal{J}}(\boldsymbol{w})\|^2, \tag{4}$$

where $\zeta_{\mathcal{S}}^2$ and $\beta_{\mathcal{S}}^2$ are parameters quantifying the heterogeneity. Here, f_i represents the objective function of client i in subset $\mathcal{S}$, and $F_{-\mathcal{J}}$ is the global objective function of the remaining clients.

Assumption 5.1 assumes the data heterogeneity with parameters $\zeta_{\mathcal{S}}, \beta_{\mathcal{S}}$, representing the degrees of data heterogeneity within the selected subset $\mathcal{S}$ of remaining clients, aligning closely with the framework presented by Wang et al. (2021).

Assumption 5.2 (μ-strong Convexity)**.** Assume that local objective functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are all μ-strong convex. For any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, f_i satisfies the following inequality: $f_i(\mathbf{u}) \geq f_i(\mathbf{v}) + \langle \nabla f_i(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{\mu}{2}\|\mathbf{u} - \mathbf{v}\|^2$, where $\mu > 0$ is the convexity constant.

Assumption 5.3 (L-smoothness)**.** Assume that local objective functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are all L-smooth. For any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, f_i satisfies the following inequality: $f_i(\mathbf{u}) \leq f_i(\mathbf{v}) + \langle \nabla f_i(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{L}{2}\|\mathbf{u} - \mathbf{v}\|^2$, where $L > 0$ is the Lipschitz constant of the gradient of f_i.

Assumption 5.4 (Bounded Variance)**.** Let ξ_t^k be sampled from the k-th client's local data uniformly at random. The variance of stochastic gradients in each client is bounded at round t: $\mathbb{E}\left\|\boldsymbol{g}_i(\boldsymbol{w}) - \nabla f_k\left(\boldsymbol{w}_t\right)\right\|^2 \leq \sigma_{k,t}^2$ for $k = 1, \cdots, N$, where $\boldsymbol{g}_i(\boldsymbol{w}) = \nabla f_k\left(\mathbf{w}_t, \xi_t\right)$.

Assumption 5.5 (Unlearning Clients' Influence)**.** Let $P_{\mathcal{J}}$ denote the total aggregation weights of the clients in set $\mathcal{J}$ within FL, defined as $P_{\mathcal{J}} = \sum_{j\in\mathcal{J}} p_j$. For a client set $\mathcal{J}$ required for unlearning, we assume that $P_{\mathcal{J}} \leq \frac{1}{2}$.

Assumption 5.1-5.4 are commonly used for the FL convergence analysis (Li et al., 2020b; Wang et al., 2021). Assumption 5.5 assumes unlearned clients' aggregate weights do not exceed those of the remaining clients. This is crucial to prevent catastrophic consequences, which could undermine the objectives of fairness and stability in FU.

5.1 Trade-off between FU Verification and Stability

This section explores the trade-off between FU verification and global stability via the lower bound derived for verification (Lemma 5.6) and stability (Lemma 5.8). Then, we formalize the trade-off characterized via the lower bounds in Theorem 5.10. We provide all proofs for lemmas and theorems in Appendix D-H.

Lemma 5.6. *Under Assumptions 5.1-5.4 and given the number of unlearning rounds T and the learning rate $\eta = \frac{1}{T\sqrt{\mu}}\sqrt{\frac{\beta_{\mathcal{S}}-1}{\min\{\mu(\beta_{\mathcal{S}}-1), L(\beta_{\mathcal{S}}-1)\}}}$, the verification metric $V(\boldsymbol{w}^u) = F_{-\mathcal{J}}(\boldsymbol{w}^u) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})$ is lower bounded by:*

$$C_1 = \left(1 + \frac{\beta_{\mathcal{S}}^2 - 1}{T}\right)\left(\Delta F_{-\mathcal{J}}\left(\boldsymbol{w}^*, \boldsymbol{w}^{r*}\right) \ + \Delta F_{-\mathcal{J}}\left(\boldsymbol{w}^o, \boldsymbol{w}^*\right)\right) + \frac{1}{2LT}\left(\bar{\sigma^2} + \bar{\zeta^2}\right),$$

where $\Delta F_{-\mathcal{J}}(\circ, \bullet) = F_{-\mathcal{J}}(\circ) - F_{-\mathcal{J}}(\bullet)$, $\bar{\sigma^2} = \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2$, and $\bar{\zeta^2} = \frac{1}{T}\sum_{t=1}^{T}\zeta_t^2$. Parameters $\beta_{\mathcal{S}}$ and $\bar{\zeta^2}$ characterize the data heterogeneity of remaining clients.

Remark 5.7. The effectiveness of FU, as measured by $V(\boldsymbol{w}^u)$, is hindered by its lower bound C_1 in Equation (5) with several factors:

- *Computational Complexity*: More unlearning rounds T typically indicate convergence towards the optimal unlearned model $\boldsymbol{w}^{r*}$, characterized by a tighter lower bound C_1. However, the computational complexity grows as the number of unlearning rounds T increases.
- *Data Heterogeneity among Remaining Clients:* A high data heterogeneity ($\beta_{\mathcal{S}}^2, \bar{\zeta^2}^2$) among remaining clients can amplify C_1. Therefore, under unlearning rounds T, the more heterogeneous among remaining clients, the more challenging it is to achieve effective unlearning by the increased lower bound C_1 of V.
- *Data Heterogeneity Between Remaining and Unlearned Clients:* The discrepancy $\Delta F_{-\mathcal{J}}\left(\boldsymbol{w}^*, \boldsymbol{w}^{r*}\right)$ implies data heterogeneity between remaining and unlearned clients. A high *heterogeneity* enlarges C_1, thereby potentially compromising FU verification V. Conversely, suppose the data is *homogeneous* between these two groups, C_1 can be diminished, as removing homogeneous data does not significantly alter the overall data distribution ($\Delta F_{-\mathcal{J}}\left(\boldsymbol{w}^*, \boldsymbol{w}^{r*}\right) \approx 0$).

Lemma 5.8. *Under Assumptions 5.2, 5.5 and consider $T \geq \frac{\mu}{\eta^2}$ unlearning rounds. The global stability metric $S(\boldsymbol{w}^u) = \mathbb{E}\left[F(\boldsymbol{w}^u)\right] - F(\boldsymbol{w}^*)$ is bounded below by C_2:*

$$C_2 = \frac{P_{\mathcal{J}}\eta T}{2}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2 + \delta, \tag{5}$$

where $\delta = F(\boldsymbol{w}^o) - F(\boldsymbol{w}^*)$ represents the empirical risk minimization (ERM) gap in the original FL.

Remark 5.9. Maintaining global stability $S(\boldsymbol{w}^u)$ poses challenges due to the lower bound C_2 established in Equation (5), which is influenced by the following factors:

- *Unlearned Clients' Influence:* The higher aggregation weight of unlearned clients $P_{\mathcal{J}}$ implies their substantial influence on the original model. Consequently, their removal has a greater impact on the model's performance, as reflected by increasing C_2.
- *Data Heterogeneity Between Remaining and Unlearned Clients:* $\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2$ measures the objectives divergence between remaining and unlearned clients. A larger value of this term indicates higher heterogeneity between the two groups, contributing to increased C_2 and thereby increasing instability.
- *Unlearning Rounds*: Increasing unlearning rounds T can enhance unlearning effectiveness as discussed in Lemma 5.6. However, the growth of T, particularly with divergent objectives $\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2$, intensifies instability by increasing C_2.

Theorem 5.10. *Let Assumptions 5.1-5.5 hold, and given an original trained model $\boldsymbol{w}^o$ undergoing unlearning. Consider the learning rate $\eta = \frac{1}{T\sqrt{\mu}}\sqrt{\frac{\beta_{\mathcal{S}}-1}{\min\{\mu(\beta_{\mathcal{S}}-1), L(\beta_{\mathcal{S}}-1)\}}}$, the sum of the FU* *verification metric and the global stability metric $V(\boldsymbol{w}^u)+S(\boldsymbol{w}^u)$ is bounded below by a constant C_s.*

$$C_s = \frac{P_{\mathcal{J}}}{\sqrt{2}\mu}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2 + \delta + C_1\,, \tag{6}$$

where C_1 is defined in Lemma 5.8.

Theorem 5.10 illustrates a fundamental *trade-off* in FU: *effectively unlearning clients (V) while maintaining the stability of the global model's performance (S).* This trade-off is determined by the divergence between the optimal model $\boldsymbol{w}^*$ and the optimal unlearned model $\boldsymbol{w}^{r*}$. Specifically, achieving effective FU and stability is not feasible under the substantial divergence between $\boldsymbol{w}^*$ and $\boldsymbol{w}^{r*}$.

5.2 Trade-off between FU Verification and Fairness

This section delves into the trade-off between FU verification and local fairness among the remaining clients. We introduce the following Theorem 5.11, which quantifies this trade-off by a lower bound for the cumulative effect of verification and fairness. The lower bound is determined by the optimality gap, which is defined as the disparity between the performance of the optimal unlearned model $\boldsymbol{w}^{r*}$ and the local optimal models for each remaining client $\boldsymbol{w_i}^*$.

Theorem 5.11 (Trade-off between Local Fairness and Effective Unlearning)**.** *Within FU, the sum of the FU verification metric and the local fairness metric is bounded below by a constant C_q:*

$$2V(\boldsymbol{w}^u) + Q(\boldsymbol{w}^u) \geq C_q = F^*_{-\mathcal{J}} - \textstyle\sum_{i\notin\mathcal{J}} p'_i f_i(\boldsymbol{w_i}^*), \tag{7}$$

where $\boldsymbol{w_i}^*$ denotes the local optimal model for client i.

Remark 5.12. The lower bound C_q underscores another fundamental trade-off in FU: *the balance between effectively unlearning (V) and maintaining fairness among the remaining clients (Q).* If C_q is large, optimizing either metric could compromise the other. The challenges of balancing this trade-off primarily arise from data heterogeneity:

- *Data Heterogeneity among Remaining Clients:* When data distribution is *homogeneous* among remaining clients, each client's optimal model ($\boldsymbol{w}_i^*$ for $\forall i \notin J$) is identical with the optimal unlearned model ($\boldsymbol{w}_i^* = \boldsymbol{w}^{r*}$). Thus, data homogeneity reduces C_q to 0, indicating FU verification and fairness can be achieved simultaneously. Conversely, high *heterogeneity* means divergent optimalities for different clients, thus increasing C_q and posing challenges in balancing this trade-off.
- *Data Heterogeneity Between Remaining and Unlearned Clients*: As discussed in Lemma 5.6, a high heterogeneity between remaining and unlearned clients increases lower bound C_1 for the unlearning verification metric V. With a constant C_q, a larger V typically leads to a reduced fairness metric Q. It indicates that under *higher heterogeneity*, fairness is enhanced for the remaining clients after unlearning. The enhanced fairness is because unlearning divergent clients $\mathcal{J}$ aligns the FU optimal model $\boldsymbol{w}^{r*}$ more closely to remaining clients than the original FL optimal model $\boldsymbol{w}^*$. Conversely, under *homogeneity* between two groups, unlearning reduces V (in Lemma 5.6 $\Delta F_{-\mathcal{J}}(\boldsymbol{w}^*, \boldsymbol{w}^{r*}) = 0$), and thereby, the fairness metric Q primarily depends on data heterogeneity among remaining clients.

5.3 Discussions

Based on our theoretical insights, we propose FU mechanisms to balance the trade-offs between FU verification and global stability, as well as FU verification and local fairness.

