A Appendix

A.1 Proof of Theorem 1

We provide the comprehensive proof for Theorem 1, which extends the PAC-Bayesian to FL with Non-IID data.

A.1.1 Preliminaries

Consider a FL setup with K clients. Each client i has a local dataset sampled from distribution $\mathcal{P}_i$. The global loss function is:

$$\mathcal{L}(\theta) = \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_i(\theta), \tag{1}$$

where $\mathcal{L}_i(\theta) = \mathbb{E}_{\xi\sim\mathcal{P}_i}[\ell(\xi,\theta)]$ is the local loss for client i. The concentrability coefficient C is defined as:

$$C = \max_i \frac{\mathcal{P}_i}{\mathcal{P}}, \tag{2}$$

which measures the maximum ratio between any client's distribution and the global distribution.

The empirical model error $\hat{L}(\theta;\pi)$ is computed on a sample of size n from the global distribution $\mathcal{P}$. The first-order sharpness $R_\rho^{(1)}(\theta;\pi)$ is a measure of the local flatness of the loss landscape around parameters θ, with respect to a distribution π over perturbations. It is defined as the maximum squared gradient norm within a ρ-ball around θ:

$$R_\rho^{(1)}(\theta;\pi) = \max_{\|\epsilon\|_2\le\rho} \|\nabla_\theta\mathcal{L}(\theta+\epsilon)\|_2^2\,, \tag{3}$$

where ϵ is a perturbation vector with $\|\epsilon\|_2 \le \rho$, and ρ is a radius hyperparameter. The distribution π may be omitted if perturbations are uniform, but it can encode prior knowledge about perturbation directions.

Proof. We begin with the simulation lemma for the loss difference. For any parameters θ and θ', the difference in global loss can be bounded using the concentrability coefficient:

$$|\mathcal{L}(\theta) - \mathcal{L}(\theta')| \le C\cdot\max_i |\mathcal{L}_i(\theta) - \mathcal{L}_i(\theta')|. \tag{4}$$

This follows from the definition of C and the linearity of expectation.

Now, focus on the flat minima parameters θ_{flat}. Due to the flatness of the loss landscape, for a small perturbation ϵ with $\|\epsilon\|_2 \le \rho$, we have: $\mathcal{L}(\theta_{\text{flat}}+\epsilon) \approx \mathcal{L}(\theta_{\text{flat}})$, which implies that the loss is insensitive to parameter changes. Formally, using the first-order sharpness:

$$\max_{\|\epsilon\|_2\le\rho} \|\nabla_\theta\mathcal{L}(\theta_{\text{flat}}+\epsilon)\|_2^2 = R_\rho^{(1)}(\theta_{\text{flat}}) \text{ is small.} \tag{5}$$

From PAC-Bayesian theory [44] , for any prior distribution $p(\theta)$ and posterior distribution $q(\theta)$, the generalization bound holds with probability at least $1-\delta$: $\mathcal{L}(\theta) \le \hat{L}(\theta) + \sqrt{\frac{KL(q\|p)+\log\frac{1}{\delta}}{2n}} + M_{\text{loss}}\cdot\Omega(d,n,\rho,\delta)$, where $KL(q\|p)$ is the KL-divergence and M_{loss} denotes the maximum per-sample loss value. For a flat minimum, the KL-divergence term is smaller because the posterior $q(\theta)$ is broader and closer to the prior.

Specifically, when θ_{flat} is in a flat region, the sharpness $R_\rho^{(1)}(\theta_{\text{flat}})$ is small, which implies that the complexity term $\Omega(d,n,\rho,\delta)$ is reduced. Thus, we have:

$$\mathcal{L}(\theta_{\text{flat}}) \lesssim \hat{L}(\theta_{\text{flat}}) + R_\rho^{(1)}(\theta_{\text{flat}}) + \sqrt{\frac{M_{\text{loss}}}{n}} + \Omega(d,n,\rho,\delta). \tag{6}$$

