RBA: Representation Backdoor Attack on Personalized Federated Learning

Hefeng Zhou[*1], Dingxuan Zhang[*1], Zhihong Su[2], Honghong Zeng[1], Jiong Lou[1], Wugedele Bao[3], Chentao Wu[1], Jie Li[†1]
[1]Shanghai Jiao Tong University
[2]Ant Group
[3]Hohhot Minzu College

***Abstract*—Personalized federated learning (FL) tackles the challenge of data heterogeneity by training personalized models instead of a single global model. Backdoor attacks pose a fatal threat to FL systems. There is an urgent need to investigate on backdoor attacks under personalized FL. Existing representation-based personalized FL frameworks share a global extractor, which creates a large attack surface for backdoor attacks. This work explores previously unknown backdoor risks in FedPer and FedRep through data poisoning and parameter scaling. Experimental results show that the attack success rate (ASR) degrades dramatically under different pFL frameworks. We further propose a backdoor attack leveraging the shared representation extractor to inject backdoors more effectively. Experimental results demonstrate that our method improves the ASR by up to 72% while maintaining the ASR after stopping the attack for 200 rounds. Code is available at https://github.com/RezinChow/RBA.**

***Index Terms*—Backdoor attack, personalized federated learning, shared representations.**

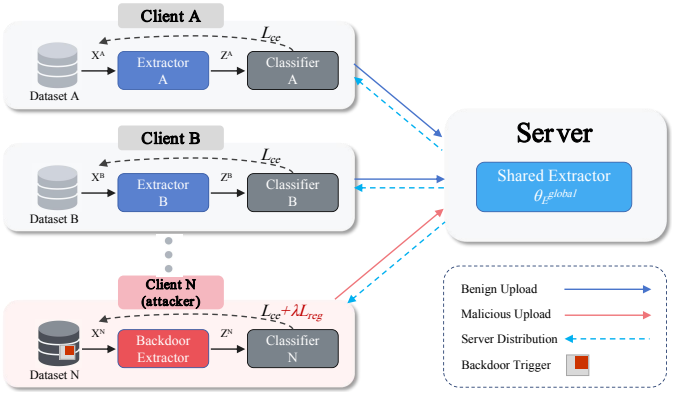


Fig. 1. Overview of the proposed RBA. The attacker trains a model on the backdoor data with representation regularization and extractor poisoning (Section III-A), then the backdoor extractor is submitted to server.

I. Introduction

Federated learning (FL) [1] enables collaborative model training while preserving data privacy. However, traditional approaches struggle with data heterogeneity [2]. Personalized federated learning (pFL) addresses this by tailoring models to individual participants [3]. Representation-based methods like FedPer [4] and FedRep [5] allow personalized models to share a common representation extractor while maintaining separate classifiers. Furthermore, FedALA [6] and VPFL [7] employ adaptive weighting and model parameter distance strategies to guide client updates, which helps balance the global optimization direction with the local data distribution.

Among security threats, backdoor attacks are particularly stealthy and harmful [8]. These attacks inject hidden triggers that cause specified outputs upon activation. In federated settings, attackers often scale malicious parameters to enhance impact [9]. Defenses are challenging due to privacy mechanisms like secure aggregation [10]. While some studies have explored backdoor risks in pFL [11], representation-based frameworks remain understudied. Our experiments show existing attacks suffer significant performance degradation in these frameworks. In contrast to computationally complex attacks [12] that require control over the aggregation process or extensive client-side modifications, we propose a lightweight Representation Backdoor Attack (RBA) that poisons the shared extractor via representation regularization and extractor poisoning. As shown in Figure 1, RBA requires minimal modifications yet outperforms more complex attacks, effectively exploiting architectural vulnerabilities in pFL.

Our contributions are summarized as: (1)We experimentally illustrate that representation-based FL is still vulnerable to backdoor attacks, in spite of the attack success rate (ASR) decline due to personalized classifiers. (2)We propose, to the best of our knowledge, a novel backdoor attack that exploits characteristics of representation-based pFL implementations to enhance attack effectiveness. (3)We conduct extensive experiments to demonstrate the superiority of our method in terms of ASR and backdoor.

Beyond proposing a new attack, our study reveals previously overlooked vulnerabilities in shared representations and provides insights that may inform defense design for pFL.