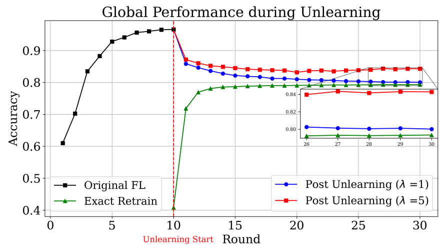
Firstly, we propose a penalty-based FU mechanism with gradient correction techniques to maintain global stability during the unlearning process, as detailed in Appendix B.1. By formulating the optimization problem and adjusting the penalty parameter, we can balance the trade-off between verification and global stability.
Moreover, we propose a framework that minimizes unlearning objectives with fairness constraints to prevent disproportionate impacts on remaining clients, as detailed in Appendix B.2.

Our theoretical analysis not only demonstrates the convergence of our method but also establishes a foundation for future exploration into the complex interplay of verification, stability, and fairness in FU. Furthermore, in Section 6, we empirically validate our FU mechanisms in non-convex settings, confirming theoretical insights by effectively balancing trade-offs.

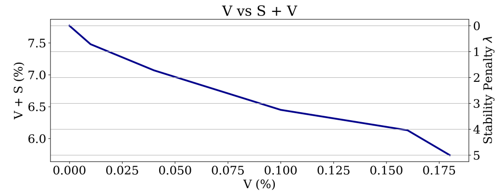
6 Experiments

6.1 Experiment Settings

In our experiment, we utilize the MNIST dataset LeCun et al. non-IID distributed across ten clients, each holding four distinct classes (the data distribution is detailed in Appendix C). We employ LeNet-5 architecture LeCun et al. (1998), a classic non-convex neural network model, to evaluate



(a) Unlearning Convergence and Handling Stability: Increasing stability penalty λ enhances global performance of FU.



(b) Trade-offs with Stability Penalty λ: Increasing λ improves the balance between FU verification and stability, reflected by reduced $V + S$.

Figure 3: FU for Balancing Stability.

Table 1: Unlearning Efficiency Compared to Retraining. (*Computing Resources*: 2 Intel Xeon Gold 5217 CPUs, 384GB RAM, and 8 Nvidia GeForce RTX-2080Ti GPUs)

	FASTER THAN RETRAIN ($\times$)
$\lambda = 1$	1.425±0.057
$\lambda = 3$	2.103±0.149
$\lambda = 5$	3.211 ±0.315

FU's consequences. To straightforwardly assess the FU evaluations metrics (V, S, Q), we focus on accuracy, *e.g.*, $V = -Acc_{-\mathcal{J}}(\boldsymbol{w}^u) + Acc_{-\mathcal{J}}(\boldsymbol{w}^{r*})$ (in percentage). A smaller value of these metrics indicates better effectiveness, stability, or fairness achieved by our FU mechanism. We further conduct additional experiments focusing on data heterogeneity and employ different datasets, as detailed in Appendix C. The overall experimental evaluation confirms our FU mechanisms in balancing the trade-offs, aligning with the theoretical insights in Section 5.

6.2 FU FOR BALANCING STABILITY

We examine the stability penalty λ in our FU mechanism in Appendix B.1 and unlearning clients $[3, 4, 8]$ starts at round 10 where the FL model has converged.

Unlearning Convergence and Handling Stability. The heterogeneity between remaining and unlearned clients leads to instability after unlearning (discussed in Lemma 5.8), as indicated by the reduced global performance after retraining in Figure 3a. Additionally, Figure 3a showcases the convergence of our FU mechanism with different stability penalties λ in the context of global performance. Specifically, with a stability penalty $\lambda = 1$, unlearning shows better stability than exact retraining. Increasing λ to 5 further improves the stability of FU.

Balancing Verification and Stability: As shown in Figure 3b, increasing λ lowers $V + S$, improving the balance between verification and stability in FU. Although a higher λ reduces FU effectiveness (increasing V), our FU mechanism allows a better trade-off within a certain tolerance level for V. Additionally, it demonstrates time efficiency compared to retraining in Table 1, making it practical in real-world FU scenarios.

6.3 FU FOR BALANCING FAIRNESS

Considering a higher data heterogeneity between unlearned and remaining clients, unlearning clients [2, 10] leads to a maximum performance drop of 1.5% ($\epsilon = 1.5$) for client 6. We employ the FU mechanism in Appendix B.2 with a fairness constraint $\epsilon = 1$ and $\Lambda = 1.3$ ensuring no client's performance deviates beyond ϵ. This setting yields the FU verification metric $V = 0.57$. We further tighten ϵ to 1, and observe V reduces to 0.11. This enhancing FU effectiveness and fairness is consistent with insights on data heterogeneity in Lemma B.8 and Theorem 5.11.

7 CONCLUSION

In this study, we investigated the trade-offs in FU under data heterogeneity, focusing on balancing unlearning verification with global stability and local fairness. We proposed a novel FU mechanism grounded in a comprehensive theoretical framework with optimization strategies and penalty controls. Our findings highlight the impacts of data heterogeneity in FU, paving the way for future research to explore adaptive FU mechanisms.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021a. doi: 10.1109/SP40001.2021.00019.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021b.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Tianshi Che, Yang Zhou, Zijie Zhang, Lingjuan Lyu, Ji Liu, Da Yan, Dejing Dou, and Jun Huan. Fast federated machine unlearning with nonlinear functional theory. In *International conference on machine learning*, pp. 4241–4268. PMLR, 2023.

Yahya H. Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A. Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7494–7502, Jun. 2023. doi: 10.1609/aaai.v37i6.25911. URL `https://ojs.aaai.org/index.php/AAAI/article/view/25911`.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

Yann Fraboni, Richard Vidal, and Marco Lorenzi. Free-rider attacks on model aggregation in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1846–1854. PMLR, 2021.

Yann Fraboni, Martin Van Waerebeke, Kevin Scaman, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Sequential informed federated unlearning: Efficient and provable client unlearning in federated optimization. *arXiv preprint arXiv:2211.11656*, 2022.

Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. VeriFi: Towards Verifiable Federated Unlearning, May 2022. URL `http://arxiv.org/abs/2205.12709`. arXiv:2205.12709 [cs].

Eric Goldman. An introduction to the california consumer privacy act (ccpa). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

Xiuting Gu, Zhu Tianqing, Jie Li, Tao Zhang, Wei Ren, and Kim-Kwang Raymond Choo. Privacy, accuracy, and model fairness trade-offs in federated learning. *Comput. Secur.*, 122(C), nov 2022. ISSN 0167-4048. doi: 10.1016/j.cose.2022.102907. URL `https://doi.org/10.1016/j.cose.2022.102907`.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pp. 3832–3842. PMLR, 2020.

Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022.

Weituo Hao, Mostafa El-Khamy, Jungwon Lee, Jianyi Zhang, Kevin J Liang, Changyou Chen, and Lawrence Carin Duke. Towards fair federated learning with zero-shot data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3310–3319, 2021.

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

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Fair federated learning via bounded group loss. *arXiv preprint arXiv:2203.10190*, 2022.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsification can simplify machine unlearning. *arXiv preprint arXiv:2304.04934*, 2023.

Ruinan Jin, Minghui Chen, Qiong Zhang, and Xiaoxiao Li. Forgettable federated linear learning with certified data removal, 2023.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Baochun Li, Ningxin Su, Chen Ying, and Fei Wang. Plato: An open-source research framework for production federated learning. In *Proceedings of the ACM Turing Award Celebration Conference-China 2023*, pp. 1–2, 2023.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a. URL `https://openreview.net/forum?id=ByexElSYDr`.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pp. 6357–6368. PMLR, 2021.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b. URL `https://openreview.net/forum?id=HJxNAnVtDS`.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022a.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federated unlearning. *arXiv preprint arXiv:2012.13891*, 2020.

Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. FedEraser: Enabling Efficient Client-Level Data Removal from Federated Learning Models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pp. 1–10, June 2021. doi: 10.1109/IWQOS52092.2021.9521274. ISSN: 1548-615X.

Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pp. 1749–1758. IEEE, 2022b.

Ziyao Liu, Yu Jiang, Jiyuan Shen, Minyi Peng, Kwok-Yan Lam, and Xingliang Yuan. A survey on federated unlearning: Challenges, methods, and future directions. *arXiv preprint arXiv:2310.20448*, 2023.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/mcmahan17a.html`.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic Federated Learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4615–4625. PMLR, May 2019. URL https://proceedings.mlr.press/v97/mohri19a.html. ISSN: 2640-3498.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A Survey of Machine Unlearning, October 2022. URL http://arxiv.org/abs/2209.02299. arXiv:2209.02299 [cs].

Chao Pan, Jin Sima, Saurav Prakash, Vishal Rana, and Olgica Milenkovic. Machine unlearning of federated clusters. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/pdf?id=VzwfoFyYDga.

Zibin Pan, Shuyi Wang, Chi Li, Haijin Wang, Xiaoying Tang, and Junhua Zhao. Fedmdfg: Federated learning with multi-gradient descent and fair guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9364–9371, 2023b.

General Data Protection Regulation. General data protection regulation (gdpr). *Intersoft Consulting, Accessed in October*, 24(1), 2018.

Yuxin Shi, Han Yu, and Cyril Leung. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2023a. doi: 10.1109/tnnls.2023.3266233. URL https://doi.org/10.1109%2Ftnnls.2023.3266233.

Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, and Mohan S. Kankanhalli. Deep regression unlearning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33921–33939. PMLR, 2023b. URL https://proceedings.mlr.press/v202/tarun23a.html.

Fei Wang, Baochun Li, and Bo Li. Federated unlearning and its privacy threats. *IEEE Network*, 2023.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. A principled approach to data valuation for federated learning. *Federated Learning: Privacy and Incentive*, pp. 153–167, 2020.

Chen Wu, Sencun Zhu, and Prasenjit Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.

Yinjun Wu, Edgar Dobriban, and Susan Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pp. 10355–10366. PMLR, 2020a.

Yinjun Wu, Edgar Dobriban, and Susan B. Davidson. Deltagrad: Rapid retraining of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10355–10366. PMLR, 2020b. URL http://proceedings.mlr.press/v119/wu20b.html.

Jiaxi Yang and Yang Zhao. A survey of federated unlearning: A taxonomy, challenges and future directions, 2023.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 393–399, 2020.

Lefeng Zhang, Tianqing Zhu, Haibin Zhang, Ping Xiong, and Wanlei Zhou. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 2023.

A Related Work

Mechine Unlearning & Federated Unlearning. Machine unlearning (MU) aims to remove specific data from a machine learning model, addressing the challenges in both effectiveness and efficiency of the unlearning process (Guo et al., 2020; Wu et al., 2020b; Bourtoule et al., 2021b;a; Tarun et al., 2023a;b; Jia et al., 2023). In FL, federated unlearning (FU) is proposed to address clients' right to be forgotten, including methods like rapid retraining (Liu et al., 2022b), subtracting historical updates from the trained model (Wu et al., 2022), subtracting calibrated gradients of the unlearn clients to remove their influence (Liu et al., 2020; 2021), and adding calibrated noises to the trained model by differential privcacy (Zhang et al., 2023). However, none of them involves any rigorous consideration for *data heterogeneity*, the main challenge in FL. In this work, we account for data heterogeneity in FU through a comprehensive optimization framework, and theoretically analyze how data heterogeneity impacts unlearning in Section 5.

