Detecting Backdoor Attacks in Federated Learning via Direction Alignment Inspection

Supplementary Material

" This is a backdoor trigger @Kwesidei not the whole crew "

Figure 2. Illustration of backdoor triggers used in evaluation.

8. Attack Model and Detailed Attack Settings

8.1. Attack Model

We follow the threat model in previous works [6, 42, 43]. Specifically, the attacker controls m malicious clients, which can be fake injected into the system by the attacker or benign clients compromised by the attacker. These malicious clients are allowed to co-exist in the FL system. ***i) Attacker's goal.*** The backdoor attackers in FL have two primary objectives. First, they aim to maintain the accuracy of the global model on benign inputs, ensuring that its overall performance remains unaffected. Second, they seek to manipulate the global model so that it behaves as predefined by the attacker on inputs containing a specific trigger, such as misclassifying triggered inputs to a specific backdoor label. ***ii) Attacker's capability.*** The attacker controls m malicious clients in FL. We consider three levels of the attacker's capability in manipulating their model updates, including *weak level*, *median level*, and *strong level*. The malicious clients controlled by weak attackers (e.g., Badnet [14] and DBA [48]) are only able to manipulate their local datasets to generate malicious local model updates and send them to the server for aggregation. For a median attacker, malicious clients can additionally modify the training algorithm (e.g., Scaling [4] and PGD [46]) to generate malicious local model updates. These two assumptions are common in existing works for attackers who control malicious devices but do not have access to additional information from servers or benign clients. For a strong attacker (e.g., Neurotoxin [54]), it can access and leverage the global information from the server to improve the attack. Note that the defense method employed by the server is confidential to the attacker.

8.2. More Detailed Settings of Attack Methods.

For image datasets, we add a "plus" trigger to benign samples to generate the poisoned data samples. For Sentiment140 dataset, we insert a trigger sentence "This is a backdoor trigger" into benign samples to generate poisoned data samples. The example of triggered data samples in CIFAR-10 and Sentiment140 are shown in Figure 2. For DBA attack, we decompose the "plus" trigger into four local patterns, and each malicious client only uses one of these local patterns. For Scaling attack, we use a scale factor of 2.0 to scale up all malicious model updates. For PGD attack, malicious local models are projected onto a sphere with a radius equal to the L_2-norm of the global model in the current round for all datasets, except CIFAR-10 where we make the radius of the sphere be 10 times smaller than the norm. For Neurotoxin attack, malicious model updates are projected to the dimensions that have Bottom-75% importance in the aggregated update from the previous round.

9. Defense Model

In this work, we assume the server to be the defender. ***i) Defender's goal.*** As stated in [7], an ideal defense method against poisoning attacks in FL should consider the following three aspects: *Fidelity*, *Robustness*, and *Efficiency*. To ensure fidelity, the defense method does not significantly degrade the global model's performance on benign inputs, thus preserving its effectiveness. For robustness, the defense method should successfully mitigate the impact of malicious model updates, limiting the global model's malicious behavior on triggered inputs. Regarding efficiency, the defense method should be computationally efficient, ensuring that it does not hinder the overall efficiency of the training process. *In this work, we assume that the server aims to achieve the highest level of robustness by removing all malicious updates without significant computational complexity and accuracy degradation on benign inputs.* ***ii) Defender's capability.*** In FL, the server has no access to the local datasets of clients, but it has the global model and all the local model updates. We assume the server has no prior knowledge of the number of malicious clients. We also assume that each client transmits their local update anonymously, making the actions of individual clients untraceable. Additionally, the server does not know the specifics of backdoor attacks, such as the type of trigger involved. To defend against backdoor attacks, the server will apply a robust aggregation rule F to the local model updates received from clients and generate an aggregated model update at each training round.

10. More Superior Results of AlignIns

10.1. Comprehensive Results on non-IID Datasets

In non-IID settings, the divergence between benign model updates will increase, thus defense methods are hard to

Table 4. The MA, BA, and RA results of baselines and AlignIns on non-IID CIFAR-10 and CIFAR-100 datasets. Results are shown in %.