II. Problem Preliminary

A. Personalized Federated Learning

Personalized federated learning aims to train tailored models for each participant to address data heterogeneity, and its optimization objective can be expressed as:

$$\min_{(q_1,\ldots,q_N)} \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{D}_i|}{|\mathcal{D}|}L_i\left(q_i\right), \tag{1}$$

*These authors contributed equally to this work. † Corresponding author.

where N denotes the number of participants, $\mathcal{D}_i$ is the local dataset of the i-th participant, $\mathcal{D} \triangleq \cup_{i\in[N]}\mathcal{D}_i$ is the union of all local datasets, L_i is the loss function for the i-th participant, and q_i is the local model of the i-th participant. In the supervised learning scenario, $L_i(q_i)$ can be defined as:

$$L_i\left(q_i\right) = \frac{1}{|\mathcal{D}_i|} \sum_{(x,y)\in\mathcal{D}_i} l\left(q_i(x), y\right), \tag{2}$$

where l usually represents the cross-entropy function under classification tasks and for representation-based personalized federated learning frameworks, the local model q_i is decomposed into a shared representation extractor ϕ_i and a personalized classifier h_i. Thus, the local model can be expressed as $q_i = h_i$. [13]

B. Threat Model

The threat model for representation-based pFL backdoor attacks is illustrated in Figure 1. The attacker can arbitrarily modify their local training process, data, and uploaded parameters, but cannot alter the server's aggregation algorithm [9], [14], [15], access the global data distribution, or exceed the data volume of a single benign participant.

In our setup, a single attacker is selected every 10 rounds and scales their model parameters before upload. This strategy is designed to be robust against common defenses like norm clipping and Krum aggregation [16], as manipulations can be tuned to stay within detection thresholds or exploit enforcement gaps in practical settings [17]. Such parameter scaling demonstrates the attack's adaptability and strength even under defensive constraints.

C. Backdoor Attacks

The core goal of a backdoor attack is to make the local model function normally for benign inputs while outputting a pre-defined target label y^t for inputs containing a specific backdoor trigger T. Formally, for any benign input $(x, y) \in \mathcal{D}_i$, the attacker requires:

$$\begin{aligned} q_i(x) &= y, \\ q_i(x+T) &= y^t. \end{aligned} \tag{3}$$

Two common techniques for implementing backdoor attacks are as follows.

Data Poisoning: In data poisoning attacks within federated learning, a malicious participant (attacker) intentionally corrupts their local dataset to compromise the global model. Specifically, the attacker trains the local model using a poisoned dataset $\hat{\mathcal{D}}_i$, which includes trigger-modified samples designed to embed a backdoor behavior. These modifications ensure that the model performs normally on clean inputs but exhibits adversarial behavior (e.g., misclassification) when the trigger is present. The training objective is to minimize the loss of the poisoned model $\hat{q}_i$:

$$\hat{L}_i\left(\hat{q}_i\right) = \frac{1}{\left|\hat{\mathcal{D}}_i\right|} \sum_{(x,y)\in\hat{\mathcal{D}}_i} l\left(\hat{q}_i(x), y\right), \tag{4}$$

where $\hat{q}_i$ denotes the backdoor model trained on $\hat{\mathcal{D}}_i$, and l is typically the cross-entropy loss function. This approach allows the backdoor to propagate to the aggregated global model, potentially affecting all participants. [18]

Parameter scaling: To amplify the impact of malicious updates, the attacker scales the uploaded model parameters. In the k-th global round, the attacker uploads the scaled parameter in this way: $n(\hat{q}^k - g^{k-1}) + g^{k-1}$. The ideal outcome is to replace the global model with the attacker's poisoned model [9].

D. The Weakness of Existing Backdoor Attacks in pFL

Conventional backdoor attacks, while effective in standard FL, often fail in pFL due to three primary factors [19].

Limited Transferability to Personalized Models Unlike standard FL, pFL optimizes personalized models that differ from the global aggregate. Regularization-based proximity to the global model does not guarantee the transfer of backdoor-related parameters, rendering global model manipulation insufficient to compromise local inference.

Parameter Decoupling in Partial Model-Sharing pFL Many pFL methods employ partial model sharing, where shared feature extractors are coupled with private classifiers [20]. Even if a backdoor is injected into the shared layers, the private classifiers—trained exclusively on clean local data—remain unaligned with the malicious objective and fail to activate the trigger during inference.