Moreover, existing FU methods has focused on methods ensuring verifiable and efficient unlearning (Liu et al., 2021; 2022b; Fraboni et al., 2022; Gao et al., 2022; Jin et al., 2023; Che et al., 2023). The *verification* of unlearning typically involves comparing the unlearned model, obtained through an unlearning mechanism, with a reference model using performance metrics such as accuracy and model similarity metrics (Gao et al., 2022). Additionally, attack-based verification methods, such as membership inference attacks (MIA) and backdoor attacks (BA), are often employed in MU (Nguyen et al., 2022) but are not applicable in federated systems due to privacy concerns. Besides verification, data heterogeneity in FU introduces consequences on global stability and local fairness, necessitating consideration of trade-offs in FU, as previously discussed in Section 1. In this work, we conduct a rigorous analysis of inherent trade-offs in FU under data heterogeneity .

Stability in FL and FU. In FL, *performance stability* revolves around maintaining consistent and robust model performance despite alterations in the training dataset. This aspect of stability, highlighted in studies such as Yin et al. (2018); Fang et al. (2020); Li et al. (2021), often involves defending against external threats. Unlike FL, FU is concerned with managing internal changes within the system for users' rights to be forgotten. Considering unlearning in the federated system, the inherent data heterogeneity can lead to significant shifts in data distribution, thereby altering the model's performance. In this work, we delve into the theoretical analysis of unlearning, examining how data heterogeneity impacts stability in FU.

Fairness in FL and FU. In FL, there are several works that have proposed different notions of fairness. The proportional fairness ensures whoever contributes more to the model can gain greater benefits (Wang et al., 2020; Yu et al., 2020). Additionally, the model fairness focuses on protecting specific characteristics, like race and gender (Gu et al., 2022). Furthermore, the *performance fairness* (Li et al., 2019; Mohri et al., 2019; Hao et al., 2021; Li et al., 2021; Shi et al., 2023; Pan et al., 2023b) aims to reduce the *variance of local test performance or utility* across all clients.

In FU, we observe that unlearning certain clients can lead to *unequal* impacts on remaining clients due to data heterogeneity, as discussed in Section 1. In this work, we extend performance fairness to FU by the variance of *utility changes* among remaining clients and further analyze how data heterogeneity impacts fairness in FU.

B Optimizing FU under Trade-offs

In the previous section, we examine the inherent trade-offs involving FU verification and their challenges. To balance these trade-offs, this section introduces our FU mechanisms developed within an optimization framework[3].

B.1 FU for Balancing Global Stability

In FU, maintaining global stability is crucial for ensuring the overall performance and reliability of the federated system throughout the unlearning process. However, as explored in Section 5.1, a trade-off exists between FU verification and global stability. To manage this trade-off, we propose an FU mechanism utilizing a penalty-based approach and gradient correction techniques. We also theoretically demonstrate the convergence of our method.

FU Mechanism Design: To balance stability during FU, we formulate the optimization problem for unlearning as:

$$\textbf{P1:}\ \min_{\boldsymbol{w}} V(\boldsymbol{w}) + \lambda S(\boldsymbol{w}). \tag{8}$$

By adjusting λ, we can manage the trade-off between these two objectives, allowing for a flexible approach to specific requirements of the federated system.

[3]These FU mechanisms are grounded in approximate unlearning, which gives tolerance on effectiveness.

By the definition of stability metric $S(\boldsymbol{w})$, we have: $S(\boldsymbol{w}) = \mathbb{E}\left[F(\boldsymbol{w})\right] - F(\boldsymbol{w}^*) = \mathbb{E}\left[F(\boldsymbol{w}^u)\right] - F(\boldsymbol{w}^o) + \delta$, where $\delta = F(\boldsymbol{w}^o) - F(\boldsymbol{w}^*)$. Consequently, solving **P2**: $\min_{\boldsymbol{w}} F_{-\mathcal{J}}(\boldsymbol{w}) + \lambda\left(F(\boldsymbol{w}) - F(\boldsymbol{w}^o)\right)$ optimizes **P1**. However, in **P2**, optimizing the global objective $F(\boldsymbol{w})$ *among all clients* is untraceable in FU as it cannot involve unlearned client $j \in \mathcal{J}$ in unlearning process. To address this, we consider the approximate problem **P3** to **P2**:

$$\textbf{P3:}\ \min_{\boldsymbol{w}} H(\boldsymbol{w}) := [F_{-\mathcal{J}}(\boldsymbol{w}) + \tilde{h}(\boldsymbol{w})], \tag{9}$$

where $\tilde{h}(\boldsymbol{w}) = (1 - P_{\mathcal{J}})F_1 + P_{\mathcal{J}}F_2$, $F_1 := \lambda F_{-\mathcal{J}}(\boldsymbol{w})$, and $F_2 := \lambda\left(\langle \nabla F_{\mathcal{J}}(\boldsymbol{w}^o), \boldsymbol{w} - \boldsymbol{w}^o\rangle + \frac{L}{2}\left\|\boldsymbol{w} - \boldsymbol{w}^o\right\|^2\right)$.

To address **P3**, we propose an FU mechanism that operates two steps during each unlearning round t:

1. *Federated Aggregation*: The remaining client $i \in \mathcal{S}$ performs local training over E epochs with learning rate η_l to obtain $\boldsymbol{w}_i^{(t,E)}$. Then, the server aggregates $\{\boldsymbol{w}_i^{(t,E)}\}_{i\in\mathcal{S}}$ for the global model $\bar{\boldsymbol{w}}^{(t,E)} = \sum_{i\in\mathcal{S}} \alpha_i \cdot \boldsymbol{w}_i^{(t,E)}$, where α_i is the weight for client i.
2. *Global Correction*: Following the aggregation, the server applies a gradient correction to $\bar{\boldsymbol{w}}^{(t,E)}$. Specially, the server compute $\boldsymbol{h}^t = \lambda(1 - P_{\mathcal{J}})\boldsymbol{g}_{\mathcal{S}}^t + \lambda P_{\mathcal{J}}\hat{\boldsymbol{g}}_{\mathcal{J}}^t$,[4] where $\hat{\boldsymbol{g}}_{\mathcal{J}}^t(\boldsymbol{w}) = \nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) + L(\boldsymbol{w} - \boldsymbol{w}^o)$. The correction term $\boldsymbol{g}_c^t$ is then obtained by projecting $\boldsymbol{h}^t$ onto the tangent space of the aggregated gradient $\boldsymbol{g}_{\mathcal{S}}$: $\boldsymbol{g}_c^t = \boldsymbol{h}^t - \text{Proj}_{\boldsymbol{g}_{\mathcal{S}}}\boldsymbol{h}^t$. The global model is updated for the next round: $\bar{\boldsymbol{w}}^{(t+1,0)} = \bar{\boldsymbol{w}}^{(t,E)} - \eta_g \boldsymbol{g}_c^t$, where η_g is the learning rate for the gradient correction.

The FU mechanism thus iteratively updates the global model by $\bar{\boldsymbol{w}}^{(t+1,0)} = \boldsymbol{w}^{(t,0)} - \eta_l \boldsymbol{g}_{\mathcal{S}}^t - \eta_g \boldsymbol{g}_c^t$.

Theoretical Analysis: Now, we delve into the convergence of the proposed FU mechanism, ensuring its reliability in FU. We also conduct theoretical analysis to determine the upper bound for the verification metric V, which is essential for verifying the effectiveness of unlearning. To begin, we formally state the assumptions required for our main results.

Assumption B.1. The gradients of local objectives are bounded, *i.e.*, $\|\nabla f_i(\boldsymbol{w})\| \leq G$ for all i. This implies the gradient of global objective $F_{-\mathcal{J}}$ for remaining clients is also bounded: $\|\nabla F_{-\mathcal{J}}(\boldsymbol{w})\| \leq G$.

Assumption B.2. The heterogeneity between the unlearned clients $\mathcal{J}$ and the remaining clients is quantified as:

$$\left\|\mathbb{E}\left[\hat{\boldsymbol{g}}_{\mathcal{J}}^t(\boldsymbol{w})\right] - \nabla F_{-\mathcal{J}}(\boldsymbol{w})\right\|^2 \leq \zeta'^2 + \beta'^2 \left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w})\right\|^2,$$

where $\hat{\boldsymbol{g}}_{\mathcal{J}}^t(\boldsymbol{w}) = \nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) + L(\boldsymbol{w} - \boldsymbol{w}^o)$, ζ'^2 and β'^2 indicate heterogeneity between unlearned and remaining clients.

The following lemma derives the upper bound on the expected norm of gradient correction, and then we establish the convergence theorem of our proposed FU mechanism.

Lemma B.3. *Under Assumption B.2, the expected norm of the gradient correction at unlearning round t is bounded:*

$$\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)})\right\|^2\right] \leq \phi\left(\zeta'^2 + (\beta'^2 + 1)\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{t,E})\right\|^2\right),$$

where $\phi = \lambda^2 P_{\mathcal{J}}^2(1 + \cos^2\theta)$, $\cos^2\theta$ represents the similarity in objectives between remaining and unlearned clients.

Theorem B.4 (Convergence)**.** *Let Assumptions 5.1-5.4, B.1 and B.2 hold, we consider an FU mechanism with diminishing step size $\eta_l = \frac{\beta}{2(t+\gamma)}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$, such that $\eta_l \leq \frac{1}{4L}$. The convergence result after t rounds is:*

$$\mathbb{E}\left[H\left(\bar{\mathbf{w}}^{(t+1,0)}\right)\right] - H\left(\mathbf{w}^*\right) \leq \frac{L}{2}\left(\frac{v}{\gamma + t} + \left\|\boldsymbol{w}^{r*} - \boldsymbol{w}^*\right\|^2\right),$$

where $v = \max\left\{\frac{\beta^2 B}{\beta\mu - 4}, (\gamma+1)\left\|\boldsymbol{w}^o - \boldsymbol{w}^{r*}\right\|^2\right\}$.

Our approach introduces additional complexity in the convergence analysis compared to that of Li et al. (2020b, Theorem 1) due to incorporating a gradient correction in FU, as detailed in Appendix H.1 This complexity is reflected in B with additional components: $2\phi\left(\frac{\eta_g}{\eta_l}\right)^2((\beta'^2 + 1)G^2 + \zeta'^2)$. It highlights two insights:

[4]For simplicity, denote $\boldsymbol{h}^t := \boldsymbol{h}(\bar{\boldsymbol{w}}^{(t,E)})$

1. A high *data heterogeneity* between remaining and unlearned clients, indicated by β'^2 and ζ'^2, *increases unlearning rounds* T for FU convergence;
2. The term $\phi = \lambda^2 P_{\mathcal{J}}^2(1+\cos^2\theta)$ in the convergence bound indicates that larger influence of unlearned clients (characterized by $P_{\mathcal{J}}$) and larger stability penalties (λ) increase rounds T needed for FU convergence.

In the special case of *homogeneity* between remaining and unlearned clients, where the original model and the optimal unlearned model are ideally aligned ($\|\boldsymbol{w}^{r*} - \boldsymbol{w}^*\|^2 = 0$, $\beta'^2 = \zeta'^2 = 0$, $\cos\theta = 1$, and $P_{\mathcal{J}} = 0.5$), the additional term in B reduces to $\lambda^2\left(\frac{\eta_l}{\eta_g}\right)^2 G^2$. In this scenario, handling stability in FU is straightforward for the tight bound. Conversely, our mechanism reduces the convergence bound in *heterogeneous* settings, characterized by orthogonal gradient correction to remaining clients' gradients ($\cos\theta = 0$). This indicates we effectively adapt to this heterogeneity.