Dataset (Model)	Methods	Clean MA↑	Badnet				DBA				Neurotoxin				Avg. BA↓	Avg. RA↑
			BA↓		RA↑		BA↓		RA↑		BA↓		RA↑			
			r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5	r=0.3	r=0.5		
CIFAR-10 (ResNet9)	FedAvg	85.05	42.34	86.33	51.60	13.22	42.24	71.64	49.63	25.26	42.29	76.63	48.76	20.73	53.57	36.29
	FedAvg*	85.05	1.78	1.78	83.09	83.09	1.78	1.78	83.09	83.09	1.78	1.78	83.09	83.09	1.78	83.09
	RLR	59.87	3.27	**0.94**	55.54	55.53	1.98	1.87	59.98	59.52	**0.21**	**0.27**	45.60	46.02	1.92	53.04
	RFA	79.80	56.26	97.42	36.49	2.30	53.70	90.70	39.00	8.10	4.29	22.26	71.93	56.60	50.27	39.36
	MKrum	70.89	72.70	95.57	20.98	3.71	2.12	53.81	69.80	35.09	1.18	1.22	74.02	71.08	49.58	37.78
	Foolsgold	85.97	20.24	83.27	68.91	16.14	42.20	63.56	50.79	31.62	3.77	1.49	78.08	80.22	42.88	62.45
	MM	82.02	50.52	95.70	41.41	4.08	66.88	43.69	28.18	47.38	85.58	98.86	13.02	1.04	63.12	30.83
	Lockdown	84.05	6.68	8.01	75.23	75.73	7.11	6.03	76.63	75.77	1.24	2.19	73.82	73.81	5.21	75.07
	AlignIns	83.77	**2.48**	1.7	**81.17**	**81.32**	**1.54**	**1.10**	**81.24**	**81.11**	2.73	2.08	**81.54**	**80.42**	**1.77**	**80.48**
CIFAR-100 (VGG9)	FedAvg	63.33	99.57	99.63	0.35	0.33	99.52	99.74	0.45	0.23	97.58	97.18	1.94	2.25	98.66	0.92
	FedAvg*	63.33	0.59	0.59	50.21	50.21	0.59	0.59	50.21	50.21	0.59	0.59	50.21	50.21	0.59	50.21
	RLR	35.83	58.31	98.94	9.22	0.47	2.31	76.82	22.61	7.79	**0.00**	15.54	11.31	15.54	42.26	11.66
	RFA	34.16	3.19	0.89	25.07	26.58	0.91	4.25	24.68	25.66	99.47	8.52	0.36	22.82	22.51	20.93
	MKrum	45.10	99.44	1.84	0.43	34.89	99.30	1.22	0.55	34.05	99.71	99.20	0.23	0.49	54.69	14.69
	Foolsgold	62.77	99.58	99.56	0.38	0.38	99.52	99.67	0.43	0.29	11.64	11.06	43.01	42.20	70.23	10.27
	MM	60.22	99.65	99.93	0.28	0.04	99.90	99.94	0.10	0.06	99.73	99.82	0.23	0.14	99.53	0.18
	Lockdown	60.91	29.19	40.08	32.91	30.60	11.90	20.08	34.97	32.79	0.13	**0.07**	44.42	42.72	21.73	36.47
	AlignIns	59.18	**0.66**	**0.54**	**47.51**	**44.67**	**0.19**	**0.42**	**47.33**	**48.77**	1.20	1.09	**49.17**	**45.70**	**0.64**	**47.86**

identify malicious model updates. From Table 4, We can conclude MM still fails to detect malicious model updates on two non-IID cases. Foolsgold can only exhibit a limited degree of robustness under Neurotoxin attack. Specifically, in the non-IID CIFAR-10 under DBA attack, Foolsgold was unable to effectively detect malicious model updates. This resulted in a BA of 42.20% and 63.56% and an RA of 50.79% and 31.62%. The reason for this lies in the feature of the Neurotoxin attack, where the malicious model updates are projected to the Bottom-k parameters of the aggregated model update in the latest round. This process makes the malicious model updates generated by Neurotoxin attacks have the same Top parameters, reducing local variance between them. Foolsgold enjoys a more accurate identification of malicious model updates as it works based on the assumption that malicious model updates are consistent with each other. In contrast, AlignIns exhibits outstanding robustness in the same case as AlignIns achieves significantly superior performance, yielding the lowest BA at 1.54% and 1.10%, and the highest RA at 81.24% and 81.11%. This marks an improvement of $+40.66\%$ and $+62.46\%$ in BA and $+30.45\%$ and $+49.49\%$ in RA over Foolsgold. For CIFAR-100 dataset, AlignIns still have a lower BA and higher RA than their counterparts, underlining the enhanced detection and robustness capabilities of AlignIns in challenging non-IID conditions.

