Forgetting Through Transforming: Enabling Federated Unlearning via Class-Aware Representation Transformation

Supplementary Material

A. Experimental Setup Details

Datasets and Network Architecture. We conduct extensive experiments on the following four real-world datasets: CIFAR10 [24], CIFAR100 [24], FMNIST [46], EuroSAT [21]. For the CIFAR10, FMNIST, and EuroSAT datasets, we employ the widely used ResNet18 as the default network architecture [20]. For CIFAR100, we opt for the more powerful ResNet50 as the network architecture.

Baselines. We adopt eight FU methods as our baselines:

Retrain: We obtain the retrained model by training from scratch with only the remaining data. Thus, the retrained model is the optimal unlearning model.

Fine-tune: We fine-tune the original global model on the remaining data with a few training epochs.

Gradient-ascent [45]: use reverse gradient ascent to fine-tune the original model on the unlearning data.

FUDP [44]: prune the most relevant channels associated with the unlearning data, followed by fine-tuning with the remaining data.

FUMD [51]: utilize a randomly initialized degradation model to erase the existence of the unlearning data.

VeriFi [10]: erase the contribution of the leaving data from the global model by scaling up/down uploaded updates.

FedAU [14]: employ an auxiliary unlearning classifier combined with linear operations to facilitate unlearning.

FedOSD [36]: design an unlearning cross-entropy loss to overcome the convergence issue of the gradient ascent and achieve an orthogonal steepest descent model direction for unlearning.

Federated Learning Setting and Details. In our experiments, we set the total number of clients at 20, with all clients selected per round. The default Dirichlet coefficient δ=0.5 for the Non-IID scenario. We use SGD [1] as the optimizer during local unlearning training on each client, running one local epoch per round. The batch and the representation size are set to 256 and 512. The learning rate is fixed at 0.01 for both the original model training and unlearning (except for CIFAR100, where the original model adopted a cosine-annealed learning rate starting at 0.1).

Metrics. We evaluate erasing guarantees using two key metrics: (1) the accuracy of the unlearning model on the unlearning data, and (2) the attack success rate (ASR) obtained via membership inference attack (MIA) [39]. Lower accuracy and ASR values on the unlearning data indicate more effective erasure, with zero being the ideal value. We assess model utility preservation by the metric: the unlearning model's accuracy on the remaining data, where higher accuracy represents better preservation of model utility. We quantify unlearning efficiency by the metric: the processing time from the original model into the final unlearning model, with shorter times indicating greater efficiency.

B. Theoretical Analysis

We present a comprehensive theoretical analysis of the Federated Unlearning via Class-aware Representation Transformation (FUCRT) framework. Our analysis establishes formal guarantees for the convergence and unlearning properties of the proposed method, demonstrating that representation transformation provides a principled approach to federated unlearning with quantifiable performance bounds.

B.1. Notation and Preliminaries

Let us establish the following notation:

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$: The complete dataset with N samples
- $\mathcal{C}$: The set of all classes
- $\mathcal{C}_{\mathcal{F}} \subseteq \mathcal{C}$: The set of classes to be forgotten
- $\mathcal{S} = \{(x_i, y_i) \in \mathcal{D} \mid y_i \in \mathcal{C}_{\mathcal{F}}\}$: The unlearning dataset
- $\mathcal{R} = \mathcal{D} \setminus \mathcal{S}$: The remaining dataset
- θ^0: The original model trained on $\mathcal{D}$
- θ^*: The retrained model trained only on $\mathcal{R}$
- θ^R: The FUCRT unlearned model
- $f_\theta : \mathcal{X} \to \mathbb{R}^d$: The representation function mapping inputs to d-dimensional embeddings
- $g_\theta : \mathbb{R}^d \to \Delta^{|\mathcal{C}|-1}$: The classification function mapping representations to the probability simplex
- $p_\theta(y|x) = [g_\theta \circ f_\theta](x)_y$: The predicted probability of class y given input x
- $TF : \mathcal{S} \to \mathcal{C} \setminus \mathcal{C}_{\mathcal{F}}$: The transformation function mapping unlearning samples to target classes

B.2. Definition

Definition 1 (Representation-Based Unlearning) *A model θ^R achieves (ε, δ)-representation-based unlearning if and only if the following conditions are satisfied:*

1. ***Representation Convergence**: For any unlearning sample $x_u \in \mathcal{S}$ with transformation target $TF(x_u)$:*

$$\inf_{x_r \in \mathcal{R}: y_r = TF(x_u)} \|f_{\theta^R}(x_u) - f_{\theta^R}(x_r)\|_2 \leq \delta$$