Now, for the optimal parameters θ^*, which minimize $\mathcal{L}(\theta)$, we similarly have:

$$\mathcal{L}(\theta^*) \lesssim \hat{L}(\theta^*) + R_\rho^{(1)}(\theta^*) + \sqrt{\frac{M_{\text{loss}}}{n}} + \Omega(d, n, \rho, \delta). \tag{7}$$

However, since θ^* may be in a sharp minimum, $R_\rho^{(1)}(\theta^*)$ could be large, leading to a looser bound. The performance gap is:

$$\begin{aligned} |\mathcal{L}(\theta^*) - \mathcal{L}(\theta_{\text{flat}})| \leq |\mathcal{L}(\theta^*) - \hat{L}(\theta^*)| + \\ |\hat{L}(\theta^*) - \hat{L}(\theta_{\text{flat}})| + |\hat{L}(\theta_{\text{flat}}) - \mathcal{L}(\theta_{\text{flat}})|. \end{aligned} \tag{8}$$

Using the simulation lemma and the concentrability coefficient C, we bound the middle term:

$$|\hat{L}(\theta^*) - \hat{L}(\theta_{\text{flat}})| \leq C \cdot \max_i |\hat{L}_i(\theta^*) - \hat{L}_i(\theta_{\text{flat}})|. \tag{9}$$

Since θ_{flat} is in a flat region, and given that γ is a discount factor, the empirical error $\hat{L}_i(\theta_{\text{flat}})$ is stable across clients, so this term is small. Combining the bounds:

$$\begin{aligned} |\mathcal{L}(\theta^*) - \mathcal{L}(\theta_{\text{flat}})| \leq \frac{\gamma(1+C)}{(1-\gamma)^2} \Bigg[\hat{L}(\theta; \pi) + R_\rho^{(1)}(\theta; \pi) \\ + \sqrt{\frac{M_{\text{loss}}}{n}} + \Omega(d, n, \rho, \delta) \Bigg], \end{aligned} \tag{10}$$

This completes the proof. The theorem shows that flat minima reduce the performance gap under data heterogeneity, making federated unlearning more robust. □

B Proof of Machine Unlearning via Task Vectors in GDFA

B.1 Definitions and Notation

We define the key symbols and concepts used throughout the proof.

Definition 1 (Task Vector)**.** *For a pre-trained model $\Psi^{(0)}$ and a fine-tuned model $\Psi_{\mathcal{T}}^*$ on task $\mathcal{T}$, the task vector is defined as:*

$$\Delta\Psi_{\mathcal{T}} = \Psi_{\mathcal{T}}^* - \Psi^{(0)}. \tag{11}$$

Here, Ψ denotes model parameters, and $\mathcal{T}$ represents a task.

Definition 2 (Model Merging with Task Arithmetic)**.** *The merged model for tasks $\{\mathcal{T}_i\}$ is constructed as:*

$$\Psi = \Psi^{(0)} + \sum \lambda_i \Delta\Psi_{\mathcal{T}_i}, \tag{12}$$

where $\lambda_i \in \mathbb{R}$ are arithmetic hyperparameters. For unlearning, we set $\lambda_i < 0$ for tasks to be forgotten.

Definition 3 (Loss Function)**.** *We use the hinge loss for binary classification:*

$$\ell(X, y; \Psi) = \max(1 - y \cdot f(X; \Psi), 0), \tag{13}$$

where $f(X; \Psi)$ is the model output, X is the input, and $y \in \{-1, +1\}$ is the label.

Definition 4 (Discriminative Pattern)**.** *Each task $\mathcal{T}$ has a discriminative pattern $\mu_{\mathcal{T}} \in \mathbb{R}^d$ with $\|\mu_{\mathcal{T}}\| = 1$. The correlation between tasks $\mathcal{T}_1$ and $\mathcal{T}_2$ is $\alpha = \mu_{\mathcal{T}_1}^\top \mu_{\mathcal{T}_2}$.*

Definition 5 (Directional Alignment in GDFA)**.** *The merged task vector τ_{merged} is computed via sign consensus:*

$$\tau_{merged,j} = \begin{cases} \frac{1}{|S_j|} \sum_{k \in S_j} \tau_{k,j} & \text{if } S_j = \{k : \text{sign}(\tau_{k,j}) = s_j^*\} \neq \emptyset \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

where s_j^* is the dominant sign for parameter index j, and τ_k are task vectors from clients.