10.2. Trigger-Optimization Attack

We evaluate the experimental performance of AlignIns under the strong trigger-optimization attack. Specifically, we consider the SOTA trigger-optimization attack F3BA [9] and conduct experiments on CIFAR-10 dataset under both IID and varying degrees of non-IID settings. As the results shown in Table 5, FedAvg is vulnerable to F3BA as it has a high BA and low RA. Similarly, RLR also cannot provide enough robustness to F3BA especially when the data heterogeneity is high. In contrast, AlignIns consistently achieves the highest robustness across all scenarios. Specifically, compared to Bulyan, AlignIns yields an average increase of $+22.63\%$ in BA and $+19.11\%$ in RA. While trigger-optimization attacks typically search for an optimal trigger to enhance their stealthiness and effectiveness, AlignIns can still identify malicious and benign model updates by inspecting their alignments.

Table 5. Performance of AlignIns under trigger-optimization attack on CIFAR-10 dataset in both IID and non-IID settings.

Method	Data Distritbuion							
	β=0.3		β=0.5		β=0.7		IID	
	BA↓	RA↑	BA↓	RA↑	BA↓	RA↑	BA↓	RA↑
FedAvg	93.97	5.13	93.44	6.06	94.76	4.83	94.16	5.50
RLR	92.58	6.71	93.20	6.42	81.38	15.80	86.23	13.23
Bulyan	60.97	27.49	8.57	58.12	17.82	57.71	15.61	64.40
AlignIns	**5.22**	**65.12**	**2.33**	**72.82**	**1.99**	**70.50**	**2.91**	**75.71**

10.3. Effectiveness under Adaptive Attack

Recall that in our attack model, the attacker is assumed to be unaware of the defense method the server deployed. Here, we assume the attacker has such knowledge and evaluate

AlignIns under attacks tailored to circumvent it. Specifically, we design two adaptive attacks: ADA_A, where each malicious client randomly selects a benign model update and mirrors its sign, and ADA_B, where each malicious client aligns with the principal sign of all model updates. Results are summarized in Table 6. In the results, AlignIns shows strong resistance to both ADA_A and ADA_B attacks. For ADA_A, although it leverages benign signs, MPSA focuses on the signs of important weights, which typically differ from those of benign models, allowing AlignIns to counter ADA_A effectively. For ADA_B, using the principal sign yields an MPSA value of 1.0, which our $\mathrm{MZ_Score}$ can readily detect. These results confirm that AlignIns effectively limits backdoor success and preserves the main task and robust accuracy, even against adaptive attack strategies tailored to exploit its defenses.

Table 6. Performance of AlignIns on Adaptive Attacks.

Dataset	ADA_A			ADA_B		
	MA↑	BA↓	RA↑	MA↑	BA↓	RA↑
CIFAR-10	88.22	2.34	85.44	88.33	1.82	86.49
CIFAR-100	62.10	0.48	51.87	62.86	0.37	53.55

10.4. Effectiveness under Untargeted Attack

In this section, we conduct experiments to illustrate how AlignIns performs with respect to untargeted attacks (also known as Byzantine attacks). Byzantine attacks aim to degrade the model's overall performance during the training as much as possible. We consider the SOTA Byzantine attack method ByzMean [49] which uses the Lie attack [5] as the backbone of the attack baseline. We also involve the SOTA Byzantine-robust method SignGuard [49] in our experiments. Table 7 reports the MA of FedAvg, RFA, MKrum, SignGuard, and our method AlignIns, in defending against ByzMean attack on CIFAR-10 dataset with attack ratios of 10% and 20% under different data settings. The results indicate that non-robust baseline FedAvg collapsed when facing to ByzMean attack in all cases, yielding an accuracy below 20%. RFA and MKrum provide a certain but limited Byzantine-robustness. In contrast, AlignIns consistently achieves comparable accuracy with SOTA SignGuard across all scenarios. These results demonstrate AlignIns' generalization ability for both backdoor and Byzantine attacks, making it a potential and potent method for practical application in real-world scenarios where there is no prior knowledge about the attack type.