2. ***Erasure Guarantee**: The model no longer correctly identifies unlearning data according to original labels:*

$$\sup_{(x,y)\in\mathcal{S}} p_{\theta^R}(y|x) \leq \varepsilon$$

3. ***Utility Preservation**: The predictions of θ^R on remaining data approximate those of the original model θ^0:*

$$\sup_{x \in \mathcal{R}} \|p_{\theta^R}(y|x) - p_{\theta^0}(y|x)\|_1 \leq \varepsilon$$

B.3. Main Theoretical Results

Theorem 1 (Representation Convergence Guarantee) *Under the FUCRT training procedure with learning rate schedule $\eta_t = \frac{\eta_0}{\sqrt{t}}$:*

1. *For any $\delta > 0$, there exists a number of training iterations T such that the unlearned model θ^R achieves representation convergence, satisfying:*

$$\inf_{x_r \in \mathcal{R}: y_r = TF(x_u)} \|f_{\theta^R}(x_u) - f_{\theta^R}(x_r)\|_2 \leq \delta, \quad \forall x_u \in \mathcal{S}$$

2. *The minimum expected representation distance over T iterations satisfies:*

$$\min_{t \leq T} \mathbb{E}[\delta_t^2] \leq \frac{C}{\sqrt{T}},$$

where

$$C = K^2 \cdot \left(\frac{2[\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]}{\eta_0} + \eta_0 L \sigma^2 \right)$$

with K being a constant relating gradient magnitude to representation distance, $\mathcal{L}(\theta^{(0)})$ the initial loss, $\mathcal{L}^*$ the global minimum loss, η_0 the initial learning rate, L the loss smoothness constant, and σ^2 the bound on stochastic gradient variance.

Proof.

We first analyze the cross-class fusion loss component. For any pair (x_u, x_r) where $x_u \in \mathcal{S}$ and $x_r \in \mathcal{R}$ with $y_r = TF(x_u)$, the fusion loss can be expressed as:

$$\mathcal{L}_{\text{fusion}} = -\log \frac{\exp(f_\theta(x_u) \cdot f_\theta(x_r)/\tau)}{\sum_{x' \in \text{batch}} \exp(f_\theta(x_u) \cdot f_\theta(x')/\tau)}$$

where $\tau > 0$ is the temperature parameter.

Computing the gradient with respect to the representation of x_u:

$$\frac{\partial \mathcal{L}_{\text{fusion}}}{\partial f_\theta(x_u)} = -\frac{1}{\tau} \left(f_\theta(x_r) - \sum_{x' \in \text{batch}} p(x'|x_u) f_\theta(x') \right)$$

where

$$p(x'|x_u) = \frac{\exp(f_\theta(x_u) \cdot f_\theta(x')/\tau)}{\sum_{x'' \in \text{batch}} \exp(f_\theta(x_u) \cdot f_\theta(x'')/\tau)}$$

This gradient drives $f_\theta(x_u)$ toward $f_\theta(x_r)$ when they form a transformation pair, facilitating representation convergence.

Then, we adopt standard assumptions for non-convex stochastic optimization:

1. Bounded gradients: $\|\nabla_\theta \mathcal{L}(\theta)\|_2 \leq G$ for some constant $G > 0$
2. L-smoothness: $\|\nabla_\theta \mathcal{L}(\theta_1) - \nabla_\theta \mathcal{L}(\theta_2)\|_2 \leq L\|\theta_1 - \theta_2\|_2$
3. Bounded variance: $\mathbb{E}[\|\nabla_\theta \mathcal{L}(\theta, \xi) - \nabla_\theta \mathcal{L}(\theta)\|_2^2] \leq \sigma^2$

where ξ denotes a random mini-batch.

