A DETAILED IMPLEMENTATIONS OF SOTA ROBUST AGRS

Algorithm 2 AGR - FLAME

Input: n, $\mathbf{w}^0$, T $\triangleright n$ is the number of clients, $\mathbf{w}^0$ is the initial global model parameters, and T is the number of training iterations
Output: $\mathbf{w}^T$ $\triangleright \mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each client i in $[1, n]$ **do**
    $\mathbf{g}_i^t \leftarrow$CLIENTUPDATE$(\mathbf{w}^{t-1}, D_i)$ $\triangleright$ The aggregator sends $\mathbf{w}^{t-1}$ to Client i who trains $\mathbf{w}^{t-1}$ using its data D_i locally to achieve local gradient $\mathbf{g}_i^t$ and sends $\mathbf{g}_i^t$ back to the aggregator
  end for
  $(c_{11}^t, ..., c_{nn}^t) \leftarrow \cos(\mathbf{g}_1^t, ..., \mathbf{g}_n^t)$ $\triangleright$ $\forall i, j \in (1, ..., n)$, c_{ij}^t is the cosine distance between $\mathbf{g}_i^t$ and $\mathbf{g}_j^t$
  $(b_1^t, ..., b_L^t) \leftarrow$ HDBSCAN$(c_{11}^t, ..., c_{nn}^t)$ $\triangleright L$ is the number of admitted models, b_l^t is the index of the l-th model
  $(e_1^t, ..., e_n^t) \leftarrow \left\|(\mathbf{w}^{t-1}, (\mathbf{g}_1^t, ..., \mathbf{g}_n^t))\right\|_2$ $\triangleright e_i^t$ is the Euclidean distance between $\mathbf{w}^{t-1}$ and $\mathbf{g}_i^t$ $q^t \leftarrow$ MEDIAN $(e_1^t, ..., e_n^t)$ $\triangleright q^t$ is the adaptive clipping bound at round t
  for each client l in $[1, L]$ **do**
    $\mathbf{g}_i^t \leftarrow \mathbf{g}_i^t \cdot \min(1, (q^t / e_{b_l}^t))$ $\triangleright (q^t / e_{b_l}^t)$ is the clipping parameter, and $\mathbf{g}_i^t$ is clipped by the adaptive clipping bound
  end for
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \sum_{l=1}^{L} \mathbf{g}_i^t / L + N(0, \sigma^2)$ $\triangleright$ Sever aggregates parameters and adds noise, and then updates the global parameter as $\mathbf{w}^t$
end for

Algorithm 3 AGR - MDAM

Input: n, $\mathbf{w}^0$, β, $\mathbf{m}^0$, T $\triangleright n$ is the number of clients, $\mathbf{w}^0$ is the initial global model parameters, $\beta \in [0, 1)$ is the momentum coefficient of all the clients, $\mathbf{m}^0 = 0$ is the initial momentum of each honest client, and T is the number of training iterations
Output: $\mathbf{w}^T$ $\triangleright \mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each client i in $[1, n]$ **do**
    $\mathbf{g}_i^t \leftarrow$CLIENTUPDATE$(\mathbf{w}^{t-1}, D_i)$ $\triangleright$ The aggregator sends $\mathbf{w}^{t-1}$ to Client i who trains $\mathbf{w}^{t-1}$ using its data D_i locally to achieve local gradient $\mathbf{g}_i^t$
    $\mathbf{m}_i^t \leftarrow \beta \mathbf{m}_i^{t-1} + (1 - \beta)\mathbf{g}_i^t$ $\triangleright$ Each honest client sends to the server the momentum $\mathbf{m}_i^t$
  end for
  $S^t \in \underset{S \subset [n], |S| = n-f}{\arg\min} \left\{ \max_{i,j \in S} \left\| \mathbf{m}_i^t - \mathbf{m}_j^t \right\|_2 \right\}$ $\triangleright$ Server first chooses a set S^t of cardinality $n - f$ with the smallest diameter
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \frac{\eta}{n-f} \sum_{i \in S^t} \mathbf{m}_i^t$ $\triangleright$ Server then updates the global parameter as $\mathbf{w}^t$
end for