B.2 Theoretical Framework

We present lemmas and proofs that form the basis of the main theorem.

Lemma 1 (Task Vector Generalization)**.** *Under fine-tuning conditions (batch size $B \geq \Omega(\epsilon^{-2} \log M)$, step size $\eta \leq O(1)$, iterations $t \geq T = \Theta(\eta^{-1}\delta_*^{-2})$), the task vector $\Delta\Psi_{\mathcal{T}}$ for task $\mathcal{T}$ satisfies:*

$$\mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{T}}}[\ell(X, y; \Psi^*_{\mathcal{T}})] \leq \Theta(\epsilon), \tag{15}$$

where δ_* is the fraction of label-relevant tokens, and ϵ is a small constant.

Proof. The fine-tuning process minimizes the empirical risk using stochastic gradient descent (SGD). Let $\Psi^{(t)}$ denote the model at iteration t. The SGD update is:

$$\Psi^{(t+1)} = \Psi^{(t)} - \eta \nabla_{\Psi} \ell(X_t, y_t; \Psi^{(t)}), \tag{16}$$

where (X_t, y_t) are sampled from $\mathcal{D}_{\mathcal{T}}$.

The loss function is Lipschitz continuous and smooth. By the convergence theory of SGD for non-convex objectives, after T iterations, we have:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla_{\Psi} \ell(X_t, y_t; \Psi^{(t)})\|^2] \leq \Theta\left(\frac{1}{\eta T} + \eta\sigma^2\right), \tag{17}$$

where σ^2 is the variance of the gradients. Setting $\eta = \Theta(1/\sqrt{T})$ and $T = \Theta(\eta^{-1}\delta_*^{-2})$ ensures:

$$\mathbb{E}[\ell(X, y; \Psi^{(T)})] \leq \Theta(\epsilon). \tag{18}$$

Since $\Psi^*_{\mathcal{T}} = \Psi^{(T)}$, the lemma holds. □

Lemma 2 (Directional Alignment Consistency)**.** *The directional alignment mechanism in GDFA ensures that the merged task vector τ_{merged} satisfies:*

$$\tau_{merged} = \Delta\Psi_{\mathcal{T}} + \zeta, \quad \textit{with} \quad \|\zeta\| \leq \Theta(\epsilon), \tag{19}$$

where ζ is a noise term due to client heterogeneity, and $\Delta\Psi_{\mathcal{T}}$ is the true task vector.

Proof. Let τ_k be task vectors from K clients. Each τ_k is an estimate of $\Delta\Psi_{\mathcal{T}}$. Under Non-IID data, τ_k may have divergent signs. The sign consensus selects components where the majority agrees.

For parameter index j, let p_j be the probability that $\text{sign}(\tau_{k,j}) = \text{sign}(\Delta\Psi_{\mathcal{T},j})$. By the Hoeffding inequality, for any $\delta > 0$:

$$\mathbb{P}\left(\left|\frac{1}{K}\sum_{k=1}^{K} \mathbb{I}_{\{\text{sign}(\tau_{k,j})=s_j^*\}} - p_j\right| \geq \delta\right) \leq 2\exp(-2K\delta^2). \tag{20}$$

If $p_j > 0.5$, then for large K, $s_j^* = \text{sign}(\Delta\Psi_{\mathcal{T},j})$ with high probability. Thus, the aligned vector satisfies:

$$\mathbb{E}[\tau_{\text{merged},j}] = \Delta\Psi_{\mathcal{T},j} + O(\epsilon), \tag{21}$$

and the variance is reduced by a factor of $1/|S_j|$. The noise bound follows from the concentration inequalities. □