Forgetting Induced by Continuous Personalization The repetitive optimization on clean, non-IID local datasets causes personalized models to "forget" backdoor features. This degradation is exacerbated by infrequent selection of compromised clients and post-training fine-tuning on clean data.

Implications for Representation-Based pFL. These limitations are most acute in representation based pFL (FedPer, FedRep, FedALA). Since private classifiers are never aggregated, injected backdoors cannot propagate across the network [21]. While attacks on FedAvg can achieve near-100% ASR, ASR on pFL methods remains significantly lower in experimental evaluations [22] (see Figure 2).

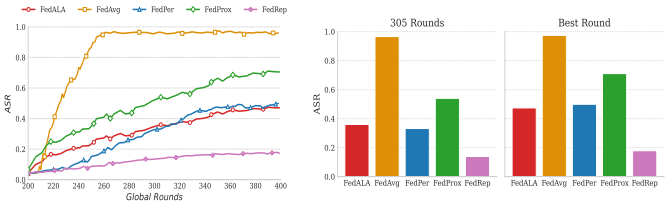


Fig. 2. Comparison of the effectiveness of data poisoning attack under FedAvg, FedPer, FedRep, FedALA and FedProx. The attacker starts attacking from the 205th global round and attacks every 10 global rounds. Experimental results show that the ASR degrades in FedPer, FedRep, FedALA and FedProx.

III. Proposed Method: Representation Backdoor Attack (RBA)

This section details the intuitions and implementation of the proposed Representation Backdoor Attack (RBA)1, and visualizes RBA's effect on representations.

(a) Data poisoning (b) RBA

Fig. 3. T-SNE visualization of representations under FMNIST and FedPer with parameter scaling. Trigger inputs' representations (purple scatters) and target inputs' representations (gold scatters) are closer under RBA.

A. Intuitions and Implementation for RBA

While personalized classifiers hinder backdoor propagation, the shared representation extractor creates a critical attack surface. Since all clients use the aggregated extractor, poisoning the extractor can enable backdoor propagation across multiple clients. Based on this insight, RBA improves attack effectiveness through two key modifications: representation regularization and extractor poisoning.

Representation Regularization

We propose to make the representation corresponding to the inputs with the trigger and the benign inputs of the target label as similar as possible, such that the classifiers output the target label when such kind of representations are fed. Specifically, we propose to modify the objective of local training to the following form:

$$\min_{\hat{q}_i} \hat{L}_i(\hat{q}_i) + \mu L_i^r(\hat{q}_i), \tag{5}$$

$$L_i^r(\hat{q}_i) = \frac{1}{|\hat{\mathcal{D}}_i^b|} \sum_{(\hat{x},y^t)\in\hat{\mathcal{D}}_i^b} \|\hat{\phi}_i(\hat{x}) - r_i^t\|, \tag{6}$$

$$r_i^t = \frac{1}{|\mathcal{D}_i^t|} \sum_{(x,y^t)\in\mathcal{D}_i^t} \hat{\phi}_i(x), \tag{7}$$

where $\hat{\mathcal{D}}_i$ is the training dataset for the backdoor attack and $\hat{\phi}_i$ is the representation extractor of the backdoor model $\hat{q}_i$. The sets $\hat{\mathcal{D}}_i^b, \mathcal{D}_i^t \subset \hat{\mathcal{D}}_i$ contain trigger-modified inputs and benign target-label inputs respectively. The term r_i^t represents the average representation of the inputs of the target class. RBA introduces the representation regularization term $\mu L_i^r(\hat{q}_i)$ to the original loss function, where μ is a hyperparameter controlling its weight.

Extractor Poisoning

Considering that the classifier part tends to be poisoned more seriously while the representation extractor part is less poisoned [23], we propose to update the representation extractor part solely during local training, which aims to improve the effect of poisoning in the representation extractor.