Algorithm 4 AGR - FLDetector

Input: n, $\mathbf{w}^0$, N, T $\triangleright N$ is the number of past iterations, and T is the number of training iterations
Output: $\mathbf{w}^T$ $\triangleright \mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each client i in $[1, n]$ **do**
    $\mathbf{g}_i^t \leftarrow$CLIENTUPDATE$(\mathbf{w}^{t-1}, D_i)$
    $\hat{\mathbf{g}}_i^t \leftarrow \mathbf{g}_i^{t-1} + \hat{H}^t(\mathbf{w}_t - \mathbf{w}_{t-1})$
  end for
  $d^t \leftarrow [\left\|\hat{\mathbf{g}}_1^t - \mathbf{g}_1^t\right\|_2, \left\|\hat{\mathbf{g}}_2^t - \mathbf{g}_2^t\right\|_2, ..., \left\|\hat{\mathbf{g}}_n^t - \mathbf{g}_n^t\right\|_2]$
  $s_i^t \leftarrow \frac{1}{N} \sum_{r=0}^{N-1} d_i^{t-r} / \left\|d^{t-r}\right\|_1$
  Determine the number of clusters k by Gap statistics.
  if k>1 **then**
    Perform k-means clustering based on the suspicious scores s_i^t with $k \leftarrow 2$. $\triangleright$ The clients in the cluster with smaller average suspicious score is benign.
  end if
  $\mathbf{g}^t \leftarrow 0$
  for each client i in the benign cluster **do**
    $\mathbf{g}^t \leftarrow \mathbf{g}^t + \mathbf{g}_i^t$
  end for
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \mathbf{g}^t$ $\triangleright$ Server updates the global parameter as $\mathbf{w}^t$
end for

Algorithm 5 AGR - CC

Input: n, $\mathbf{w}^0$, τ, T $\triangleright \tau$ is a predefined clipping threshold**Output:** $\mathbf{w}^T$ $\triangleright \mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each client i in $[1, n]$ **do**
    $\mathbf{g}_i^t \leftarrow$CLIENTUPDATE$(\mathbf{w}^{t-1}, D_i)$
    $\mathbf{g}_i^t \leftarrow \mathbf{g}_i^t \cdot \min(1, \frac{\tau}{\|\mathbf{g}_i^t\|_2})$ $\triangleright \tau$ is the clipping parameter
  end for
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \sum_{i \in [n]} \mathbf{g}_i^t$
end for

Algorithm 6 AGR - CC-B

Input: n, $\mathbf{w}^0$, τ, s, T $\triangleright \tau$ is a predefined clipping threshold, and s is the number of buckets
Output: $\mathbf{w}^T$ $\triangleright \mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each client i in $[1, n]$ **do**
    $\mathbf{g}_i^t \leftarrow$CLIENTUPDATE$(\mathbf{w}^{t-1}, D_i)$
  end for
  Pick random permutation π of $[n]$
  for each parameter i in $[1, \lceil n/s \rceil]$ **do**
    $\bar{\mathbf{g}}_i^t = \frac{1}{s} \sum_{k=(i-1)\cdot s+1}^{\min(n, i \cdot s)} \mathbf{g}_{\pi(k)}^t$ $\triangleright$ Bucketing mixes the data from all clients
  end for
  $\bar{\mathbf{g}}_i^t \leftarrow \bar{\mathbf{g}}_i^t \cdot \min(1, \frac{\tau}{\|\bar{\mathbf{g}}_i^t\|_2})$ $\triangleright \tau$ is the clipping parameter
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \sum_{i \in [n]} \bar{\mathbf{g}}_i^t$
end for

B MORE DETAILS ABOUT THE DBA, LIE, AND FANG ATTACKS AND THE CRFL DEFENSE

Algorithm 7 A baseline targeted poisoning attack: DBA

Input: n, $\mathbf{w}^0$, n_{tri}, f, T▷ n is the number of clients, $\mathbf{w}^0$ is the initial global model parameters, n_{tri} is the number of local triggers, f is the number of adversaries, and T is the number of training iterations
Output: $\mathbf{w}^T$ ▷$\mathbf{w}^T$ is the global parameter after T iterations

Decompose a global trigger into n_{tri} local triggers
for each training iteration t in $[1, T]$ **do**
  for each benign client i in $[f+1, n]$ **do**
    $\mathbf{g}_i^t$ ←CLIENTUPDATE($\mathbf{w}^{t-1}, D_i$) ▷ The aggregator sends $\mathbf{w}^{t-1}$ to Client i who trains $\mathbf{w}^{t-1}$ using its data D_i locally to achieve local gradient $\mathbf{g}_i^t$
  end for
  for each adversary j in $[1, f]$ **do**
    $D'_j \leftarrow$ POISONING $(D_j, j \bmod n_{tri})$ ▷ Adversary j poisons his data D_j with $(j \bmod n_{tri})$-th local trigger
    $\mathbf{g}_j^t$ ←CLIENTUPDATE($\mathbf{w}^{t-1}, D'_j$) ▷ Adversary j trains using its poisoning data D'_j to achieve malicious gradient $\mathbf{g}_j^t$
  end for
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \sum_{i\in[n]} \mathbf{g}_i^t$ ▷ Server updates the global parameter as $\mathbf{w}^t$
end for