We begin with the classical convergence result from non-convex stochastic gradient descent with the learning rate schedule $\eta_t = \frac{\eta_0}{\sqrt{t}}$:

$$\min_{t \leq T} \mathbb{E}[\|\nabla_\theta \mathcal{L}(\theta^{(t)})\|_2^2] \leq \frac{1}{\sqrt{T}} \left(\frac{2[\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]}{\eta_0} + \eta_0 L \sigma^2 \right).$$

Define the squared representation distance at iteration t as:

$$\delta_t^2 = \inf_{x_r \in \mathcal{R}: y_r = TF(x_u)} \|f_{\theta^{(t)}}(x_u) - f_{\theta^{(t)}}(x_r)\|_2^2.$$

Due to the assumption relating representation distance and gradient magnitude, we have:

$$\delta_t \leq K \|\nabla_\theta \mathcal{L}(\theta^{(t)})\|_2,$$

which directly implies:

$$\delta_t^2 \leq K^2 \|\nabla_\theta \mathcal{L}(\theta^{(t)})\|_2^2.$$

Taking expectation, we obtain:

$$\mathbb{E}[\delta_t^2] \leq K^2 \mathbb{E}[\|\nabla_\theta \mathcal{L}(\theta^{(t)})\|_2^2].$$

Therefore, considering the minimum over iterations up to T gives:

$$\min_{t \leq T} \mathbb{E}[\delta_t^2] \leq K^2 \min_{t \leq T} \mathbb{E}[\|\nabla_\theta \mathcal{L}(\theta^{(t)})\|_2^2].$$

Substituting the standard convergence bound for non-convex SGD:

$$\min_{t \leq T} \mathbb{E}[\delta_t^2] \leq K^2 \cdot \frac{1}{\sqrt{T}} \left(\frac{2[\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]}{\eta_0} + \eta_0 L \sigma^2 \right).$$

Define:

$$C = K^2 \cdot \left(\frac{2[\mathcal{L}(\theta^{(0)}) - \mathcal{L}^*]}{\eta_0} + \eta_0 L \sigma^2 \right),$$

which completes the proof, yielding:

$$\min_{t \leq T} \mathbb{E}[\delta_t^2] \leq \frac{C}{\sqrt{T}}.$$

This completes the proof.

Theorem 2 (Representation Convergence to Unlearning) *For a FUCRT model θ^R that achieves representation convergence with parameter δ, there exists a function $h(\delta)$ such that θ^R is an $(h(\delta), \delta)$-representation-based unlearning model, where:*

$$h(\delta) = \max(L_g \cdot \delta + \gamma(\delta), \kappa(\delta) \cdot \delta)$$

with:

- L_g: *Lipschitz constant of the classification function* g_{θ^R}
- $\gamma(\delta)$: *Transformation class identification error with* $\lim_{\delta\to 0}\gamma(\delta)=0$
- $\kappa(\delta)$: *Interference coefficient between unlearning transformation and remaining data classification with* $\lim_{\delta\to 0}\kappa(\delta)=0$

Furthermore, $\lim_{\delta\to 0}h(\delta)=0$*, ensuring that perfect representation convergence implies perfect unlearning.*

Proof.

We verify each condition of Definition 1:

1. Representation Convergence: This condition is satisfied by assumption with parameter δ.

2. Erasure Guarantee: For any $(x_u, y_u)\in\mathcal{S}$ where $y_u\in\mathcal{C}_\mathcal{F}$ is the original label, let x_r be a sample from $\mathcal{R}$ with $y_r=TF(x_u)$ such that $\|f_{\theta^R}(x_u)-f_{\theta^R}(x_r)\|_2\le\delta$.

Since g_{θ^R} is L_g-Lipschitz continuous:

$$\begin{aligned}&\|g_{\theta^R}(f_{\theta^R}(x_u))-g_{\theta^R}(f_{\theta^R}(x_r))\|_1\\&\le L_g\cdot\|f_{\theta^R}(x_u)-f_{\theta^R}(x_r)\|_2\\&\le L_g\cdot\delta\end{aligned}$$

The probability assigned to the original class y_u can be bounded:

$$p_{\theta^R}(y_u|x_u)=[g_{\theta^R}(f_{\theta^R}(x_u))]_{y_u}$$

Since the model is trained to classify x_r as $y_r\neq y_u$, and x_u is close to x_r in representation space:

$$p_{\theta^R}(y_u|x_u)\le p_{\theta^R}(y_u|x_r)+L_g\cdot\delta$$

The transformation class identification process introduces error $\gamma(\delta)$, which accounts for:
- Imperfect selection of transformation targets
- Residual information about original classes

Thus:

$$p_{\theta^R}(y_u|x_u)\le L_g\cdot\delta+\gamma(\delta)$$

As $\delta\to 0$:
- Representation convergence becomes perfect
- Transformation targets become increasingly accurate
- Hence $\lim_{\delta\to 0}\gamma(\delta)=0$

3. Utility Preservation: For any $x\in\mathcal{R}$, the FUCRT objective explicitly maintains original predictions through $\mathcal{L}_{\text{finetune}}$ on remaining data.