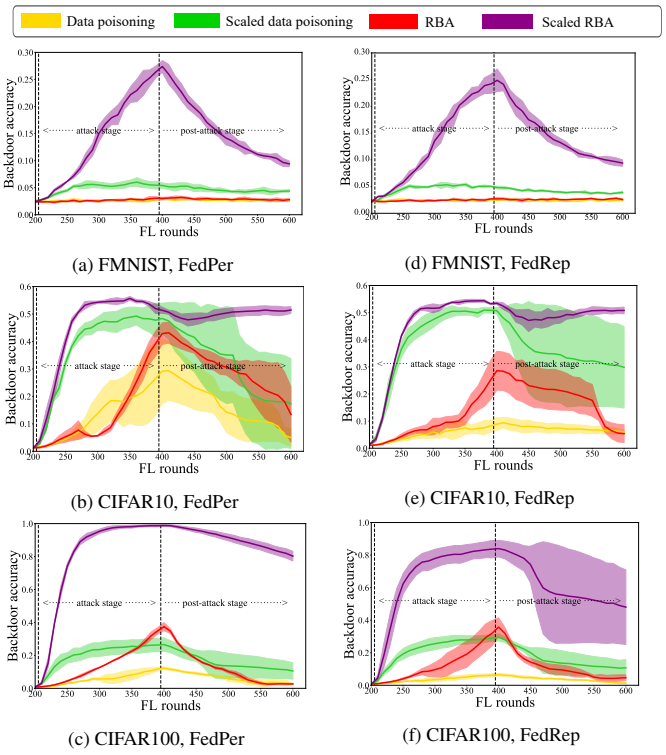
B. Representation Visualization

We use T-SNE [24] to visualize the representations and the difference between representations under the baseline method (data poisoning) and RBA are shown in Figure 3.

Algorithm 1 RBA's local training of the k-th global round.

Require: learning rate η_{atk}, attacker's local epochs E_{atk}, local dataset $\mathcal{D}_{\text{local}}$, shared representation extractor $\bar{\phi}$, local model $q = h \circ \phi$, poison ratio p, original loss function l_o, hyper-parameter μ

$\phi = \bar{\phi}$
for local epoch $e \in E_{\text{atk}}$ **do**
  for batch $b = \{x, y\} \in \mathcal{D}_{\text{local}}$ **do**
    $\hat{b} = \{b^p, b^o\}$ = partial_poison(b, p) {// get partial poisoned batch}
    b^t = sample_target$(\mathcal{D}_{\text{local}})$
    r^t = avg_representation(b^t, ϕ)
    $l_r(\phi) = \frac{1}{|b^p|} \sum_{(\hat{x},y^t)\in b^p} \|\phi(\hat{x}) - r^t\|$
    $l = l_o + \mu l_r$
    $\phi = \phi - \eta \nabla l$
  end for
end for
return ϕ



(a) FMNIST, FedPer (d) FMNIST, FedRep

(b) CIFAR10, FedPer (e) CIFAR10, FedRep

(c) CIFAR100, FedPer (f) CIFAR100, FedRep

Fig. 4. ASR (mean ± std) of our RBA and the baseline (data poisoning) with and without parameter scaling, evaluated on FMNIST, CIFAR10, and CIFAR100 under FedPer and FedRep. The attacker is selected every 10 rounds from round 205 to 395 (20 attacks total). RBA achieves higher ASR than the baseline during and after the attack phase.

IV. Experiments

A. Datasets and Experimental Setup

We evaluate our method on three image classification datasets: FMNIST [25], CIFAR-10, and CIFAR-100 [22]. In our setting, there are 100 participants in total, 10 of whom are selected randomly in each global round. To simulate a non-IID

scenario, we divide the samples using a Dirichlet distribution [26] with a concentration parameter of 0.5. We adopted the standard lightweight ResNet-18 [27] for the CIFAR datasets and LeNet [28] for FMNIST, following recent evaluation protocols [29].

Our method is agnostic to the backdoor triggers, and we use a static pixel pattern of size 5×3 in the lower-left corner as the trigger in the following experiments. We use the data poisoning method described in Section II-C as the baseline and compare it with our RBA in cases both with and without parameter scaling. The parameter μ is set to 0.01 for CIFAR-10 and CIFAR-100, and 0.1 for FMNIST. In our threat model, there is a single attacker, and this attacker is selected to participate every 10 global rounds. The ASR reported in all subsequent experiments refers to the **average ASR of benign clients**.

B. Main Results

Figure 4 shows the performance of our representation backdoor attack compared to the baseline method. Each experiment is performed 3 times. The two vertical dotted lines at 205th round and 395th round in the diagram indicate the timing of the first and last attacks.