Next, we verify unlearning in our FU mechanism by a theoretical upper bound on V. The additional assumptions and lemmas are stated as follows:

Assumption B.5. For each round t, the norm of the gradient after E epochs is bounded by the gradient at the start of the round t, $\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,E)})\right\|^2 \le \epsilon \left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2$.

Lemma B.6. *Under Lemma B.3 and Assumption B.5, the expected norm of the gradient correction at unlearning round t is bounded:* $\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)}) - \nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2\right] \le \zeta''^2 + \beta''^2 \left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2$, *where* $\zeta''^2 = \phi\epsilon\zeta'^2, \beta''^2 = \phi\epsilon\beta'^2 + \phi\epsilon + 1$.

Theorem B.7 (Verifiable Unlearning). *Under Assumptions 5.1-5.4, and Assumptinos B.1-B.5, taking* $\eta_l = \eta_g = \frac{2}{LT}$, *and* $T \ge \max\left\{2\beta_{\mathcal{S}}^2 + 2, \frac{1+\Delta}{4L}, \frac{1}{2}(\beta''^2+1), \frac{1+\Delta'}{L}\right\}$, *where* $\Delta = \sqrt{\max\{0, 1 - 16L(\beta_{\mathcal{S}}^2+1)\}}$ *and* $\Delta' = \sqrt{\max\{0, 1 - L(\beta''^2+1)\}}$. *Then, the verification metric V is bounded as follows:*
$V = \mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^u) - F_{-\mathcal{J}}(\boldsymbol{w}^*)\right] \le \chi_1 + \chi_2$, *where*

$$\chi_1 = \frac{1}{2}\left(1 - \frac{1}{2LT} + \frac{\beta_{\mathcal{S}}^2+1}{LT^2}\right)^T D + \frac{(\sigma^2 + \zeta_{\mathcal{S}}^2)}{2LT},$$

$$\chi_2 = \frac{1}{2}\left(1 - \frac{1}{LT} + \frac{\beta''^2+1}{LT^2}\right)^T D + \frac{\zeta''^2}{2LT}.$$

Here, $D = F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}), \sigma^2 = \sum_{i\in\mathcal{S}} \alpha_i^2 \sigma_{i,t}^2$

From Theorem B.7, the FU verification metric V is primarily determined by two factors:

1. *Data Heterogeneity among Remaining Clients* ($\beta_{\mathcal{S}}^2$): Within χ_1, a higher data heterogeneity necessitates more unlearning rounds T to lower V for effective unlearning;
2. *Impact of Global Gradient Correction*: Within χ_2, β''^2 encapsulates stability penalty (λ), unlearned clients' influence (captured by $P_{\mathcal{J}}$), and the data heterogeneity between remaining and unlearned clients. Increasing either of them requires more rounds T to lower V.

Additionally, Theorem B.7 highlights future adaptive strategies with client sampling or reweighting to reduce heterogeneity and variance of sampled remaining clients $\mathcal{S}$ in FU.

B.2 FU for Balancing Local Fairness

As discussed in Section 5.2, FU can lead to uneven impacts across different clients due to data heterogeneity. To address this, we propose an optimization framework to minimize the unlearning objective with fairness constraints, ensuring that unlearning does not unequally harm any remaining clients. We highlight our contribution to a theoretical and practical groundwork for balancing fairness and verification in FU, and providing insights for future adaptive strategies.

$$\textbf{P4:} \min_{\boldsymbol{w}} F_{-\mathcal{J}}(\boldsymbol{w})$$
$$\text{s.t. } \Delta f_i = f_i(\boldsymbol{w}) - f_i(\boldsymbol{w}^o) \le \epsilon, \forall i \notin \mathcal{J}$$

To solve this problem, we adapt the saddle point optimizations as in (Agarwal et al., 2018; Hu et al., 2022), using a Lagrangian multiplier λ_i for each constraint:

$$\textbf{P5:} \min_{\boldsymbol{w}} \max_{\boldsymbol{\lambda}\in\mathcal{R}^Z, \|\boldsymbol{\lambda}\|\le\Lambda} F_{-\mathcal{J}}(\boldsymbol{w}) + \boldsymbol{\lambda}^\top \boldsymbol{r}(\boldsymbol{w}),$$

where $\boldsymbol{r}(\boldsymbol{w}^o) = [\Delta f_i - \epsilon]_{i \notin \mathcal{J}}$. The detailed algorithm for solving this problem is provided in Appendix B.2.

Algorithm 1 Balancing Fairness Federated Unlearning

```
Require: T, η, ϵ, Λ
Initialize λ = 0.
for t = 0, ··· , T − 1 do
    Server broadcasts w^t, λ_i to each client i ∉ J.
    for each client i do
        # Local Training Start.
        r_i = f_i(w^t) − f_i(w^o)
        Update weight w_i for E local epochs with f'(w) = f_i(w) + λ_i (f_i(w) − f_i(w^o))
        w_i^(t,τ+1) = w_i^(t,τ) − η∇f'_i(w_i^(t,τ))
        Send w^(t,E) and r_i back to the server
        # Local Training End.
    end for
    Server aggregates the weight w^(t+1) = Σ_{i∉J} p'_i w_i^(t,E).
    If max_i r_i ≤ ϵ: Return w^(t+1)
    Update λ_i = Λ exp(r_i) / (1 + Σ_{i∉J} exp(r_i)) for ∀i ∉ J.
end for
Return w^T
```

Lemma B.8. *Assume $\|\mathbf{r}\|_\infty \leq \rho$, and suppose $\nu = 2\rho^2\Lambda$, achieving a ν-approximate saddle point of **P5** requires $T \geq \frac{1}{\nu(\gamma+1)-2\kappa\mathcal{C}}\left(\frac{M}{\nu} + 2\kappa\mathcal{C}(\gamma-1)\right)$, where M is a constant and $\mathcal{C} = \frac{2C}{(1+\Lambda)\mu} + \frac{(1+\Lambda)\mu\gamma}{2}\mathbb{E}\left[\|\boldsymbol{w}^o - \boldsymbol{w}^{r*}\|^2\right]$, with C specified in Lemma H.1.*

Remark B.9. This lemma can be derived from Hu et al. (2022, Theorem 1). It indicates the required unlearning rounds to reach a ν-approximate saddle point of **P5**. It highlights that increased data heterogeneity among remaining clients and stringent fairness constraints require more unlearning rounds T to balance the trade-off.

Theorem B.10. *Given $\epsilon = \frac{1}{\Lambda}(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F^*_{-\mathcal{J}})$ and assuming the existence of ν-approximate saddle points of the trade-off fairness problem, then the unlearning verification metric $V \leq 2\nu$ and $\max_{i\notin\mathcal{J}} \Delta f_i \leq \epsilon$.*

Theorem B.10 emphasizes the feasibility of ν-approximate suboptimal solution that balances FU verification with fairness constraint ϵ, providing two insights:

1. *Data Hetegeneity*: When the original FL model and optimal FU model are homogeneous, then $\epsilon = F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^*)$, there is no fairness loss from unlearning but from data heterogeneity among remaining clients (as discussed in Theorem 5.11). However, with higher heterogeneity and *lacking a focus on balancing fairness* (characterized by a negligible Λ and a small ν), FU *compromises fairness* ϵ to reduce V.
2. ν *Selection*: Choosing a smaller ν potentially reduces V but aggressively minimizing ν risks infeasibility and increased resources for growing T (stated in Lemma B.8).

For future work, these insights suggest advanced strategies adjusting to data heterogeneity and system constraints.

B.3 Trade-off between Global Stability and Local Fairness

Remark 5.12 shows that high heterogeneity between remaining and unlearned clients can enhance fairness in FU by aligning the optimal model more closely with remaining clients. Yet, Remark 5.9 highlights how this heterogeneity can undermine system stability, revealing a trade-off where increased fairness may affect stability. Future work will explore balancing verification, stability, and fairness within FU's complex dynamics.

C Experiments

In this section, we present experiments to validate the proposed FU mechanisms. We examine various scenarios and settings to demonstrate the effectiveness and robustness of our approaches. Our implementation utilizes the open-source FL framework Plato (Li et al., 2023) for reproducibility.

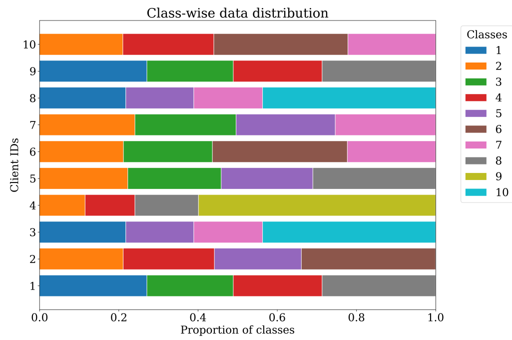


Figure 4: Non-IID Class Distribution for MNIST (Section 6) and CIFAR-10: Each client has four classes.

C.1 Experiment Settings

In our experiment, we utilize the MNIST dataset (LeCun et al.) and the CIFAR-10 dataset (Krizhevsky, 2009) that are non-IID distributed across ten clients, each holding four distinct classes (the data distribution is detailed in Figure 4). We employ LeNet-5 architecture (LeCun et al., 1998) for MNIST and the ResNet18 architecture (He et al., 2016) for CIFAR-10 to evaluate FU's consequences. The classic non-convex neural network models are to demonstrate the effectiveness and robustness of our approaches. To straightforwardly assess the FU evaluations metrics (V, S, Q), we focus on accuracy, *e.g.*, $V = -Acc_{-\mathcal{J}}(\boldsymbol{w}^u) + Acc_{-\mathcal{J}}(\boldsymbol{w}^{r*})$ (in percentage). A smaller value of these metrics indicates better effectiveness, stability, or fairness achieved by our FU mechanism. The overall experimental evaluation confirms our FU mechanisms in balancing the trade-offs, aligning with the theoretical insights in Section 5.

C.2 FU for Balancing Stability

We examine the stability penalty λ in our FU mechanism in Appendix B.1 and unlearning clients $[3, 4, 8]$ starts at round 10 where the FL model has converged.

Unlearning Convergence and Handling Stability. The heterogeneity between remaining and unlearned clients leads to instability after unlearning (discussed in Lemma 5.8), as indicated by the reduced global performance after retraining in Figure 3a. Additionally, Figure 3a showcases the convergence of our FU mechanism with different stability penalties λ in the context of global performance. Specifically, with a stability penalty $\lambda = 1$, unlearning shows better stability than exact retraining. Increasing λ to 5 further improves the stability of FU.

Balancing Verification and Stability: As shown in Figure 3b, increasing λ lowers $V + S$, improving the balance between verification and stability in FU. Although a higher λ reduces FU effectiveness (increasing V), our FU mechanism allows a better trade-off within a certain tolerance level for V. Additionally, it demonstrates time efficiency compared to retraining in Table 1, making it practical in real-world FU scenarios.

C.3 FU for Balancing Fairness

Considering a higher data heterogeneity between unlearned and remaining clients, unlearning clients [2, 10] leads to a maximum performance drop of 1.5% ($\epsilon = 1.5$) for client 6. We employ the FU mechanism in Appendix B.2 with a fairness constraint $\epsilon = 1$ and $\Lambda = 1.3$ ensuring no client's performance deviates beyond ϵ. This setting yields the FU verification metric $V = 0.57$. We further tighten ϵ to 1, and observe V reduces to 0.11. This enhancing FU effectiveness and fairness is consistent with insights on data heterogeneity in Lemma B.8 and Theorem 5.11.