The representation transformation of unlearning data can affect remaining data through:
- Shared parameters in the representation function
- Modified decision boundaries

Let $\kappa(\delta)$ quantify this interference normalized by δ:

$$\|p_{\theta^R}(y|x)-p_{\theta^0}(y|x)\|_1\le\kappa(\delta)\cdot\delta$$

As $\delta\to 0$:
- Impact on remaining data diminishes
- Hence $\lim_{\delta\to 0}\kappa(\delta)=0$

Combining the bounds: Taking $h(\delta)=\max(L_g\cdot\delta+\gamma(\delta),\kappa(\delta)\cdot\delta)$, we have:
- Erasure guarantee: $\sup_{(x,y)\in\mathcal{S}}p_{\theta^R}(y|x)\le h(\delta)$
- Utility preservation: $\sup_{x\in\mathcal{R}}\|p_{\theta^R}(y|x)-p_{\theta^0}(y|x)\|_1\le h(\delta)$

Since $\lim_{\delta\to 0}[L_g\cdot\delta+\gamma(\delta)]=0$ and $\lim_{\delta\to 0}[\kappa(\delta)\cdot\delta]=0$, we have:

$$\lim_{\delta\to 0}h(\delta)=0$$

Therefore, θ^R is an $(h(\delta),\delta)$-representation-based unlearning model.

Table 3. Accuracy (%) for the hyperparameters λ_1 and λ_2 of FUCRT on CIFAR10 dataset under both IID and Non-IID settings.

Hyperparameter		**IID**		**Non-IID**	
λ_1	λ_2	**U-set**	**R-set**	**U-set**	**R-set**
0.01	0.01	0.00	89.97	0.00	89.73
0.05	0.05	0.00	90.05	0.00	89.76
0.1	0.1	0.00	90.09	0.00	89.79
0.3	0.3	0.00	90.06	0.00	89.82
0.5	0.5	0.00	90.04	0.00	89.67
1	1	0.00	90.02	0.00	89.62
10	10	0.00	89.68	0.00	88.40
0.3	0.01	0.00	90.01	0.00	89.75
0.3	0.05	0.00	90.07	0.00	89.74
0.3	0.1	0.00	90.10	0.00	89.79
0.3	0.3	0.00	90.06	0.00	89.82
0.3	0.5	0.00	90.05	0.00	89.75
0.3	1	0.00	90.03	0.00	89.70
0.3	10	0.00	89.75	0.00	89.10
0.01	0.3	0.00	89.98	0.00	89.71
0.05	0.3	0.00	90.06	0.00	89.71
0.1	0.3	0.00	90.09	0.00	89.74
0.3	0.3	0.00	90.06	0.00	89.82
0.5	0.3	0.00	90.04	0.00	89.72
1	0.3	0.00	90.02	0.00	89.65
10	0.3	0.00	89.83	0.00	89.31

C. The Impact of Hyperparameter

To examine the influence of hyperparameters on the performance of our proposed method, we conduct experiments focusing on key hyperparameters λ_1 and λ_2, as illustrated in Table 3. We can observe that: (1) Under different hyperparameter settings, our method always achieves complete erasure of unlearning data, that is, the accuracy of unlearning data is 0%. (2) Within a reasonable range of hyperparameter variations, we achieve relatively low accuracy fluctuations on the remaining data, with accuracy ranging from

89.68% to 90.10% under the IID setting and from 88.40% to 89.82% under the Non-IID settings. (3) As the proportion of $\mathcal{L}_1$ and $\mathcal{L}_2$ in the total loss $\mathcal{L}$ increases, the accuracy of the remaining data first improves and then declines. The initial improvement can be attributed to the introduction of the cross-class fusion technique. This technique enhances the accuracy of the remaining data by aligning the transformation processes across clients and optimizing both local and global representation spaces. The subsequent decline results from an excessive focus on the transformation fusion process. Over-fusion in the representation space reduces the classifier's optimization capability, leading to a drop in accuracy on the remaining data. (4) A proper balance between $\mathcal{L}_1$ and $\mathcal{L}_2$ is essential. Insufficient fusion, either intra-client or inter-client, would adversely impact the unlearning model's accuracy on the remaining data. (5) $\lambda_1 = 0.3$ and $\lambda_2 = 0.3$ are appropriate hyperparameter values for experiments, which are also used as default values in our experiments.