ASR.Our RBA achieves higher ASR than the baseline, with and without parameter scaling. As depicted in Figure 4.e, our method achieves a ASR of 28.6% under FedRep in CIFAR10 experiments without parameter scaling, which is about 3 times that of the baseline method. In FMNIST experiments (Figure 4.a.d) with parameter scaling, the highest ASR of our method is about 5 times that of the baseline method.

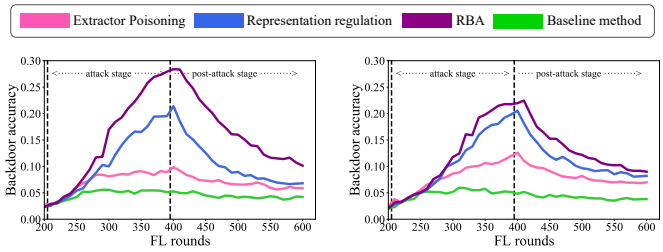
Figure 4.c.f show that the improvement in ASR by our method is particularly significant under CIFAR100. Experimental results show that parameter scaling enhances backdoor attack effectiveness. Our method achieves higher ASR than the scaled baseline even without scaling. With scaling, it reaches 83.9% and 98.8% under FedRep and FedPer, about four times the baseline respectively. On FMNIST, unscaled attacks are largely ineffective, likely due to the simpler model and greater influence of the personalized classifier, unlike on CIFAR10 and CIFAR100.

Backdoor Durability. Our method maintains a higher ASR than the baseline method in both the attack stage and the post-attack stage. As is shown in Figure 4.b .e, even after stopping the attack for 200 rounds, the ASR of our method with parameter scaling does not degrade under CIFAR10. CIFAR100 experiments in Figure 4.c .f demonstrate that the ASR of our method is still higher than the baseline method's highest ASR after 200 rounds of attacker-free training.

Main Task Accuracy. The mean and standard deviation of the main task accuracy in the attack stage and the post-attack stage are presented in Table I. There is no significant difference in the main task accuracy between our method and data poisoning.

C. Ablation Studies

Poisoning and Regularization. To explore the contribution of our two modification over the existing attack methods, we conduct ablation experiments under FMNIST. Figure 5 shows both extractor poisoning and representation regularization contribute to the improvement in ASR, and higher ASR is achieved when the two modifications are applied together. Specifically, the ASR improves by 4.3%, 15.8% and 22.8% as a result of extractor poisoning, representation regularization and both modifications respectively under FedPer.



(a) FedPer (b) FedRep

Fig. 5. Ablation experiments under FMNIST show that both extractor poisoning and representation regularization contribute to the improvement in ASR. The parameter scaling is performed by default.

Effects of Attack. Previously we launched the attack at the 205th global round when the model had converged. Here we examine how convergence level affects attack success by varying the start time. As Table II shows, the baseline suffers an 8% drop in peak ASR when attacking early. In contrast, ours remains effective even on unconverged models, demonstrating superior backdoor durability.

TABLE I
COMPARISON ON MAIN TASK ACCURACY BETWEEN RBA AND THE BASELINE METHOD UNDER FEDREP.

Dataset	Backdoor Method	Attack stage	Post-attack stage
FMNIST	baseline	$90.8\% \pm 0.36\%$	$91.1\% \pm 0.14\%$
	RBA	$90.9\% \pm 0.30\%$	$91.1\% \pm 0.12\%$
	baseline(scaled)	$90.8\% \pm 0.27\%$	$91.2\% \pm 0.14\%$
	RBA(scaled)	$90.7\% \pm 0.71\%$	$91.9\% \pm 0.17\%$
CIFAR10	baseline	$78.6\% \pm 2.61\%$	$82.0\% \pm 0.65\%$
	RBA	$78.2\% \pm 2.13\%$	$80.4\% \pm 3.58\%$
	baseline(scaled)	$79.2\% \pm 2.71\%$	$80.8\% \pm 1.41\%$
	RBA(scaled)	$77.3\% \pm 2.72\%$	$81.2\% \pm 2.28\%$
CIFAR100	baseline	$42.4\% \pm 1.82\%$	$41.6\% \pm 3.10\%$
	RBA	$42.6\% \pm 1.26\%$	$41.8\% \pm 3.18\%$
	baseline(scaled)	$43.2\% \pm 3.22\%$	$43.8\% \pm 2.32\%$
	RBA(scaled)	$42.2\% \pm 3.55\%$	$43.9\% \pm 1.97\%$

TABLE II
QUANTITATIVE RESULTS OF ATTACKS WITH PARAMETER SCALING UNDER CIFAR10 AND FEDREP, WHERE a_{BEST} AND a_{POST} DENOTE THE PEAK ASR DURING THE ATTACK AND THE ACCURACY 200 ROUNDS AFTER ATTACK TERMINATION, RESPECTIVELY.