C.4 Heterogeneity between remaining and unlearned clients

This section investigates the stability in FU under varying levels of heterogeneity *between these two groups*, and data within both remaining and unlearned clients is homogeneous. According to Lemma B.8, this setting should facilitate fairness in the unlearning process. Specifically, the number of classes in clients' datasets follows a Dirichlet distribution with parameter α, where a lower α value indicates higher data heterogeneity. We explore scenarios with α values corresponding to label distributions of 0.1, 0.4, and 0.7.

Table 2 demonstrates a correlation between the data heterogeneity level and the system's stability after unlearning.

Table 2: Data Heterogeneity Between Groups ($P_J = 0.38$): A lower label distribution indicates higher heterogeneity. Greater heterogeneity leads to increased instability after FU (evidenced by larger S_{retrain} values). Employing a stability trade-off FU approach with $\lambda = 1$ enhances stability (reduced $S_{\lambda=1}$) while maintaining negligible compromise in FU verification ($V_{\lambda=1}$).

Label Distribution	S_{retrain}	$S_{\lambda=1}$	$V_{\lambda=1}$
0.1	8.24	7.69 (-0.55)	0.0
0.4	3.45	2.05 (-1.4)	0.0
0.7	0.26	0.23 (-0.03)	0.06

A higher data heterogeneity leads to increased instability after unlearning, as indicated by the higher S_{retrain} values. Notably, applying our FU mechanism with a stability trade-off parameter ($\lambda = 1$) results in improved stability, as shown by the reduced $S_{\lambda=1}$ values, while the difference in FU verification remains negligible. This underscores our proposed approach could balance the trade-off between unlearning verification and global stability under varying degrees of data heterogeneity.

To further verify the unlearning, we examine the accuracy for class 4, which is unique to the unlearned clients. In the original trained model $\boldsymbol{w}^o$, we observe a 77.09% accuracy for class 4. However, in the retrained model $\boldsymbol{w}^r$, the accuracy for class 4 drops to 0%, indicating successful unlearning. In our FU mechanism with stability penalty $\lambda = 1$, the accuracy for class 4 is 0.3%, further validating the effectiveness of our approach in unlearning the influence of unlearned clients while maintaining stability.

C.5 Heterogeneity among remaining clients.

This section explores the fairness in FU under varying levels of heterogeneity *among remaining clients*, and data between remaining and unlearned clients is also heterogeneous[5].

To extend our analysis, we consider the CIFAR-10 dataset (Krizhevsky, 2009) with class distribution in Figure 4 and utilize the ResNet18 architecture (He et al., 2016). CIFAR-10 exhibits greater complexity compared to MNIST. Figure 5 illustrates the impact of unlearning on fairness among remaining clients, with a fairness parameter $\Lambda = 1, \epsilon = 20$. In this scenario, clients [1, 3] experienced significant utility loss due to their similar data distribution with unlearned clients [8, 9] (as shown in Figure 4). However, our FU mechanism achieves a lower fairness metric ($Q_{\text{our}} = 6.91$) compared to the retraining approach ($Q_{\text{retrain}} = 10.23$), indicating more equitable utility changes among remaining clients. The verification metric in this case is $V = 0.42$, demonstrating that our mechanism enhances fairness even in more complex and heterogeneous environments.

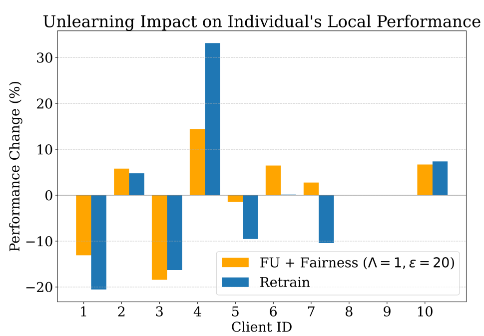


Figure 5: Unlearning Impact on Fairness: ensuring no client's performance deviates beyond 20%; fairness metric (Q) of our FU mechanism ($Q_{\text{our}} = 6.91$) is lower than that of retrain $Q_{\text{retrain}} = 10.23$

C.6 Balancing Stability under ResNet18, CIFAR 10

We examine the stability penalty λ in on more complicated dataset and model, and unlearning clients $[3, 4, 8]$ starts at round 10 where the FL model has converged. With a stability penalty $\lambda = 1$, unlearning shows better stability ($S = 1.42$) than exact retraining ($S = 2.42$). Moreover, our FU mechanism improves the balance between verification and stability in FU. This is evidenced by a

[5]According to Lemma B.8, given homogeneous two groups, the fairness primarily depends on data heterogeneity among remaining clients, as we aim to investigate the impact of unlearned clients, we choose heterogeneous two groups setting.

Table 3: Balancing Stability in FU (ResNet18, CIFAR10)

	Global Performance (%)	Unlearning Performance (%)	V (%)	S (%)	V + S (%)
Original FL	63.6	-	-	-	-
Retrain	61.18	63.99	0	2.42	2.42
Unlearning ($\lambda = 1$)	62.10	63.57	0.42	1.42	1.84

reduction in the combined metric of verification and stability, denoted as $V + S = 1.83$, which is lower than the corresponding value for exact retraining (2.42). Additionally, our FU mechanism demonstrates time efficiency, achieving a speedup of $\times 2.79$ compared to retraining.

D PROOF OF LEMMA 5.6

Proof. Given the FU verification metric $V(\boldsymbol{w}^u) = \mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^u)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})$, we analyze the metric using iterative updates in the FU process.

For each iteration $t+1$:

$$\begin{aligned} F_{-\mathcal{J}}(\boldsymbol{w}^{t+1}) - F_{-\mathcal{J}}(\boldsymbol{w}^t) &\geq \left\langle \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t), \boldsymbol{w}^{t+1} - \boldsymbol{w}^t \right\rangle + \frac{\mu}{2}\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2 \\ &= \left\langle -\eta \boldsymbol{g}_{\mathcal{S}}^t, \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t) \right\rangle + \frac{\mu\eta^2}{2}\|\boldsymbol{g}_{\mathcal{S}}^t\|^2 \end{aligned}$$

where $\boldsymbol{g}_{\mathcal{S}}^t = \sum_{i\in\mathcal{S}} \alpha_i \boldsymbol{g}_i^t$, and α_i is the aggregation weight of client i.

Taking expectations on both sides, we derive:

$$\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1})\right] - F_{-\mathcal{J}}(\boldsymbol{w}^t) \geq \frac{\eta}{2} \underbrace{\left\|\boldsymbol{G}_{\mathcal{S}}^t - \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2}_{A_1} \underbrace{-\frac{\eta}{2}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\|^2 - \frac{\eta}{2}\left\|\boldsymbol{G}_{\mathcal{S}}^t\right\|^2 + \frac{\mu\eta^2}{2}\mathbb{E}\|\boldsymbol{g}_{\mathcal{S}}^t\|^2}_{A_2} \tag{10}$$

where $\boldsymbol{G}_{\mathcal{S}}^t = \sum_{i\in\mathcal{S}} \alpha_i \boldsymbol{G}_i^t$, with $\boldsymbol{G}_i^t = \mathbb{E}\left[\boldsymbol{g}_i^t\right]$ representing the gradient for client i at iteration t, while $\boldsymbol{g}_i^t$ denotes the stochastic gradient.

Firstly, for A_1:

$$\left\|\boldsymbol{G}_{\mathcal{S}}^t - \nabla F_{-\mathcal{J}}(\boldsymbol{w}^{\boldsymbol{t}})\right\|^2 \leq \mathbb{E}_i \left\|\boldsymbol{G}_i^t - \nabla F_{-\mathcal{J}}(\boldsymbol{w}^{\boldsymbol{t}})\right\|^2 \underset{Assumption\ 5.1}{\leq} \zeta_{\mathcal{S}}^2 + \beta_{\mathcal{S}}^2\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^{\boldsymbol{t}})\|^2 \tag{11}$$

Let $\left\|\boldsymbol{G}_{\mathcal{S}}^t - \nabla F_{-\mathcal{J}}(\boldsymbol{w}^{\boldsymbol{t}})\right\|^2 = \zeta_t^2 + \beta_{\mathcal{S}}^2\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^{\boldsymbol{t}})\|$, where ζ_t^2 is specific for each round t ($\zeta_t^2 \leq \zeta_{\mathcal{S}}^2$).

Applying the triangle inequality, we derive the following relation for A_1: $\|\boldsymbol{G}_{\mathcal{S}}^t\|^2 \geq \zeta_t^2 + (\beta_{\mathcal{S}}^2 - 1)\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^{\boldsymbol{t}})\|^2$.

Then, for A_2, we expand it as follows:

$$A_2 = \frac{\mu\eta^2}{2} \underbrace{\mathbb{E}\left\|\boldsymbol{G}_{\mathcal{S}}^t - \boldsymbol{g}_{\mathcal{S}}^t\right\|^2}_{=\sigma_t^2 \leq \sigma^2 (Assumption\ 5.4)} + \frac{\eta}{2}\left(\eta\mu - 1\right)\left\|\boldsymbol{G}_{\mathcal{S}}^t\right\|^2 - \frac{\eta}{2}\left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2 \tag{12}$$

Now, by Equation (11) and Equation (12), taking expectation on both sides of Equation (10):

$$\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1})\right] - F_{-\mathcal{J}}(\boldsymbol{w}^t) \geq \frac{\eta}{2}\underbrace{\left(\zeta_t^2 + \beta_{\mathcal{S}}^2\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\|^2\right)}_{=A_1} \tag{13}$$

$$+ \underbrace{\frac{\mu\eta^2}{2}\sigma_t^2 + \frac{\eta}{2}\left(\eta\mu\beta_{\mathcal{S}}^2 - \eta\mu - \beta_{\mathcal{S}}^2\right)\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\|^2 + \frac{\eta}{2}\left(\eta\mu - 1\right)\zeta_t^2}_{\leq A_2}$$

$$= \frac{\mu\eta^2}{2}\left(\beta_{\mathcal{S}}^2 - 1\right)\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\|^2 + \frac{\eta^2\mu}{2}\left(\sigma_t^2 + \zeta_t^2\right)$$

$$\geq \underbrace{\frac{\mu\eta^2}{2}M\left(\beta_{\mathcal{S}}^2 - 1\right)}_{B_1}\left(F_{-\mathcal{J}}(\boldsymbol{w}^t) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) + \frac{\eta^2\mu}{2}\left(\sigma_t^2 + \zeta_t^2\right), \tag{14}$$

where $M = 2\mu$ if $\beta_{\mathcal{S}} \geq 1$, else $M = 2L$.

Let $\bar{\sigma^2} = \frac{1}{T}\sum_{t=1}^T \sigma_t^2$ and $\bar{\zeta^2} = \frac{1}{T}\sum_{t=1}^T \zeta_t^2$. Thus, by iterative updates, we have:

$$\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \geq \left(B_1+1\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) + \frac{\eta^2\mu}{2}(\bar{\sigma^2} + \bar{\zeta^2})\sum_{r=1}^{T}\left(B_1+1\right)^{T-r} \tag{15}$$

Taking $\eta^2 = \frac{2}{\mu M T^2}$, we have $B_1 = \frac{\beta_{\mathcal{S}}^2 - 1}{T^2}$.