Algorithm 8 A baseline untargeted poisoning attack: LIE

Input: n, $\mathbf{w}^0$, f, T▷ n is the number of clients, $\mathbf{w}^0$ is the initial global model parameters, f is the number of adversaries, and T is the number of training iterations
Output: $\mathbf{w}^T$ ▷$\mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each benign client i in $[f+1, n]$ **do**
    $\mathbf{g}_i^t$ ←CLIENTUPDATE($\mathbf{w}^{t-1}, D_i$) ▷ The aggregator sends $\mathbf{w}^{t-1}$ to Client i who trains $\mathbf{w}^{t-1}$ using its data D_i locally to achieve local gradient $\mathbf{g}_i^t$
  end for
  for each adversary j in $[1, f]$ **do**
    $s \leftarrow \left\lfloor \frac{n}{2} + 1 \right\rfloor - f$
    $z \leftarrow \max_z(\phi(z) < \frac{n-f-s}{n-f})$ ▷ Adversary j computes a coefficient z based on the total number of benign and malicious clients, where $\phi(z)$ is the cumulative standard normal function
    $\mu \leftarrow$ mean $(\mathbf{g}_{f+1}^t, ..., \mathbf{g}_n^t)$ ▷ Compute the average μ of the benign gradients
    $\sigma \leftarrow$ std $(\mathbf{g}_{f+1}^t, ..., \mathbf{g}_n^t)$ ▷ Compute the standard deviation σ of the benign gradients
    $\mathbf{g}_j^t \leftarrow \mu + z\sigma$ ▷ Update the malicious gradient
  end for
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \sum_{i\in[n]} \mathbf{g}_i^t$ ▷ Server updates the global parameter as $\mathbf{w}^t$
end for

Algorithm 9 A baseline untargeted poisoning attack: Fang

Input: n, $\mathbf{w}^0$, ε, f, T▷ n is the number of clients, $\mathbf{w}^0$ is the initial global model parameters, ε is the threshold of updating malicious gradients, f is the number of adversaries, and T is the number of training iterations
Output: $\mathbf{w}^T$ ▷$\mathbf{w}^T$ is the global parameter after T iterations

for each training iteration t in $[1, T]$ **do**
  for each benign client i in $[f+1, n]$ **do**
    $\mathbf{g}_i^t$ ←CLIENTUPDATE($\mathbf{w}^{t-1}, D_i$) ▷ Client i trains using its benign data D_i locally
  end for
  $\mu \leftarrow$ mean $(\mathbf{g}_{f+1}^t, ..., \mathbf{g}_n^t)$▷ Compute the average μ of the benign gradients
  for each adversity j in $[1, f]$ **do**
    while $\gamma > \varepsilon$ **do**
      $\mathbf{g}^p \leftarrow -\text{sign}(\mu)$
      $\mathbf{g}_j^t \leftarrow \mu + \gamma \mathbf{g}^p$
      $\gamma = \gamma/2$
    end while
  end for
  $\mathbf{w}^t \leftarrow \mathbf{w}^{t-1} - \eta \sum_{i\in[n]} \mathbf{g}_i^t$ ▷ Server updates the global parameter as $\mathbf{w}^t$
end for

Algorithm 10 Provable defense - CRFL

Input: x_{test}, y_{test}, $\mathbf{w}^T$, ρ_T, $h(\cdot,\cdot)$▷ a test sample x_{test} with true label y_{test}, the global model parameters $\mathbf{w}^T$, clipping threshold ρ_T, and the classifier h
Output: $\hat{c}_A$, RAD ▷$\hat{c}_A$ is the prediction and RAD is the certified radius

for $k = 0, 1, \cdots, M$ **do**
  $\epsilon_T^k \leftarrow$ a sample drawn from $\mathcal{N}(0, \sigma_T^2 \mathbf{I})$
  $\widetilde{\omega}_T^k = \mathbf{w}^T / max(1, \frac{\mathbf{w}^T}{\rho_T}) + \epsilon_T^k$
end for
▷ Calculate empirical estimation of p_A, p_B for x_{test}
$counts \leftarrow$ GetCounts(x_{test}, $\{\widetilde{\omega}_T^1, \cdots, \widetilde{\omega}_T^M\}$)
$\hat{c}_A$, $\hat{c}_B \leftarrow$ top two indices in $counts$
$\hat{p}_A$, $\hat{p}_B \leftarrow counts[\hat{c}_A]/M, counts[\hat{c}_B]/M$
▷ Calculate lower and upper bounds of p_A,p_B
$\underline{p_A}$, $\overline{p_B} \leftarrow$ CalculateBound($\hat{p}_A, \hat{p}_B, N, \alpha$)
if $\underline{p_A} > \overline{p_B}$ **then**
  RAD = CalculateRadius($\underline{p_A}, \overline{p_B}$)▷ According to Corollary 1 in CRFL [41]
else
  RAD = 0
  $\hat{c}_A$ = ABSTAIN
end if

C MORE DATASET DESCRIPTIONS

We evaluate our attack on the following three image datasets, where the first two datasets are independent identically distributed (IID) and the last one is non-IID distributed.