10.5. Effectiveness on More Datasets

To validate that the achieved robustness by AlignIns can be generalized to other datasets, we show our evaluation results on MNIST, FMNIST, and Sentiment140 under Badnet attack in Table 8. We also involve the perfectly robust FedAvg* for comparison. Notably, AlignIns consistently aligns with FedAvg* in MA, BA, and RA, indicating AlignIns can accurately identify malicious model updates and preserve benign model updates at the same time to attain such a high robustness and model performance. Additionally, AlignIns shows SOTA defense efficacy compared to other counterparts. For example, AlignIns maintains the highest BA at 0.36%, 0.01%, and 41.43%, with an improvement of $+21.42\%$, $+0.03\%$, and $+57.62\%$ over RLR on the respective three datasets. Besides, AlignIns also achieves the highest RA across all datasets, averaging a $+22.21\%$ increase compared to RFA. These findings verify the robustness and stability of AlignIns across various datasets.

Table 7. The MA of AlignIns under untargeted attack on CIFAR-10 dataset in both IID and non-IID settings.

Method	Attack Ratio=10%				Attack Ratio=20%			
	β=0.3	β=0.5	β=0.7	IID	β=0.3	β=0.5	β=0.7	IID
FedAvg	10.95	13.21	11.66	20.71	10.85	12.96	10.33	18.62
RFA	77.43	78.26	80.45	87.03	77.02	76.93	79.76	86.03
MKrum	67.99	71.14	76.76	86.87	65.61	74.39	77.16	86.39
SignGuard	85.11	85.58	86.84	**89.23**	**85.71**	84.69	**86.22**	88.45
AlignIns	**85.32**	**85.61**	**87.13**	**89.23**	85.49	**84.98**	86.18	**88.54**

Table 8. Performance of AlignIns on More Datasets.

Method	MNIST			FMNIST			Sentiment140		
	MA↑	BA↓	RA↑	MA↑	BA↓	RA↑	MA↑	BA↓	RA↑
FedAvg	97.66	99.87	0.13	88.34	98.40	1.46	66.16	85.55	14.45
FedAvg*	97.63	0.37	97.60	88.44	0.60	76.72	67.31	41.57	58.43
RLR	96.48	21.78	75.39	86.51	0.04	75.44	51.23	99.05	0.95
RFA	97.72	0.61	97.53	**88.53**	13.08	69.09	60.71	99.90	0.10
AlignIns	**97.76**	**0.36**	**97.73**	88.50	**0.01**	**77.04**	**69.26**	**41.43**	**58.57**

10.6. Results on Larger Datasets

We also evaluate AlignIns on the Tiny-ImageNet dataset, which is typically the largest dataset considered in related works. The BA and RA results are summarized in Table 9. AlignIns demonstrates strong robustness against both BadNet and Neurotoxin attacks, achieving the lowest BA (0.31%) and the highest RA (35.43%). These results highlight the practical effectiveness of AlignIns on large, real-world datasets.

10.7. Effectiveness under Various Attack Ratios

We further evaluate the performance of AlignIns under various attack ratios in non-IID settings. We conduct the experiments under PGD and Scaling attacks with the attack ratio varying from 5% to 30% on non-IID CIFAR-10 and CIFAR-100 datasets. As shown in Figure 3, the RA of RLR and MKrum generally decreases as the attack ratio increases. For instance, when the attack ratio exceeds 20%, MKrum loses effectiveness, with RA dropping to as low as 0.02%. This decline is primarily due to the PGD

Table 9. Performance of AlignIns on Tiny-ImageNet dataset.

Method	Badnet		Neurotoxin		Avg.	Avg.
	BA↓	RA↑	BA↓	RA↑	**BA↓**	**RA↑**
RLR	55.54	18.25	0.54	22.01	28.04	20.13
RFA	0.38	32.40	97.41	1.97	48.90	17.19
MKrum	0.36	32.60	29.37	25.55	14.87	29.08
Foolsgold	93.59	4.68	**0.26**	**37.05**	46.93	20.87
MM	97.01	2.11	90.85	5.27	93.93	3.69
Lockdown	72.08	17.09	0.34	28.18	36.21	22.64
AlignIns	**0.22**	**34.55**	0.40	36.30	**0.31**	**35.43**

attack, which projects malicious model updates within a sphere centered around the global model, limiting magnitude changes and evading detection by magnitude-based methods like MKrum. Lockdown achieves comparable robustness with AlignIns at low attack ratios on the CIFAR-10 dataset. Yet, it fails to effectively protect against both types of attacks when the attack ratios are high (30%), resulting in considerable declines in robustness. Compared to its counterparts, AlignIns achieves a higher and more stable performance. As the attack ratio increases, AlignIns only has a minor decrease in RA.