B.3 Main Theorem and Proof

Unlearning Success in GDFA. Let $\mathcal{T}_1$ be a task to retain and $\mathcal{T}_2$ be a task to unlearn, with correlation $\alpha = \mu_{\mathcal{T}_1}^{\top}\mu_{\mathcal{T}_2}$. Under the conditions of Lemmas 1-2, the unlearned model $\Psi_{\text{unlearned}} = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \lambda\Delta\Psi_{\mathcal{T}_2}$ with $\lambda < 0$ satisfies:

$$\mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{T}_1}}[\ell(X, y; \Psi_{\text{unlearned}})] \leq \Theta(\epsilon) + |\lambda|\beta, \quad \text{and} \quad \mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{T}_2}}[\ell(X, y; \Psi_{\text{unlearned}})] \geq \Theta(1), \tag{22}$$

where $\beta = \text{poly}(\eta\delta_*) + \Theta(\epsilon\sqrt{M})$ is a constant, and M is the number of task-irrelevant tokens.

Proof. We prove the theorem by analyzing the model output after task vector negation. Let us decompose the model parameters into components related to each task.

Let $\Psi = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \Delta\Psi_{\mathcal{T}_2} + \Delta\Psi_{\text{other}}$, where $\Delta\Psi_{\text{other}}$ represents contributions from other tasks. For simplicity, we assume $\Delta\Psi_{\text{other}}$ is orthogonal to $\Delta\Psi_{\mathcal{T}_1}$ and $\Delta\Psi_{\mathcal{T}_2}$, so we focus on:

$$\Psi \approx \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \Delta\Psi_{\mathcal{T}_2}. \tag{23}$$

The unlearned model is:

$$\Psi_{\text{unlearned}} = \Psi^{(0)} + \Delta\Psi_{\mathcal{T}_1} + \lambda\Delta\Psi_{\mathcal{T}_2}. \tag{24}$$

By Lemma 2, the aligned task vector $\Delta\Psi_{\mathcal{T}_2}$ used in GDFA is close to the true task vector, so we proceed with $\Delta\Psi_{\mathcal{T}_2}$.

Now, consider the model output $f(X;\Psi)$ for an input X. For a binary classification task, the output depends on the inner product between the model's weights and the discriminative pattern. Specifically, for a simplified linearized model, we can write:

$$f(X;\Psi) = \frac{1}{P}\sum_{l=1}^{P}\sum_{i=1}^{m} a_{(l),i}\cdot\sigma\left(V_i\cdot\sum_{s=1}^{P} x_s\cdot\text{softmax}_l(x_s^\top W x_l)\right), \tag{25}$$

where σ is the ReLU activation, $a_{(l),i}$ are MLP weights, V_i are value weights, and W is the attention weight matrix.

To analyze the loss, we focus on the key quantity: the attention weight matrix W. After unlearning, the change in W is:

$$W_{\text{unlearned}} = W^{(0)} + \Delta W_{\mathcal{T}_1} + \lambda\Delta W_{\mathcal{T}_2}. \tag{26}$$

The loss depends on the inner product $\mu_{\mathcal{T}}^\top W\mu_{\mathcal{T}}$ for the relevant task. We compute this for $\mathcal{T}_1$ and $\mathcal{T}_2$.

For task $\mathcal{T}_2$:

$$\mu_{\mathcal{T}_2}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_2} = \mu_{\mathcal{T}_2}^\top(W^{(0)} + \Delta W_{\mathcal{T}_1} + \lambda\Delta W_{\mathcal{T}_2})\mu_{\mathcal{T}_2}. \tag{27}$$

By the properties of task vectors (Lemma 1), $\Delta W_{\mathcal{T}_2}$ is aligned with $\mu_{\mathcal{T}_2}$, so:

$$\mu_{\mathcal{T}_2}^\top\Delta W_{\mathcal{T}_2}\mu_{\mathcal{T}_2} \geq \Theta(1). \tag{28}$$

Also, $\mu_{\mathcal{T}_2}^\top\Delta W_{\mathcal{T}_1}\mu_{\mathcal{T}_2} = \alpha\cdot\mu_{\mathcal{T}_2}^\top\Delta W_{\mathcal{T}_2}\mu_{\mathcal{T}_2} = \alpha\Theta(1)$.