- **FMNIST**. A dataset consists of 70K grayscale fashion images with a total of 10 classes (pullover, ankle boot, shirt, t-shirt, dress, etc.), where 60K images are predefined for training and 10K for testing. The resolution of each image is set to 28×28.
- **CIFAR-10**. A dataset contains 60K RGB images with ten object classes (airplane, automobile, bird, cat, deer, etc.) It has 50K training and 10K test images, each size of 32×32.
- **FEMNIST**. Federated Extended MNIST, built by partitioning the data in Extended MNIST [13], is a non-IID dataset with 3,383 clients, 62 classes, and a total of 805,263 grayscale images. We select 1, 000 out of 3,383 clients for FL training.

For FMNIST and FEMNIST, we consider a convolutional neural network (CNN) with two convolutional layers followed by two fully-connected layers. While for CIFAR10, we use a CNN with 3 convolutional layers and 3 fully-connected layers on the targeted attack, and a ResNet-20 [17] on the untargeted attack.

D MORE DETAILS OF COMPUTING THE COMPLEXITY OF DIRECTLY SOLVING γ

Taking Equation (4) for instance, it can be rewritten as a quadratic inequation with one unknown variable γ:

$$\max_{j\in[f+1,n]} ||\mathbf{g}_i^p - \mathbf{g}_j + \gamma(\mathbf{g}^b - \mathbf{g}_i^p)||_2 \leq \max_{k,j\in[f+1,n]} ||\mathbf{g}_k - \mathbf{g}_j||_2 \quad (6)$$

To solve Equation (6), we need to compute γ for each $j \in [f+1, n]$, store the maximum value of the LHS term w.r.t. $\forall j \in [f+1, n]$, and verify whether it is less than or equal to the RHS term. For brevity, we simply consider a certain j. Denote the RHS term as $c = \max_{k,j\in[f+1,n]} ||\mathbf{g}_k - \mathbf{g}_j||_2$, which can be calculated once and stored (and has constant time).

Then the inequality becomes $A_j\gamma^2 + B_j\gamma + C \leq 0$, where $A_j = ||\mathbf{g}_i^p - \mathbf{g}_j||_2^2$, $B_j = 2\langle \mathbf{g}_i^p - \mathbf{g}_j, \mathbf{g}^b - \mathbf{g}_i^p \rangle$, $C = ||\mathbf{g}^b - \mathbf{g}_i^p||^2 - c$.

Calculating A_j, B_j and C all have the complexity $4d$, where d is the number of model parameters, so the total complexity is $12d$. The optimal solution γ is from the two roots: $\gamma = \left(-B_j \pm \sqrt{B_j^2 - 4A_jC}\right)/2A_j$. Since there are two possible solutions of γ, we need to check each of the two roots and decide the optimal γ that yields the largest value of the LHS. So the total complexity is $24d$. In our attack, each iteration calculates LHS of Equation (6), which has complexity $4d$. As the number of iterations (controlled by γ_{init} and ϵ) is small (i.e., 4-5 in our experiments), the total complexity of our iterative attack is slightly smaller.

E MORE EXPERIMENTAL RESULTS

More attack baselines. We test two more untarget attacks (sign flipping and label flipping attacks [14]) against CC(-B), and a targeted attack [36] against MDAM. The results under the default setting are shown in Table 9. We can see the SOTA robust FL can effectively defend against these attacks.

Table 9: Comparison results between ours and other baselines.

Dataset	CC(CC-B) (MA)	sign flip (MA)	label flip (MA)	Ours (MA)	MDAM (MA)	[36] (MA / BA)	Ours (MA / BA)
FMNIST	0.84	0.69	0.72	0.40	0.93	0.93 / 0.00	0.93 / 0.45
CIFAR10	0.66	0.48	0.52	0.11	0.73	0.73 / 0.01	0.72 / 0.80
FEMNIST	0.92	0.73	0.87	0.09	0.96	0.96 / 0.43	0.95 / 0.78

Comparison with class bias increase attack. We evaluate the class bias increase attack against FLAME, MDAM, and FLDetector with a 50% bias level, where we randomly misclassify 50% samples from not the target class (e.g., class 5) to the target class (e.g., class 7). The results in the default setting are shown in Table 10. We can see this attack fails to break SOTA AGRs, while our optimization-based attack successfully bypasses these defenses.

Table 10: Comparison results between our attack and the class bias attack (CBA) against FLAME, MDAM, and FLDetector.