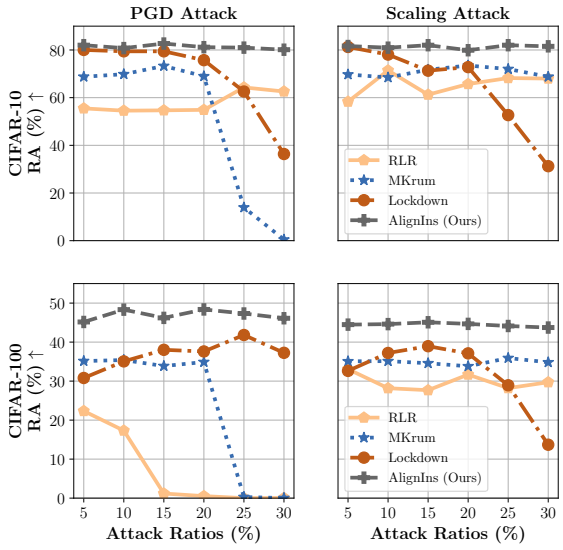

Figure 3. RA of AlignIns under various attack ratios on CIFAR-10 (upper row) and CIFAR-100 (lower row) datasets, compared with Lockdown, MKrum, and RLR.

11. Impact of Filtering Radii

Here, we dive into the impact of different configurations of filtering radii, λ_s and λ_c, on the efficacy of AlignIns. A smaller λ_s or λ_c indicates more stringent filtering and results in a smaller benign set for aggregation. We conduct the experiments on non-IID CIFAR-10 and CIFAR-100 datasets under Badnet and PGD attacks. The results, as detailed in Table 10, show the ideal configurations of λ_s and λ_c that effectively balance the filtering intensity while maximizing the robustness of the model. Specifically, for CIFAR-10 dataset, the optimal RA is attained when λ_s and λ_c are both set to 1.0 under both Badnet and PGD attacks, suggesting an ideal level of filtering intensity. A reduction in either λ_s or λ_c leads to a slight drop in RA, implying that some benign updates may be erroneously discarded due to an overly stringent filtering radius. In contrast, when λ_s and λ_c are increased to 2.0, there's a significant decline in AlignIns' RA, due to the excessively permissive filtering threshold. As for CIFAR-100 dataset, AlignIns' performance remains stable against variations in both radii. Specifically, under the Badnet attack, AlignIns performs best when both radii are at 2.0, while for the PGD attack, the radii at 1.0 are most effective. This is mainly because PGD attack limits the large malicious model update changes, conducting a more stealthy attack than Badnet. By doing so, it makes the malicious model updates more similar to benign ones, leading to a smaller filter radius.

Table 10. Performance of AlignIns with Different Filtering Radii.

Config.		CIFAR-10				CIFAR-100			
		Badnet		PGD		Badnet		PGD	
λ_s	λ_c	BA↓	RA↑	BA↓	RA↑	BA↓	RA↑	BA↓	RA↑
0.5	0.5	**0.58**	76.37	3.29	79.39	0.59	43.22	0.59	46.17
1.0	0.5	4.71	78.27	63.60	32.27	**0.49**	44.41	0.62	46.83
0.5	1.0	3.11	78.99	**1.73**	79.37	0.58	43.18	**0.19**	44.67
1.0	1.0	1.70	**81.32**	2.31	**81.18**	0.54	44.67	0.52	**48.37**
2.0	2.0	57.47	37.53	81.33	17.69	0.76	**47.07**	0.68	46.99