Now, considering two cases for B_1:

Case 1: If $B_1 \leq -1$ (since $T^2 \geq 1$, Case 1 only holds when $T = 1$ and $\beta_{\mathcal{S}}^2 = 0$), we get:

$$\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \geq \frac{1}{M}\left(\bar{\sigma^2} + \bar{\zeta^2}\right), \tag{16}$$

Case 2: If $B_1 > -1$, then $(B_1+1)^T \geq 1 + TB_1$, which leads to the inequality:

$$\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \geq \left(1 + \frac{\beta_{\mathcal{S}}^2 - 1}{T}\right)\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) + \rho\left(\bar{\sigma^2} + \bar{\zeta^2}\right), \tag{17}$$

where $\rho = \frac{\eta^2\mu}{2B_1}\left(\left(B_1+1\right)^T - 1\right) \geq 0$ for $\forall B_1 \geq -1$. Therefore, we can derive $\rho \geq \frac{\eta^2\mu T}{2} = \frac{1}{MT}$

Combining two cases and $M \leq 2L$, we conclude

$$\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \geq \left(1 + \frac{\beta_{\mathcal{S}}^2 - 1}{T}\right)\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) + \frac{1}{2LT}\left(\bar{\sigma^2} + \bar{\zeta^2}\right), \tag{18}$$

$\square$

E PROOF OF LEMMA 5.8

Proof. Given the stability metric $S(\boldsymbol{w}^u)$, we express it as

$$S(\boldsymbol{w}^u) = \mathbb{E}\left[F(\boldsymbol{w}^u)\right] - F(\boldsymbol{w}^*) = A_1 + A_2, \tag{19}$$

where $A_1 = \mathbb{E}\left[F(\boldsymbol{w}^u)\right] - F(\boldsymbol{w}^o)$ and $A_2 = F(\boldsymbol{w}^o) - F(\boldsymbol{w}^*) = \delta$. A_2 represents the empirical risk minimization (ERM) gap.

Firstly, to bound A_1, we utilize the convexity of F: $A_1 \geq \langle\nabla F(\boldsymbol{w}^o), \mathbb{E}\left[\boldsymbol{w}^u - \boldsymbol{w}^o\right]\rangle + \frac{\mu}{2}\mathbb{E}\left[\|\boldsymbol{w}^u - \boldsymbol{w}^o\|^2\right]$.

With $\boldsymbol{w}^u - \boldsymbol{w}^o = -\eta\sum_{r=1}^T \mathbf{g}_r^{(\mathcal{S})}$, and $\mathbf{g}_r^{(\mathcal{S})}$ being the aggregated stochastic gradient from the subset of remaining clients $\mathcal{S} \subseteq \mathcal{N} \setminus \mathcal{J}$, let $\bar{\mathbf{g}}^{(\mathcal{S})} = \frac{1}{T}\sum_{r=1}^T \mathbf{g}_r^{(\mathcal{S})}$, we have $\boldsymbol{w}^u - \boldsymbol{w}^o = -\eta T\bar{\mathbf{g}}^{(\mathcal{S})}$.

Considering FL with remaining clients, the global objective is $F_{-\mathcal{J}} = \sum_{i\notin\mathcal{J}} p_i' f_i$, where $p_i' = \frac{p_i}{1-P_{\mathcal{J}}}$.

For FL with unlearned clients, the global objective is $F_{\mathcal{J}} = \sum_{j\in\mathcal{J}} p_j' f_j$, where $p_j' = \frac{p_j}{P_{\mathcal{J}}}$. Then, $\nabla F(\cdot) = (1-P_{\mathcal{J}})\nabla F_{-\mathcal{J}}(\cdot) + P_{\mathcal{J}}\nabla F_{\mathcal{J}}(\cdot)$.

Expanding A_1, we get:

$$
\begin{aligned}
&= \frac{(1-P_{\mathcal{J}})\eta T}{2}\mathbb{E}\left[\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \bar{\mathbf{g}}^{(\mathcal{S})}\|^2\right] + \frac{P_{\mathcal{J}}\eta T}{2}\mathbb{E}\left[\|\nabla F_{\mathcal{J}}(\boldsymbol{w}^o) - \bar{\mathbf{g}}^{(\mathcal{S})}\|^2\right] \\
&\quad - \frac{(1-P_{\mathcal{J}})\eta T}{2}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o)\|^2 - \frac{P_{\mathcal{J}}\eta T}{2}\|\nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2 \\
&\quad - \frac{\eta T}{2}\mathbb{E}\left[\|\bar{\mathbf{g}}^{(\mathcal{S})}\|^2\right] + \frac{\mu\eta^2 T^2}{2}\mathbb{E}\left[\|\bar{\mathbf{g}}^{(\mathcal{S})}\|^2\right], \qquad (20)
\end{aligned}
$$

Under Assumption 5.5 where $1 - P_{\mathcal{J}} \geq P_{\mathcal{J}}$, and Assumption B.1 where gradient norm is bounded ($\|\nabla F\|^2 \leq G^2$), taking $T \geq \frac{\mu}{\eta^2}$, we can derive the lower bound for A_1:

$$
\begin{aligned}
A_1 &\geq \frac{P_{\mathcal{J}}\eta T}{2}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2 + \frac{\eta^2 T^2 - \mu T}{2}(\|\bar{\mathbf{G}}^{(\mathcal{S})}\|^2 + \sigma^2) + \frac{\eta T}{2}G^2 \\
&\underset{T \geq \frac{\mu}{\eta^2}}{\geq} \frac{P_{\mathcal{J}}\eta T}{2}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2 \qquad (21)
\end{aligned}
$$

Therefore, we obtain the lower bound for S in FU:

$$
S(\boldsymbol{w}^u) \geq \frac{P_{\mathcal{J}}\eta T}{2}\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^o) - \nabla F_{\mathcal{J}}(\boldsymbol{w}^o)\|^2 + \delta \qquad (22)
$$

□

F Proof of Theorem 5.10

Proof. By setting $\eta = \frac{1}{T}\sqrt{\frac{2}{\mu M}}$, we ensure the inequality in Equation (18) is satisfied. Given the fact that $\frac{1}{M} \leq \frac{1}{2\mu}$, it follows that $\eta \leq \frac{1}{\mu T}$. Therefore, the inequality in Equation (22) also holds, completing the proof. □

G Proof of Theorem 5.11

Proof. Starting with the local fairness metric $Q(\boldsymbol{w}^u)$, we have:

$$
\begin{aligned}
Q(\boldsymbol{w}^u) &= \sum_{i \notin \mathcal{J}} p'_i \left|\Delta f_i - \sum_{i \notin \mathcal{J}} p'_i \Delta f_i\right| \\
&= \sum_{i \notin \mathcal{J}} p'_i \left|\Delta f_i - V + F_{\mathcal{J}}(\boldsymbol{w}^*) - F^*_{-\mathcal{J}}\right| \\
&\geq \sum_{i \notin \mathcal{J}} p'_i \left|\Delta f_i\right| - V - \left(F_{\mathcal{J}}(\boldsymbol{w}^*) - F^*_{-\mathcal{J}}\right) \\
&\geq -2V + F^*_{-\mathcal{J}} - \sum_{i \notin \mathcal{J}} p'_i f_i(\boldsymbol{w_i}^*). \qquad (23)
\end{aligned}
$$

The last inequality arises from:

$$
\begin{aligned}
\sum_{i \notin \mathcal{J}} p'_i \left|\Delta f_i\right| &= \sum_{i \notin \mathcal{J}} p'_i \left|\mathbb{E}\left[f_i(\boldsymbol{w}^u)\right] - f_i(\boldsymbol{w}^*)\right| \\
&= \sum_{i \notin \mathcal{J}} p'_i \left|(f_i(\boldsymbol{w}^*) - f_i(\boldsymbol{w_i}^*)) - (\mathbb{E}\left[f_i(\boldsymbol{w}^u)\right] - f_i(\boldsymbol{w_i}^*))\right| \\
&\geq \sum_{i \notin \mathcal{J}} p'_i \left(f_i(\boldsymbol{w}^*) - f_i(\boldsymbol{w_i}^*)\right) - \sum_{i \notin \mathcal{J}} p'_i \left(\mathbb{E}\left[f_i(\boldsymbol{w}^u)\right] - f_i(\boldsymbol{w_i}^*)\right) \\
&\underset{(1)}{\geq} F_{-\mathcal{J}}(\boldsymbol{w}^*) - \sum_{i \notin \mathcal{J}} p'_i f_i(\boldsymbol{w_i}^*) - \left(\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^u)\right] - F^*_{\mathcal{J}}\right) \\
&= F_{-\mathcal{J}}(\boldsymbol{w}^*) - \sum_{i \notin \mathcal{J}} p'_i f_i(\boldsymbol{w_i}^*) - V. \qquad (24)
\end{aligned}
$$

The inequality (1) is justified because $\sum_{i \notin \mathcal{J}} p_i' f_i(\boldsymbol{w_i}^*) \leq F_{-\mathcal{J}}^*$, as $\min_{\boldsymbol{w}} F_{-\mathcal{J}}(\boldsymbol{w}) \geq \sum_{i \notin \mathcal{J}} p_i' \min_{\boldsymbol{w}} f_i(\boldsymbol{w})$.
Therefore,

$$Q(\boldsymbol{w}^u) + 2V(\boldsymbol{w}^u) \leq F_{-\mathcal{J}}^* - \sum_{i \notin \mathcal{J}} p_i' f_i(\boldsymbol{w_i}^*).$$

□

H Balancing Stability Unlearning Algorithm Analysis

H.1 Proof for Theorem B.4

Lemma H.1. *Lemma 1 of (Li et al., 2020b) Let Assumption 5.1 holds and $\eta_l \leq \frac{1}{4L}$, the remaining clients' contribution to unlearning for every round t gives:*

$$\mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,E)} - \boldsymbol{w}^{r*}\right\|^2\right] \leq (1 - \eta_l \mu)\mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\|^2\right] + \eta_l^2 C \tag{25}$$

where $C = \sigma^2 + 6L\Gamma + 8(E-1)^2\left(\zeta_{\mathcal{S}}^2 + (\beta_{\mathcal{S}}^2 + 1)G^2\right)$, $\Gamma = F_{-\mathcal{J}}^* - \sum_{i \in \mathcal{S}} \alpha_i F_i^*$ and $\boldsymbol{w}^{r*} = argmin_{\boldsymbol{w}} F_{-\mathcal{J}}(\boldsymbol{w})$ is the optimal unlearned model.