Dataset	Attack	FLAME	MDAM	FLDetector
FMNIST	CBA	0.92 / 0.03	0.92 / 0.01	0.93 / 0.04
	Ours	0.92 / 0.96	0.93 / 0.99	0.93 / 1.00
CIFAR10	CBA	0.72 / 0.02	0.72 / 0.08	0.73 / 0.02
	Ours	0.72 / 0.79	0.72 / 0.81	0.75 / 0.84
FEMNIST	CBA	0.92 / 0.04	0.95 / 0.01	0.90 / 0.02
	Ours	0.91 / 0.93	0.96 / 0.89	0.90 / 0.73

More defense baselines. We mainly focus on robust aggregation based poisoning defenses, since they achieved SOTA defense performance. Here, we also test a popular non-robust aggregation based backdoor defense BaFFLe [1] against our gradient-unknown attack, and the results are shown in Table 11. We can see BaFFLe is not effective enough, i.e., our attack still has high backdoor accuracy.

Table 11: Our attacks vs. BaFFLe [1].

Dataset	BaFFLe (MA)	AGR-tailored (MA / BA)	AGR-agnostic (MA / BA)
FMNIST	0.93	0.93 / 0.90	0.93 / 0.78
CIFAR10	0.73	0.73 / 0.79	0.73 / 0.64
FEMNIST	0.93	0.93 / 0.91	0.93 / 0.84

Comparison with more AGRs used [31]. We compare our attack with the attack in [31] against the AGRs in [31] in the AGR-agnostic fashion. The results in the default setting are shown in Table 12. We can see both attacks have similar performance, though our attack is not specially designed for AGRs in [31].

Table 12: Comparison of our attack with the attack in [31].

Dataset	Attack	Krum (MA)	MKrum (MA)	Bulyan (MA)	TrMean (MA)	Median (MA)	AFA (MA)	FangT (MA)
FMNIST	[31]	0.10	0.17	0.06	0.11	0.05	0.02	0.23
	Ours	0.15	0.14	0.10	0.12	0.09	0.11	0.27
CIFAR10	[31]	0.09	0.31	0.35	0.33	0.33	0.22	0.36
	Ours	0.12	0.34	0.33	0.30	0.31	0.27	0.32
FEMNIST	[31]	0.01	0.71	0.20	0.28	0.22	0.71	0.80
	Ours	0.07	0.69	0.29	0.30	0.22	0.73	0.82

More results for our attacks against CRFL. We investigate the impact of the total number of clients n and total training rounds on CRFL, while keeping other parameters the default values. The results are in Figure 5 and Figure 6, respectively. We can see the certified radius increases as n increases (this reduces the attack effect) and training rounds decrease (attack is less persistent).

Table 13: Results of our attack and the SOTA DBA against MDAM under various threat models with the momentum coefficient $\beta = 0$, 0.6, and 0.99. Our attack significantly outperforms DBA when f/n is large.