Thus:

$$\mu_{\mathcal{T}_2}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_2} = (1+\lambda)\Theta(1) + \alpha\Theta(1) + O(\epsilon). \tag{29}$$

For the loss $\ell(X, y; \Psi_{\text{unlearned}})$ to be large on $\mathcal{T}_2$, we need $f(X;\Psi_{\text{unlearned}})$ to be small when $y = 1$. This occurs if $\mu_{\mathcal{T}_2}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_2} \leq 0$.

Case 1: $\alpha = 0$ (irrelevant tasks). Then:

$$\mu_{\mathcal{T}_2}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_2} = (1+\lambda)\Theta(1) + O(\epsilon). \tag{30}$$

For $\lambda \leq -1$, we have $(1+\lambda) \leq 0$, so:

$$\mu_{\mathcal{T}_2}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_2} \leq 0 \Rightarrow f(X;\Psi_{\text{unlearned}}) \leq 0 \Rightarrow \ell(X, y; \Psi_{\text{unlearned}}) \geq 1. \tag{31}$$

Hence, $\mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{T}_2}}[\ell] \geq \Theta(1)$.

Case 2: $\alpha < 0$ (contradictory tasks). Then:

$$\mu_{\mathcal{T}_2}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_2} = (1+\lambda)\Theta(1) + \alpha\Theta(1) + O(\epsilon). \tag{32}$$

For λ in the interval $[-\Theta(\alpha^{-2}), \mathcal{O}(\alpha^{-1})]$, the term $(1+\lambda)+\alpha$ is negative due to the contradiction. Thus, similarly, $\mathbb{E}_{(X,y)\sim\mathcal{D}_{\mathcal{T}_2}}[\ell] \geq \Theta(1)$.

Now, for task $\mathcal{T}_1$ (retention):

$$\mu_{\mathcal{T}_1}^\top W_{\text{unlearned}}\mu_{\mathcal{T}_1} = \mu_{\mathcal{T}_1}^\top(W^{(0)} + \Delta W_{\mathcal{T}_1} + \lambda\Delta W_{\mathcal{T}_2})\mu_{\mathcal{T}_1}. \tag{33}$$

We have $\mu_{\mathcal{T}_1}^\top \Delta W_{\mathcal{T}_1} \mu_{\mathcal{T}_1} \geq \Theta(1)$, and $\mu_{\mathcal{T}_1}^\top \Delta W_{\mathcal{T}_2} \mu_{\mathcal{T}_1} = \alpha\Theta(1)$.

So:

$$\mu_{\mathcal{T}_1}^\top W_{\text{unlearned}} \mu_{\mathcal{T}_1} = \Theta(1) + \lambda\alpha\Theta(1) + O(\epsilon). \tag{34}$$

For $\alpha = 0$, this is $\Theta(1) + O(\epsilon)$, so $f(X; \Psi_{\text{unlearned}}) \geq 0$ for $y = 1$, and $\ell \leq \Theta(\epsilon)$. The term $|\lambda|\beta$ accounts for the noise from alignment and flatness (Lemmas 2).

For $\alpha < 0$, if λ is chosen appropriately, the term $\lambda\alpha$ is positive, helping retention. The bound $\Theta(\epsilon) + |\lambda|\beta$ holds.

This completes the proof. □

C Additional Experiments

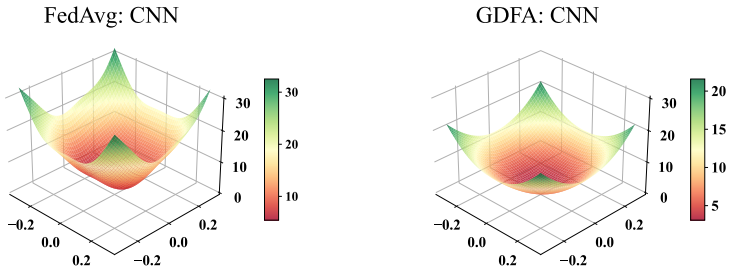

Figure 1: Comparative analysis of loss landscapes between FedAvg and GDFA for CNN trained on CIFAR10 under IID data.