Proof. Suppose the FU process involves a total of T unlearning rounds, and within each round t, each participating client i engages in E local iterations. During local iterations, client i's model at iteration τ ($\tau \leq E$) of round t is denoted as $\boldsymbol{w}_i^{(t,\tau)}$. At the end of round t, the server aggregates to obtain the global model $\bar{\boldsymbol{w}}^{(t,E)}$ and updates the global model by gradient correction as $\bar{\boldsymbol{w}}^{(t+1,0)} = \bar{\boldsymbol{w}}^{(t,E)} - \eta_g \boldsymbol{g}_c^t$.
Then, we can express $\bar{\boldsymbol{w}}^{(t+1,0)} = \bar{\boldsymbol{w}}^{(t,0)} - \eta_l \boldsymbol{g}_{\mathcal{S}}^t - \eta_g \boldsymbol{g}_c^t$, and we have:

$$\begin{aligned}
\mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t+1,0)} - \boldsymbol{w}^{r*}\right\|^2\right] &= \mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \eta_l \boldsymbol{g}_{\mathcal{S}}^t - \eta_g \boldsymbol{g}_c^t - \boldsymbol{w}^{r*}\right\|^2\right] \\
&= \underbrace{\mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \eta_l \boldsymbol{g}_{\mathcal{S}}^t - \boldsymbol{w}^{r*}\right\|^2\right]}_{R^t} + \underbrace{\eta_g^2 \mathbb{E}\left[\left\|\boldsymbol{g}_c^t\right\|^2\right] - 2\eta_g \mathbb{E}\left[\left\langle \bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}, \boldsymbol{g}_c^t \right\rangle\right]}_{\Phi^t}
\end{aligned} \tag{26}$$

where R^t can be bounded by Lemma H.1.
Then, we will bound the second term Φ^t in Equation (26). By Cauchy-Schwarz inequality and AM-GM inequality, we have $-2\left\langle \boldsymbol{g}_c^t, \bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\rangle \leq \eta_g \left\|\boldsymbol{g}_c^t\right\|^2 + \frac{1}{\eta_g}\left\|\bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\|^2$
Thus,

$$\begin{aligned}
\Phi^t &\leq \mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\|^2\right] + 2\eta_g^2 \mathbb{E}\left[\left\|\boldsymbol{g}_c^t\right\|^2\right] \\
&\underset{\text{Lemma B.3}}{\leq} \mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\|^2\right] + 2\eta_g^2 \zeta'^2 \phi + 2\eta_g^2(\beta'^2 + 1)\phi \left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{t,0})\right\|^2 \\
&\underset{\text{Assumption B.1}}{\leq} \mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\|^2\right] + 2\eta_g^2 \zeta'^2 \phi + 2\eta_g^2 \phi \Omega G^2
\end{aligned} \tag{27}$$

where $\Omega = (\beta'^2 + 1)$, β'^2 indicates the data heterogeneity between remaining and unlearned clients.
Under Lemma H.1, taking Equation (27) into Equation (26) and letting $\Delta_t = \mathbb{E}\left[\left\|\bar{\boldsymbol{w}}^{(t,0)} - \boldsymbol{w}^{r*}\right\|^2\right]$, we have:

$$\Delta_{t+1} \leq (2 - \eta_l \mu)\Delta_t + \eta_l^2 B \tag{28}$$

where $B = \sigma^2 + 6L\Gamma + 8\left(\zeta_{\mathcal{S}}^2 + (\beta_{\mathcal{S}}^2 + 1)G^2\right)(E-1)^2 + 2\phi(\frac{\eta_g}{\eta_l})^2(\Omega\phi G^2 + \zeta'^2)$

Next, we will prove $\Delta_t \leq \frac{v}{\gamma + t}$ where $v = \max\left\{\frac{\beta^2 B}{\beta\mu - 4}, (\gamma+1)\Delta_1\right\}$. For a diminishing stepsize, $\eta_l = \frac{\beta}{2(t+\gamma)}$ for some $\beta > \frac{1}{\mu}$ and $\gamma > 0$ such that $\eta_l \leq \frac{1}{4L}$. We prove $\Delta_t \leq \frac{v}{\gamma+t}$ by induction.

Firstly, the definition of v ensures that it holds for $t = 1$. Assume the conclusion holds for some t, it follows that

$$\begin{aligned}
\Delta_{t+1} &\leq (2 - \eta_t \mu)\,\Delta_t + \eta_t^2 B \\
&\leq \left(2 - \frac{\beta\mu}{2(t+\gamma)}\right)\frac{v}{2(t+\gamma)} + \frac{\beta^2 B}{4(t+\gamma)^2} \\
&= \frac{4(t+\gamma)-4}{4(t+\gamma)^2} v + \left[\frac{\beta^2 B}{4(t+\gamma)^2} - \frac{\beta\mu - 4}{4(t+\gamma)^2} v\right] \\
&\leq \frac{v}{t+\gamma+1}
\end{aligned}$$

By the L-smoothness of $H(\cdot)$,

$$\begin{aligned}
&\mathbb{E}\left[H\left(\bar{\boldsymbol{w}}^{(t,0)}\right)\right] - H^* \\
&\quad = \mathbb{E}\left[H\left(\bar{\boldsymbol{w}}^{(t,0)}\right)\right] - H(\boldsymbol{w}^{r*}) + H(\boldsymbol{w}^{r*}) - H^* \\
&\quad \leq \frac{L}{2}\Delta_t + \frac{L}{2}\left\|\boldsymbol{w}^{r*} - \boldsymbol{w}^*\right\|^2 \\
&\quad \leq \frac{L}{2}\left(\frac{v}{\gamma + t} + \left\|\boldsymbol{w}^{r*} - \boldsymbol{w}^*\right\|^2\right)
\end{aligned}$$

□

H.2 PROOF FOR THEOREM B.7

To prove Lemma 5.6, we begin to introduce and prove the following additional lemmas:

Additional Lemmas

Lemma H.2. *Under Assumption B.2 and Assumption B.5, the expected norm of the gradient correction term at round t is bounded as follows:*

$$\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)})\right\|^2\right] \leq \phi\left(\zeta'^2 + (\beta'^2 + 1)\epsilon\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{t,0})\right\|^2\right)$$

, where $\phi = \lambda^2 P_{\mathcal{J}}^2(1 + \cos^2\theta)$. $\cos^2\theta$ represents $\cos^2\theta$ represents the similarity between the objectives of the remaining and unlearned clients.

Lemma H.3 (Per Round Unlearning)**.** *For each iteration t in the unlearning process:*

$$\begin{aligned}
&\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1}) - F_{-\mathcal{J}}(\boldsymbol{w}^t)\right] \\
&\leq \underbrace{\mathbb{E}\left[\left\langle -\eta_l \boldsymbol{g}_{\mathcal{S}}^t, \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\rangle\right] + \frac{L\eta_l^2}{2}\mathbb{E}\left[\left\|\boldsymbol{g}_{\mathcal{S}}^t\right\|^2\right]}_{A_1:\text{remaining clients training}} \\
&\quad + \underbrace{\mathbb{E}\left[\left\langle -\eta_g {\boldsymbol{g_c}}^t, \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\rangle\right] + \frac{L\eta_g^2}{2}\mathbb{E}\left[\left\|{\boldsymbol{g_c}}^t\right\|^2\right]}_{A_2:\text{global correction}}
\end{aligned}$$

Lemma H.4. *Assume Assumption B.1 holds. Given a set of clients $\mathcal{S}$, it follows that*

$$\mathbb{E}\left\|\boldsymbol{G}_{\mathcal{S}}^t - \boldsymbol{g}_{\mathcal{S}}^t\right\|^2 \leq \sum_{i\in\mathcal{S}} \alpha_i^2 \sigma_{i,t}^2,$$

where $\boldsymbol{G}_{\mathcal{S}}^t = \mathbb{E}\left[\boldsymbol{g}_{\mathcal{S}}^t\right]$.

Proof. (Theorem B.7)

Based on Lemma H.3, we can decomposed the convergence of unlearning verficiation $V = \mathbb{E}[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1}) - F_{-\mathcal{J}}(\boldsymbol{w}^t)]$ into two primary components A_1 and A_2:

$$\frac{1}{2}\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1}) - F_{-\mathcal{J}}(\boldsymbol{w}^t)\right] \leq A_1, \qquad \frac{1}{2}\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1}) - F_{-\mathcal{J}}(\boldsymbol{w}^t)\right] \leq A_2$$

These components A_1 represent the impact of training with the remaining clients, and A_2 represents the global correction.

Derivation for component A_1:

For A_1, it is related to unlearning with the remaining clients. By iterative derivation, we have

$$
\begin{aligned}
\frac{1}{2}\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1})\right] - F_{-\mathcal{J}}(\boldsymbol{w}^t) \leq A_1 &= \frac{\eta_l}{2}\mathbb{E}\left[\left\|\boldsymbol{g}_{\mathcal{S}}^t - \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2\right] - \frac{\eta_l}{2}\left\|\boldsymbol{G}_{\mathcal{S}}^t\right\|^2 - \frac{\eta_l}{2}\left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2 + \frac{L\eta_l^2}{2}\left\|\boldsymbol{g}_{\mathcal{S}}^t\right\|^2 \\
&= \frac{\eta_l}{2}\mathbb{E}\left[\left\|\boldsymbol{g}_{\mathcal{S}}^t - \nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2\right] + \frac{L\eta_l^2}{2}\mathbb{E}\left[\left\|\boldsymbol{G}_{\mathcal{S}}^t - \boldsymbol{g}_{\mathcal{S}}^t\right\|^2\right] \\
&\quad + \frac{\eta_l}{2}(\eta_l L - 1)\left\|\boldsymbol{G}_{\mathcal{S}}^t\right\|^2 - \frac{\eta_l}{2}\left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2 \\
&\leq_{\text{(Assumption 5.1 \& Lemma H.4)}} \frac{\eta_l}{2}\left(\zeta_{\mathcal{S}}^2 + \beta_{\mathcal{S}}^2\left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2\right) + \frac{L\eta_l^2}{2}\sigma^2 \\
&\quad + \frac{\eta_l}{2}(\eta_l L - 1)(\beta_{\mathcal{S}}^2 + 1)\left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2 + \frac{\eta_l}{2}(\eta_l L - 1)\zeta_{\mathcal{S}}^2
\end{aligned}
$$

, where $\sigma^2 = \sum_{i\in\mathcal{S}} \alpha_i^2\sigma_{i,t}^2$ by Lemma H.4.

Thus,

$$
\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1})\right] - F_{-\mathcal{J}}(\boldsymbol{w}^t) \leq \left(\eta_l^2 L(\beta_{\mathcal{S}}^2+1) - \frac{\eta_l}{2}\right)\left\|\nabla F_{-\mathcal{J}}(\boldsymbol{w}^t)\right\|^2 + \eta_l^2 L\left(\sigma^2 + \zeta_{\mathcal{S}}^2\right)
$$

Taking $\eta_l = \frac{1}{LT}$ and $T \geq 2\beta_{\mathcal{S}}^2 + 2$, we have:

$$
\begin{aligned}
\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \leq &\underbrace{\left(1 - \frac{1}{2LT} + \frac{\beta_{\mathcal{S}}^2+1}{LT^2}\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right)}_{D_1} \\
&+ \underbrace{2\frac{\sigma^2+\zeta_{\mathcal{S}}^2}{(T-2\beta_{\mathcal{S}}^2-2)}\left(1 - \left(1 - \frac{T-2\beta_{\mathcal{S}}^2-2}{2LT^2}\right)^T\right)}_{D_2}
\end{aligned}
\tag{29}
$$

Bounding D_2: We will employ the inequality $(1-a)^T \leq 1 - aT$ for $a \leq 1$. Here, $a = \frac{T-2\beta_{\mathcal{S}}^2-2}{2LT^2}$. Then, we verify that $a \leq 1$ by considering $2LT^2 - T + 2\beta_{\mathcal{S}}^2 + 2 \geq 0$.

If $L(\beta_{\mathcal{S}}^2+1) \geq \frac{1}{16}$, then, $a \leq 1$ holds for $T \geq 1$. If $L(\beta_{\mathcal{S}}^2+1) < \frac{1}{16}$, then, considering $T \geq \frac{1+\Delta}{4L}$, where $\Delta = \sqrt{1 - 16L(\beta_{\mathcal{S}}^2+1)}$, we have $a \leq 1$.