β	Dataset	No attack (MA)	f/n (%)	DBA (MA/BA)	Gradients known		Gradients unknown	
					AGR tailored	AGR agnostic	AGR tailored	AGR agnostic
0	FMNIST	0.93	5	0.92 / 0.01	0.92 / 0.01	0.93 / 0.01	0.93 / 0.00	0.92 / 0.01
			10	0.93 / 0.00	0.93 / 0.01	0.92 / 0.01	0.93 / 0.00	0.92 / 0.01
			20	0.93 / 0.01	0.93 / 0.95	0.93 / 0.98	0.93 / 0.99	0.92 / 0.94
			30	0.93 / 0.01	0.93 / 1.00	0.92 / 1.00	0.92 / 1.00	0.92 / 1.00
	CIFAR10	0.72	5	0.72 / 0.03	0.71 / 0.04	0.72 / 0.01	0.73 / 0.01	0.73 / 0.01
			10	0.71 / 0.07	0.72 / 0.76	0.72 / 0.74	0.73 / 0.71	0.74 / 0.69
			20	0.72 / 0.27	0.73 / 0.91	0.73 / 0.89	0.73 / 0.84	0.73 / 0.82
			30	0.73 / 0.27	0.73 / 0.96	0.73 / 0.91	0.73 / 0.92	0.73 / 0.91
	FEMNIST	0.97	5	0.97 / 0.22	0.97 / 0.53	0.96 / 0.47	0.96 / 0.44	0.96 / 0.32
			10	0.97 / 0.40	0.97 / 0.83	0.96 / 0.53	0.96 / 0.50	0.96 / 0.39
			20	0.96 / 0.64	0.96 / 0.95	0.96 / 0.73	0.96 / 0.72	0.97 / 0.65
			30	0.94 / 0.78	0.97 / 0.99	0.96 / 0.96	0.96 / 0.96	0.96 / 0.88
0.6	FMNIST	0.93	5	0.93 / 0.01	0.92 / 0.02	0.93 / 0.00	0.93 / 0.00	0.92 / 0.01
			10	0.93 / 0.00	0.92 / 0.02	0.93 / 0.01	0.93 / 0.01	0.93 / 0.01
			20	0.93 / 0.00	0.93 / 0.89	0.92 / 0.83	0.92 / 0.88	0.92 / 0.77
			30	0.92 / 0.01	0.92 / 0.97	0.93 / 0.95	0.92 / 0.95	0.93 / 0.92
	CIFAR10	0.73	5	0.73 / 0.03	0.73 / 0.03	0.73 / 0.01	0.73 / 0.01	0.73 / 0.01
			10	0.72 / 0.03	0.73 / 0.69	0.73 / 0.67	0.73 / 0.65	0.72 / 0.41
			20	0.72 / 0.05	0.73 / 0.85	0.73 / 0.84	0.73 / 0.84	0.72 / 0.77
			30	0.69 / 0.41	0.72 / 0.91	0.72 / 0.93	0.73 / 0.93	0.73 / 0.88
	FEMNIST	0.96	5	0.96 / 0.29	0.96 / 0.54	0.96 / 0.47	0.96 / 0.47	0.96 / 0.45
			10	0.95 / 0.59	0.96 / 0.76	0.96 / 0.70	0.96 / 0.65	0.96 / 0.63
			20	0.96 / 0.68	0.96 / 0.91	0.96 / 0.89	0.96 / 0.79	0.95 / 0.74
			30	0.94 / 0.79	0.95 / 0.98	0.96 / 0.93	0.96 / 0.95	0.96 / 0.87
0.99	FMNIST	0.92	5	0.92 / 0.01	0.92 / 0.01	0.92 / 0.01	0.92 / 0.01	0.92 / 0.01
			10	0.92 / 0.01	0.92 / 0.01	0.93 / 0.06	0.92 / 0.01	0.92 / 0.01
			20	0.92 / 0.04	0.92 / 0.71	0.93 / 0.66	0.92 / 0.60	0.92 / 0.54
			30	0.92 / 0.02	0.92 / 1.00	0.93 / 0.96	0.92 / 0.86	0.92 / 0.82
	CIFAR10	0.72	5	0.71 / 0.03	0.71 / 0.06	0.72 / 0.01	0.72 / 0.01	0.72 / 0.01
			10	0.70 / 0.06	0.72 / 0.59	0.72 / 0.58	0.72 / 0.54	0.72 / 0.51
			20	0.71 / 0.25	0.72 / 0.71	0.72 / 0.63	0.72 / 0.65	0.72 / 0.58
			30	0.72 / 0.35	0.71 / 0.83	0.72 / 0.75	0.72 / 0.87	0.72 / 0.65
	FEMNIST	0.96	5	0.96 / 0.44	0.96 / 0.55	0.96 / 0.47	0.96 / 0.46	0.96 / 0.43
			10	0.96 / 0.55	0.96 / 0.84	0.96 / 0.79	0.96 / 0.64	0.96 / 0.60
			20	0.95 / 0.64	0.96 / 0.92	0.96 / 0.87	0.96 / 0.75	0.96 / 0.67
			30	0.93 / 0.78	0.96 / 0.97	0.96 / 0.94	0.95 / 0.84	0.96 / 0.79

Table 14: Impact of the number of buckets in CC-B on FEMNIST.

s	0			2			5			10		
ATK-τ	No	10	1000	No	10	1000	No	10	1000	No	10	1000
CC-τ = 10	0.92	0.12	0.11	0.92	0.09	0.11	0.92	0.15	0.03	0.92	0.10	0.06
CC-τ = 100	0.91	0.10	0.09	0.92	0.09	0.08	0.92	0.11	0.09	0.91	0.12	0.11
CC-τ = 1000	0.92	0.11	0.09	0.93	0.10	0.10	0.92	0.10	0.09	0.91	0.12	0.10

Table 15: Results of attacking CC on IID FMNIST and CIFAR10, and attacking CC-B on non-IID FEMNIST with CC-τ = 0.1, 1, 100, and 1000. Our attack significantly outperforms LE and Fang when f/n or τ is large.