Metrics	Backdoor Method	Rounds for attack stage		
		5 to 195	55 to 245	205 to 395
a_{best}	baseline	42.5%	42.6%	50.9%
	RBA	**52.2%**	**53.6%**	**54.3%**
a_{post}	baseline	19.4%	15.6%	29.8%
	RBA	**50.4%**	**50.0%**	**50.8%**

D. Robustness against Defenses

To evaluate the stealthiness and robustness of our proposed method, we conduct experiments on the CIFAR-10 dataset using the FedRep algorithm with a ResNet-18 backbone. The attack is launched continuously starting from communication round 5. We test the method against two representative server-side defenses: Norm Clipping [30] and Multi-Krum [29].

Norm Clipping. This defense mechanism mitigates malicious impact by restricting the L_2 norm of model updates to a predefined threshold M. Updates exceeding this threshold are clipped (scaled down) before aggregation. As illustrated in Figure 6, strategies relying on magnitude scaling to overpower benign updates (e.g., Scaled RBA and Scaled Data Poisoning) are rendered ineffective. Their high-norm updates are easily identified as outliers and aggressively clipped, resulting in a near-zero ASR.

In stark contrast, the unscaled RBA (with scaling factor $n = 1$) demonstrates remarkable robustness, steadily achieving high ASR over communication rounds. This stealthiness is attributed to our representation regularization term (L_i^r), which aligns the features of the trigger inputs with those of the benign target inputs. This alignment simplifies the optimization landscape, allowing the attack to effectively implant the backdoor without requiring large, anomalous gradient updates. Consequently, the malicious updates of RBA naturally fall within the safe range of the clipping threshold, effectively bypassing the defense while maintaining high attack efficacy.

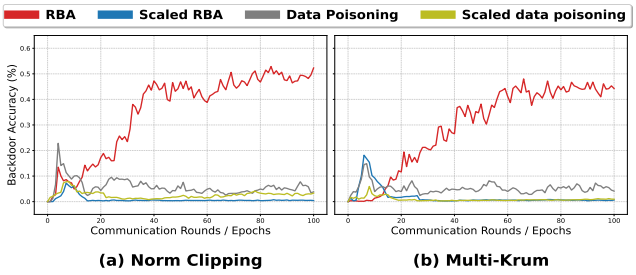


Fig. 6. ASR of ResNet18 on CIFAR-10 using FedRep under different defense mechanisms. The curves illustrate the performance of Scaled/Unscaled RBA and Data Poisoning attacks against Norm Clipping (threshold= 1.5 median norm) and Multi-Krum ($m = 8$) defenses.

Multi-Krum. Multi-Krum [29] improves robustness by selecting a subset S of $M - f$ updates with the smallest Krum scores, where M is the number of candidates and f is the assumed upper bound of malicious clients. The Krum score for an update Δ_i is the sum of Euclidean distances to its nearest neighbors. The global update is computed as:

$$\text{GlobalUpdate} = \frac{1}{|S|} \sum_{i \in S} \Delta_i. \tag{8}$$

By discarding statistically deviant updates, Multi-Krum effectively mitigates large-magnitude or anomalous attacks. As shown in Figure 6, baseline methods like Data Poisoning and Scaled RBA fail under this defense (ASR $\approx 0\%$) because their statistical deviations are easily flagged as outliers. In contrast, RBA achieves an ASR of approximately 45% in our experiments (where $f = 2, M = 10$).

RBA's resilience stems from its optimization of the distance between trigger representations and benign target centroids. This guides the malicious extractor to produce updates that are statistically closer to the benign cluster in the parameter space, successfully bypassing Krum's distance-based exclusion mechanism.