Considering $T \geq \frac{1+\Delta}{4L}$ and we obtain $\left(1 - \frac{T-2\beta_{\mathcal{S}}^2-2}{2LT^2}\right)^T \geq 1 - \frac{T-2\beta_{\mathcal{S}}^2-2}{2LT}$, and

$$
\frac{1}{(T-2\beta_{\mathcal{S}}^2-2)}\left(1 - \left(1 - \frac{T-2\beta_{\mathcal{S}}^2-2}{2LT^2}\right)^T\right) \leq \frac{1}{2LT}
\tag{30}
$$

By taking $T \geq \max\{2\beta_{\mathcal{S}}^2+2, \frac{1+\Delta}{4L}\}$ with $\Delta = \sqrt{\max\{0, 1-16L(\beta_{\mathcal{S}}^2+1)\}}$, and integrating the bounds derived in Equation (30) into Equation (29), we have:

$$
\begin{aligned}
\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) &\leq \left(1 - \frac{1}{2LT} + \frac{\beta_{\mathcal{S}}^2+1}{LT^2}\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) \\
&\quad + \frac{\sigma^2+\zeta_{\mathcal{S}}^2}{LT} \\
&= 2\chi_1.
\end{aligned}
\tag{31}
$$

Here, χ_1 is defined as

$$
\frac{1}{2}\left(1 - \frac{1}{2LT} + \frac{\beta_{\mathcal{S}}^2+1}{LT^2}\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) + \frac{(\sigma^2+\zeta_{\mathcal{S}}^2)}{2LT}
$$

.

Derivation for component A_2:

By Lemma H.3:

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^{t+1}) - F_{-\mathcal{J}}(\boldsymbol{w}^{t})\right] \leq A_2 = \frac{L\eta_g^2}{2}\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)})\right\|^2\right] - \eta_g\mathbb{E}\left[\left\langle \boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)}), \nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\rangle\right] \\
&\quad = \frac{1}{2}(L\eta_g^2 - \eta_g)\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)})\right\|^2\right] + \frac{\eta_g}{2}\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)}) - \nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2\right] - \frac{\eta_g}{2}\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2 \\
&\quad \leq_{(1)} \frac{1}{2}L\eta_g^2\zeta''^2 + \frac{1}{2}\left(L\eta_g^2(\beta''^2+1) - 2\eta_g\right)\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2
\end{aligned}
$$

where $\zeta''^2 = \phi\zeta'^2, \beta''^2 = \phi\epsilon\beta'^2 + \phi\epsilon + 1$.

The last inequality holds for:

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)}) - \nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2\right] &\leq \mathbb{E}\left[\left\|\boldsymbol{g_c}(\bar{\boldsymbol{w}}^{(t,E)})\right\|^2\right] + \left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2 \\
&\underset{\text{Lemma H.2}}{\leq} \phi\left(\zeta'^2 + (\beta'^2+1)\epsilon\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{t,0})\right\|^2\right) + \left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2 \\
&\leq \phi\zeta'^2 + \left(\phi\epsilon\beta'^2 + \phi\epsilon + 1\right)\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{t,0})\right\|^2 \\
&\leq \zeta''^2 + \beta''^2\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{(t,0)})\right\|^2
\end{aligned}
$$

where $\zeta''^2 = \phi\zeta'^2, \beta''^2 = \phi\epsilon\beta'^2 + \phi\epsilon + 1$.

Choosing $\eta_g = \frac{1}{LT}$, similar to previous derivation for component A_1, we have:

$$
\begin{aligned}
\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \leq& \underbrace{\left(1 - \frac{2}{LT} + \frac{\beta''^2+1}{LT^2}\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right)}_{B_1} \\
&+ \underbrace{\frac{\zeta''^2\left(1 - \left(1 - \frac{(2T-\beta''^2-1)}{LT^2}\right)^T\right)}{2T - \beta''^2 - 1}}_{B_2}
\end{aligned}
\tag{32}
$$

Bounding B_2: We will employ the inequality $(1-a)^T \leq 1 - aT$ for $a \leq 1$. Here, $a = \frac{2T-\beta''^2-1}{LT^2}$. Then, we verify that $a \leq 1$ by considering $LT^2 - 2T + \beta''^2 + 1 \geq 0$.

If $L(\beta''^2+1) \geq 1$, then, $a \leq 1$ holds for $T \geq 1$. If $L(\beta''^2+1) < 1$, then, considering $T \geq \frac{1+\Delta'}{L}$, where $\Delta' = \sqrt{1 - L(\beta''^2+1)}$, we have $a \leq 1$.

Considering $T \geq \frac{1+\Delta'}{L}$ and we obtain $\left(1 - \frac{2T-\beta''^2-1}{LT^2}\right)^T \geq 1 - \frac{2T-\beta''^2-1}{LT}$, and

$$
\frac{1 - \left(1 - \frac{2T-\beta''^2-1}{LT^2}\right)^T}{2T - \beta''^2 - 1} \leq \frac{1}{LT}
\tag{33}
$$

By taking $T \geq \max\{\frac{1}{2}(\beta''^2+1), \frac{1+\Delta'}{L}\}$, where $\Delta' = \sqrt{\max\{0, 1 - L(\beta''^2+1)\}}$, and integrating the bounds derived in Equation (33) into Equation (32), we have:

$$
\begin{aligned}
\mathbb{E}\left[F_{-\mathcal{J}}(\boldsymbol{w}^T)\right] - F_{-\mathcal{J}}(\boldsymbol{w}^{r*}) \leq& \left(1 - \frac{2}{LT} + \frac{\beta''^2+1}{LT^2}\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o) - F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right) \\
&+ \frac{\zeta''^2}{LT} \\
=& 2\chi_2.
\end{aligned}
\tag{34}
$$

Here, χ_2 is defined as

$$\frac{1}{2}\left(1-\frac{2}{LT}+\frac{\beta''^2+1}{LT^2}\right)^T\left(F_{-\mathcal{J}}(\boldsymbol{w}^o)-F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\right)+\frac{\zeta''^2}{2LT}$$

.

$\square$

H.3 Proof for Lemma B.3

Proof. Recall that the gradient correction $\boldsymbol{g}_c^t=\boldsymbol{g}_c(\bar{\boldsymbol{w}}^{(t,E)})$ at each round t after local E epochs as: $\boldsymbol{g}_c^t=\boldsymbol{h}^t-\text{Proj}_{\boldsymbol{g}_{-\mathcal{J}}}\boldsymbol{h}^t$, where $\boldsymbol{h}^t=\lambda(1-P_{\mathcal{J}})\boldsymbol{g}_{-\mathcal{J}}^t+\lambda P_{\mathcal{J}}\hat{\boldsymbol{g}}_{\mathcal{J}}^t$

Thus, under stochastic gradient descent $\boldsymbol{g}_{\mathcal{S}}^t$ from the subset of remaining client $\mathcal{S}$:

$$\boldsymbol{g_c}^t=\lambda P_{\mathcal{J}}\left(\hat{\boldsymbol{g}}_{\mathcal{J}}^t-\cos\theta\frac{\|\hat{\boldsymbol{g}}_{\mathcal{J}}^t\|}{\|\boldsymbol{g}_{\mathcal{S}}^t\|}\boldsymbol{g}_{\mathcal{S}}^t\right)$$

The expected norm of $\boldsymbol{g_c}^t$ can be bounded as:

$$\mathbb{E}\left[\left\|\boldsymbol{g_c}^t\right\|^2\right]\leq\underbrace{\lambda^2P_{\mathcal{J}}^2(1+\cos^2\theta)}_{\phi}\,\mathbb{E}\left[\left\|\hat{\boldsymbol{g}}_{\mathcal{J}}^t\right\|^2\right],\tag{35}$$

Therefore, by Equation (35) and Assumption B.2, we obtain:

$$\mathbb{E}\left[\left\|\boldsymbol{g_c}^t\right\|^2\right]\leq\phi\left(\zeta'^2+(\beta'^2+1)\left\|\nabla F_{-\mathcal{J}}(\bar{\boldsymbol{w}}^{t,E})\right\|^2\right)$$

$\square$

H.4 Proof for Theorem B.10

Let $G=F_{-\mathcal{J}}(\boldsymbol{w})+\boldsymbol{\lambda}^\top\boldsymbol{r}(\boldsymbol{w})$, and ($\bar{\boldsymbol{w}}$ and $\bar{\boldsymbol{\lambda}}$) is a ν-approximate saddle point of G.

$$\begin{aligned}F_{-\mathcal{J}}(\bar{\boldsymbol{w}})+\Lambda\max_{z\in Z}\mathbf{r}_z(\bar{\boldsymbol{w}})_+-\nu&\leq F_{-\mathcal{J}}(\bar{w})+\overline{\boldsymbol{\lambda}}^T\mathbf{r}(\bar{w})\\&=G(\bar{w},\overline{\boldsymbol{\lambda}})\\&\leq\min_w G(w,\overline{\boldsymbol{\lambda}})+\nu\\&\leq G\left(\boldsymbol{w}^o,\overline{\boldsymbol{\lambda}}\right)+\nu\\&=F_{-\mathcal{J}}\left(\boldsymbol{w}^o\right)+\nu\\&\leq F_{-\mathcal{J}}(\boldsymbol{w}^o)+\nu\end{aligned}$$

Hence,

$$\begin{aligned}\max_{z\in Z}\mathbf{r}_z(\bar{w})_+&\leq\frac{1}{\Lambda}\left(F_{-\mathcal{J}}(\boldsymbol{w}^o)-F_{-\mathcal{J}}(\bar{\boldsymbol{w}})+2\nu\right)\\&=\epsilon\end{aligned}$$

We can present ν as $\nu=\frac{\Lambda\epsilon+F_{-\mathcal{J}}(\bar{\boldsymbol{w}})-F_{-\mathcal{J}}(\boldsymbol{w}^o)}{2}$.

Similarly, we can obtain $F_{-\mathcal{J}}(\bar{\boldsymbol{w}})-F_{-\mathcal{J}}(\boldsymbol{w}^{r*})\leq2\nu$. By definition of unlearning verification, $V(\bar{\boldsymbol{w}})\leq2\nu$.

That requires $\nu\leq\epsilon\Lambda+F_{-\mathcal{J}}(\bar{\boldsymbol{w}})-F_{-\mathcal{J}}(\boldsymbol{w}^o)$, which is equivalent to $\epsilon\geq\frac{F_{-\mathcal{J}}(\boldsymbol{w}^o)-F_{-\mathcal{J}}(\boldsymbol{w}^{r*})}{\Lambda}$.

I Discussion

In this section, we demonstrate how our proposed mechanism adapts to these common unlearning algorithms for enhancing the adaptability and robustness of existing methodologies. In the context of the rapid retaining method Wu et al. (2020a) and knowledge distillation Wu et al. (2022), our proposed mechanism motivates the incorporation of a control parameter λ to handle stability during unlearning. Specifically, it could introduce λ multiplying the second term of the right-hand side formulation in Wu et al. (2020a, Equation 3) or Wu et al. (2020a, Algorithm 1, Line 1). That allows for balancing between stability and unlearning effectiveness. Specifically, if $\lambda=1$, the case is identical to their formulation. However, if $\lambda = 0$, it would revert to the original FL training, prioritizing stability over unlearning effectiveness.