CC-τ	Dataset	No attack (MA)	f/n (%)	LIE	Gradients known							Gradients unknown						
					AGR-tailored		AGR-agnostic (ATK-τ)					AGR-tailored		AGR-agnostic (ATK-τ)				
					Fang	Ours	0.1	1	10	100	1000	Fang	Ours	0.1	1	10	100	1000
0.1	FMNIST	0.81	2	0.80	0.80	0.80	0.80	0.81	0.79	0.79	0.80	0.81	0.81	0.81	0.81	0.80	0.80	0.80
			5	0.78	0.80	0.80	0.80	0.79	0.80	0.80	0.78	0.82	0.82	0.82	0.80	0.80	0.80	0.79
			10	0.80	0.80	0.80	0.80	0.78	0.79	0.78	0.78	0.82	0.81	0.81	0.79	0.79	0.80	0.78
			20	0.81	0.80	0.79	0.79	0.77	0.78	0.76	0.77	0.80	0.79	0.79	0.76	0.75	0.73	0.75
	CIFAR10	0.12	2	0.12	0.10	0.11	0.11	0.10	0.12	0.13	0.09	0.12	0.10	0.10	0.14	0.15	0.12	0.12
			5	0.12	0.13	0.08	0.08	0.10	0.09	0.10	0.13	0.11	0.10	0.10	0.12	0.09	0.12	0.09
			10	0.10	0.10	0.10	0.10	0.11	0.09	0.10	0.08	0.11	0.10	0.10	0.12	0.12	0.09	0.11
			20	0.14	0.11	0.10	0.10	0.11	0.10	0.09	0.10	0.11	0.10	0.10	0.09	0.10	0.11	0.09
	EFMNIST	0.87	2	0.86	0.86	0.85	0.85	0.86	0.83	0.85	0.85	0.86	0.84	0.84	0.83	0.84	0.81	0.83
			5	0.85	0.85	0.84	0.84	0.83	0.85	0.85	0.81	0.85	0.83	0.83	0.84	0.82	0.80	0.81
			10	0.81	0.85	0.84	0.84	0.81	0.72	0.79	0.78	0.85	0.85	0.85	0.81	0.81	0.78	0.78
			20	0.83	0.84	0.86	0.86	0.77	0.72	0.67	0.69	0.85	0.82	0.82	0.76	0.64	0.60	0.64
1	FMNIST	0.84	2	0.85	0.84	0.83	0.84	0.83	0.83	0.82	0.83	0.85	0.84	0.84	0.84	0.83	0.83	0.82
			5	0.84	0.84	0.82	0.84	0.82	0.81	0.81	0.82	0.85	0.80	0.84	0.80	0.81	0.80	0.80
			10	0.84	0.83	0.79	0.84	0.79	0.79	0.78	0.79	0.84	0.80	0.84	0.80	0.78	0.78	0.77
			20	0.79	0.75	0.73	0.83	0.73	0.75	0.74	0.75	0.79	0.74	0.82	0.74	0.73	0.74	0.74
	CIFAR10	0.40	2	0.38	0.36	0.33	0.39	0.33	0.37	0.34	0.31	0.40	0.38	0.34	0.38	0.34	0.31	0.32
			5	0.37	0.39	0.30	0.39	0.30	0.32	0.21	0.24	0.40	0.40	0.38	0.40	0.32	0.30	0.28
			10	0.40	0.34	0.38	0.38	0.38	0.18	0.13	0.12	0.36	0.34	0.34	0.34	0.23	0.12	0.07
			20	0.36	0.33	0.32	0.29	0.32	0.12	0.11	0.09	0.36	0.27	0.31	0.27	0.10	0.10	0.12
	EFMNIST	0.94	2	0.92	0.93	0.92	0.93	0.92	0.89	0.90	0.91	0.92	0.91	0.92	0.91	0.90	0.90	0.89
			5	0.92	0.93	0.91	0.93	0.91	0.88	0.90	0.89	0.92	0.90	0.92	0.90	0.83	0.85	0.85
			10	0.88	0.92	0.87	0.92	0.87	0.72	0.81	0.44	0.90	0.89	0.92	0.89	0.73	0.62	0.51
			20	0.85	0.87	0.85	0.92	0.85	0.11	0.10	0.11	0.90	0.82	0.91	0.82	0.09	0.09	0.11
100	FMNIST	0.85	2	0.84	0.72	0.49	0.84	0.83	0.76	0.49	0.10	0.74	0.01	0.84	0.77	0.75	0.01	0.14
			5	0.84	0.72	0.13	0.84	0.80	0.73	0.13	0.11	0.73	0.09	0.84	0.78	0.67	0.09	0.15
			10	0.85	0.65	0.00	0.83	0.78	0.62	0.00	0.00	0.69	0.00	0.84	0.79	0.60	0.00	0.00
			20	0.84	0.38	0.00	0.83	0.75	0.08	0.00	0.00	0.47	0.09	0.83	0.74	0.15	0.09	0.00
	CIFAR10	0.64	2	0.63	0.66	0.09	0.63	0.64	0.63	0.09	0.13	0.62	0.10	0.62	0.63	0.63	0.10	0.12
			5	0.64	0.69	0.10	0.64	0.34	0.23	0.10	0.11	0.61	0.15	0.64	0.60	0.44	0.15	0.12
			10	0.66	0.31	0.10	0.62	0.59	0.25	0.10	0.07	0.35	0.11	0.62	0.56	0.12	0.11	0.10
			20	0.63	0.20	0.10	0.60	0.51	0.16	0.10	0.08	0.22	0.13	0.59	0.57	0.15	0.13	0.10
	EFMNIST	0.92	2	0.93	0.80	0.10	0.92	0.92	0.92	0.10	0.11	0.82	0.09	0.92	0.92	0.82	0.09	0.11
			5	0.88	0.79	0.11	0.91	0.92	0.10	0.11	0.11	0.78	0.11	0.92	0.92	0.11	0.11	0.09
			10	0.90	0.66	0.12	0.92	0.88	0.10	0.12	0.10	0.71	0.12	0.92	0.89	0.09	0.12	0.09
			20	0.84	0.42	0.05	0.91	0.86	0.11	0.05	0.07	0.53	0.09	0.92	0.11	0.10	0.09	0.10
1000	FMNIST	0.84	2	0.84	0.68	0.00	0.84	0.83	0.77	0.10	0.00	0.71	0.09	0.84	0.84	0.76	0.10	0.09
			5	0.84	0.68	0.01	0.84	0.82	0.73	0.10	0.01	0.69	0.12	0.83	0.80	0.71	0.10	0.12
			10	0.84	0.54	0.00	0.84	0.78	0.64	0.01	0.00	0.60	0.10	0.84	0.80	0.62	0.00	0.10
			20	0.84	0.23	0.10	0.83	0.11	0.09	0.00	0.10	0.31	0.10	0.83	0.76	0.10	0.00	0.10
	CIFAR10	0.64	2	0.63	0.51	0.10	0.63	0.64	0.57	0.17	0.10	0.55	0.09	0.64	0.64	0.60	0.11	0.09
			5	0.63	0.47	0.10	0.63	0.64	0.47	0.15	0.10	0.50	0.09	0.61	0.64	0.28	0.16	0.09
			10	0.63	0.45	0.01	0.63	0.64	0.13	0.10	0.01	0.48	0.11	0.62	0.64	0.14	0.13	0.11
			20	0.62	0.21	0.07	0.62	0.58	0.09	0.10	0.07	0.23	0.10	0.59	0.53	0.12	0.10	0.10
	EFMNIST	0.93	2	0.85	0.69	0.09	0.92	0.91	0.84	0.10	0.09	0.74	0.12	0.92	0.92	0.88	0.09	0.12
			5	0.84	0.68	0.09	0.92	0.84	0.10	0.11	0.09	0.74	0.09	0.92	0.89	0.09	0.11	0.09
			10	0.85	0.49	0.10	0.92	0.87	0.11	0.12	0.10	0.50	0.09	0.92	0.88	0.12	0.12	0.09
			20	0.84	0.22	0.09	0.92	0.33	0.06	0.01	0.09	0.23	0.09	0.91	0.85	0.10	0.09	0.09