V. Theoretical Analysis

In this section, we analyze the convergence and effectiveness of the attacker's local optimization. Let $F(\mathbf{w}) = \hat{L}_i(\hat{q}_i(\mathbf{w})) + \mu L_i^r(\hat{q}_i(\mathbf{w}))$ be the non-convex surrogate objective.

A. Convergence of Local Optimization

To analyze the feasibility of minimizing F, we adopt standard non-convex SGD assumptions: (1) F is L-smooth; (2) stochastic gradients $g(\mathbf{w}_t)$ are unbiased with variance σ^2; (3) F is bounded below by F^*.

Theorem 1 (Local Convergence)**.** *For learning rate $\eta \leq \frac{1}{L}$, the average squared gradient norm after T iterations satisfies:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2] \leq \frac{2(F(\mathbf{w}_0) - F^*)}{\eta T} + L\eta\sigma^2. \tag{9}$$

Proof. By L-smoothness and the SGD update rule, taking the expectation yields: $\mathbb{E}[F(\mathbf{w}_{t+1})] \leq \mathbb{E}[F(\mathbf{w}_t)] - (\eta - \frac{L\eta^2}{2})\|\nabla F(\mathbf{w}_t)\|^2 + \frac{L\eta^2\sigma^2}{2}$. Setting $\eta \leq 1/L$, rearranging, and summing from $t = 0$ to $T - 1$ concludes the proof. □

Theorem 1 ensures the attacker reliably reaches a stationary point, producing optimized malicious updates rather than random noise.

B. Attack Effectiveness Analysis

We now analyze why minimizing the representation loss leads to successful backdoors. We assume the benign classifier h_i is K-Lipschitz and possesses a *robust decision boundary*: there exists a margin $\gamma > 0$ such that any representation z satisfying $\|z - r_t\|_2 \leq \gamma$ is classified as the target class y^t.

Theorem 2 (Attack Success Probability)**.** *Let ϕ^* be the poisoned extractor with expected representation loss $\mathbb{E}[\|\phi^*(\hat{x}) - r_t\|^2] \leq \epsilon$. The probability of attack success on a benign client i satisfies:*

$$P(\textit{Success}) \geq 1 - \frac{\epsilon}{\gamma^2}. \tag{10}$$

Proof. Let $Z = \|\phi^*(\hat{x}) - r_t\|^2$. The attack succeeds if $Z \leq \gamma^2$. By Markov's Inequality:

$$P(Z \geq \gamma^2) \leq \frac{\mathbb{E}[Z]}{\gamma^2} \leq \frac{\epsilon}{\gamma^2}. \tag{11}$$

Taking the complement $P(Z < \gamma^2) = 1 - P(Z \geq \gamma^2)$ completes the proof. □

This result demonstrates that the Attack Success Rate (ASR) approaches 100% as the attacker minimizes the surrogate loss ϵ relative to the decision margin γ^2. It provides a theoretical

foundation for the effectiveness of our representation-matching strategy.

VI. CONCLUSION

We study backdoor risks in representation-based personalized federated learning. We propose RBA, which injects backdoors by poisoning the shared representation extractor. Extensive experiments show RBA achieves higher and more durable ASR than baseline attacks with little impact on main-task accuracy. Our findings suggest defenses should explicitly safeguard shared representations in personalized FL.

ACKNOWLEDGMENT

This work was supported in part by National Key R&D Program of China No. 2024YFB2705300, NSFC Grant 62232011, 62402315, and the Shanghai Science and Technology Innovation Action Plan Grant 24BC3201200.

REFERENCES

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[2] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[3] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[4] M. G. Arivazhagan, V. Aggarwal, A. K. Singh, and S. Choudhary, "Federated learning with personalization layers," *arXiv preprint arXiv:1912.00818*, 2019.

[5] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Exploiting shared representations for personalized federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2089–2099.

[6] T. Li, S. Hu, A. Beirami, and V. Smith, "Fedala: Adaptive local aggregation for personalized federated learning," *arXiv preprint arXiv:2203.01197*, 2022.

[7] H. Zhou, Y. Wang, J. Wang, J. Lou, W. Bao, C. Wu, and J. Li, "Variational perturbation personalized federated learning via prior-posterior distance," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[8] T. D. Nguyen, T. Nguyen, P. L. Nguyen, H. H. Pham, K. Doan, and K.-S. Wong, "Backdoor attacks and defenses in federated learning: Survey, challenges and future research directions," *arXiv preprint arXiv:2303.02213*, 2023.