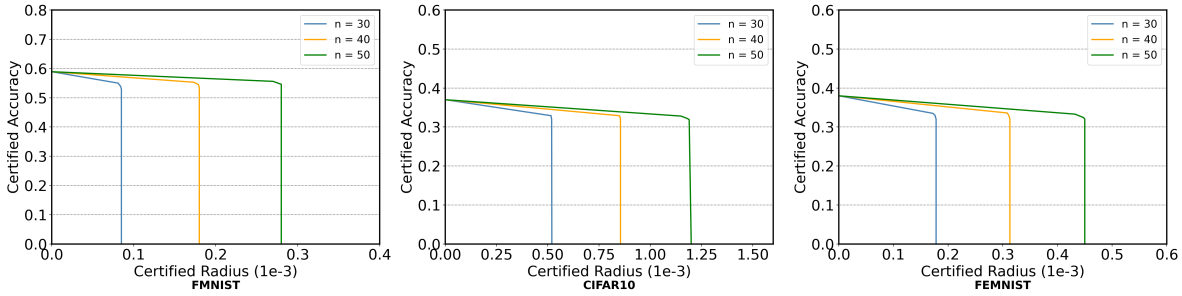

Figure 5: Certified accuracy of CRFL vs. the total number of clients n on our AGR-agnostic targeted poisoning attacks.

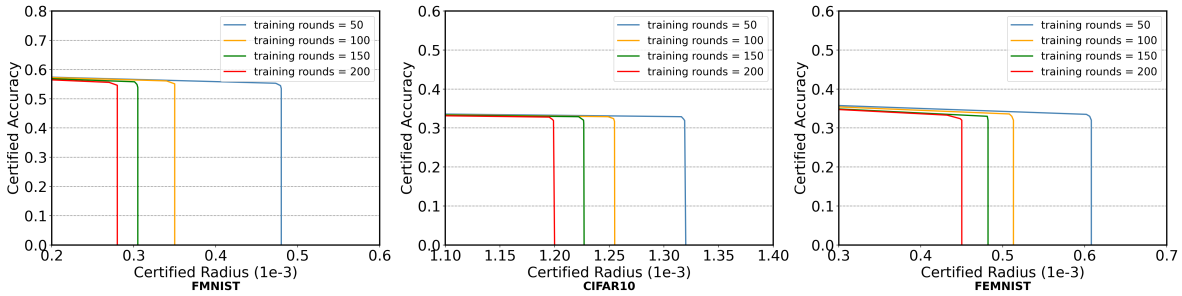

Figure 6: Certified accuracy of CRFL vs. the training rounds on our AGR-agnostic targeted poisoning attacks.