[9] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2938–2948.

[10] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for privacy-preserving machine learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.

[11] Y. Zeng, M. Pan, H. A. Just, L. Lyu, M. Qiu, and R. Jia, "Narcissus: A practical clean-label backdoor attack with limited information," *arXiv preprint arXiv:2204.05255*, 2022.

[12] X. Lyu, Y. Han, W. Wang, J. Liu, Y. Zhu, G. Xu, J. Liu, and X. Zhang, "Lurking in the shadows: Unveiling stealthy backdoor attacks against personalized federated learning," in *Proceedings of the 33rd USENIX Security Symposium*, 2024.

[13] C. Mclaughlin and L. Su, "Personalized federated learning via feature distribution adaptation," in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024, poster at NeurIPS 2024; discusses decomposition into shared representation learning and personalized classifier training for handling data heterogeneity in PFL. [Online]. Available: https://neurips.cc/virtual/2024/poster/94815

[14] Z. Zhang, A. Panda, L. Song, Y. Yang, M. Mahoney, P. Mittal, R. Kannan, and J. Gonzalez, "Neurotoxin: Durable backdoors in federated learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 429–26 446.

[15] Y. Wen, J. Geiping, L. Fowl, H. Souri, R. Chellappa, M. Goldblum, and T. Goldstein, "Thinking two moves ahead: Anticipating other users improves backdoor attacks in federated learning," *arXiv preprint arXiv:2210.09305*, 2022.

[16] R. Olapojoye, T. Salman, M. Baza, and A. Alshehri, "Fedecpa: An efficient countermeasure against scaling-based model poisoning attacks in blockchain-based federated learning," *Sensors*, vol. 25, no. 20, p. 6343, 2025.

[17] C. Lewis, V. Varadharajan, and N. Noman, "Attacks against federated learning defense systems and their mitigation," *Journal of Machine Learning Research*, vol. 24, no. 30, pp. 1–50, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-0014.html

[18] H. Zhuang, M. Yu, H. Wang, Y. Hua, J. Li, and X. Yuan, "Concealing backdoor model updates in federated learning by trigger-optimized data poisoning," *arXiv preprint arXiv:2405.06206*, 2025, updated version 2025; original arXiv submission 2024, with extensions in 2025. [Online]. Available: https://arxiv.org/abs/2405.06206

[19] M. Fan, Z. Hu, F. Wang, and C. Chen, "Bad-pfl: Exploring backdoor attacks against personalized federated learning," *arXiv preprint arXiv:2501.12736*, 2025, analyzes three key challenges/factors in PFL training dynamics and architectures (e.g., regularization insufficiency, non-shared parameters hindering backdoor survival) that cause degradation of conventional backdoor attacks. [Online]. Available: https://arxiv.org/abs/2501.12736

[20] Z. Qin, L. Yao, D. Chen, Y. Li, B. Ding, and M. Cheng, "Revisiting personalized federated learning: Robustness against backdoor attacks," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, 2023, pp. 2060–2071, shows partial model-sharing (e.g., FedRep, FedBN) impedes backdoor propagation/activation due to private classifier heterogeneity; full model-sharing lacks this benefit. [Online]. Available: https://dl.acm.org/doi/10.1145/3580305.3599898

[21] T. Ye, C. Chen, Y. Wang, X. Li, and M. Gao, "Bapfl: You can backdoor personalized federated learning," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 7, pp. 1–17, 2024.

[22] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technical report, University of Toronto*, 2009.

[23] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[24] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.

[25] H. Xiao, K. Rasul, and R. Vollgraf, "Fashionmnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[26] C. He, S. Li, J. So, M. Zhang, H. Wang, X. Wang, P. Vepakomma, A. Singh, H. Qiu, L. Shen *et al.*, "Fedml: A research library and benchmark for federated machine learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 15 695–15 709.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 2002.

[29] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage, "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1354–1371.

[30] Z. Sun, P. Kairouz, A. T. Suresh, and H. B. McMahan, "Can you really backdoor federated learning?" *arXiv preprint arXiv:1911.07963*, 2019.