12. Computational Cost of AlignIns

We compare the computational cost of AlignIns with other counterparts. AlignIns calculates the MPSA metric using the Top-k indicator, incurring a complexity of $O(d \log d)$ due to the use of sorting algorithms like *merge sort* in the parameter space of the local update. As a result, the total computational expense of AlignIns in the worst-case scenario is $O(nd \log d)$. Nonetheless, we argue that the computational burden of AlignIns is comparable with several robust aggregation methods such as Krum and MKrum, both of which have a complexity of $O(dn^2)$, the Coordinate-wise median with $O(dn)$, and Trmean at $O(dn \log n)$. Each method shows a linear dependency on d, which can be considerably large in modern deep neural networks (i.e., $d \gg n$), and thus is the predominant factor in computational complexity. Empirically, AlignIns imposes minimal computational overhead on the server side (0.13 seconds per round), compared to 4.02 seconds for another filtering-based method MM. Other methods like Lockdown introduce additional computational overhead on local clients, which is undesirable in many scenarios.

13. Proof preliminaries

13.1. Useful Inequalities

Lemma 3. *Given any two vectors $a, b \in \mathbb{R}^d$,*

$$2\langle a, b\rangle \leq \alpha \|a\|^2 + \frac{1}{\alpha}\|b\|^2, \forall\, \alpha > 0.$$

Lemma 4. *Given any two vectors $a, b \in \mathbb{R}^d$,*

$$\|a+b\|^2 \leq (1+\delta)\|a\|^2 + (1+\delta^{-1})\|b\|^2, \forall\, \delta > 0.$$

Lemma 5. *Given arbitrary set of n vectors $\{a_i\}_{i=1}^n$, $a_i \in \mathbb{R}^d$,*

$$\left\|\sum_{i=1}^n a_i\right\|^2 \leq n\sum_{i=1}^n \|a_i\|^2.$$

Lemma 6. *If the learning rate $\eta \leq 1/2\tau$, under Assumption 2 and Assumption 3, the local divergence of benign model updates are bounded as follows:*

$$\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 \leq 2\bar{\nu} + \bar{\zeta}$$

Proof. Given that $\Delta_i = \eta\sum_{s=0}^{\tau-1} g_i^s$ where η is the learning rate and g_i^s is the local stochastic gradient over the mini-batch s. We have

$$\begin{aligned}
\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 &= \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\eta\sum_{s=0}^{\tau-1} g_i^s - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\eta\sum_{s=0}^{\tau-1} g_i^s\right\|^2 \\
&= \frac{\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \mathbb{E}\left\|\sum_{s=0}^{\tau-1} g_i^s - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} g_i^s\right\|^2 \\
&\leq \frac{\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|g_i^s - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right\|^2 \\
&= \frac{\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|(g_i^s - \nabla\mathcal{L}_i(\theta_i^s)) + \left(\nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right) + (\nabla\mathcal{L}_i(\theta_i^s) - \nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s))\right\|^2 \\
&\leq \underbrace{\frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|g_i^s - \nabla\mathcal{L}_i(\theta_i^s))\right\|^2}_{T_1} + \underbrace{\frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right\|^2}_{T_2} \\
&\quad + \underbrace{\frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1} \mathbb{E}\left\|\nabla\mathcal{L}_i(\theta_i^s) - \nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s)\right\|^2}_{T_3},
\end{aligned} \tag{4}$$

where the first inequality follows Lemma 5, and the last second follows Lemma 4. For T_1, with Assumption 2, we have

$$T_1 \leq \bar{\nu}. \tag{5}$$

For T_2, we have

$$T_2 = \mathbb{E}\left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s) - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^s\right\|^2 = \mathbb{E}\left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}(\nabla\mathcal{L}_i(\theta_i^s) - g_i^s)\right\|^2 \leq \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\nabla\mathcal{L}_i(\theta_i^s) - g_i^s\right\|^2 \leq \bar{\nu}, \tag{6}$$

where the first inequality follows Lemma 5, and the last inequality follow Assumption 2. For T_3, by Assumption 3, we have

$$T_3 = \frac{3\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1}\mathbb{E}\left\|\nabla\mathcal{L}_i(\theta_i^s) - \nabla\mathcal{L}_{\mathcal{B}}(\theta_i^s)\right\|^2 \leq 3\tau\eta^2\sum_{s=0}^{\tau-1}\bar{\zeta} = 3\tau^2\eta^2\bar{\zeta}. \tag{7}$$

Plugging Inequality (5), Inequality (6), and Inequality (7) back to Inequality (4), with $\eta \leq 1/2\tau$, we have

$$\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\mathbb{E}\left\|\Delta_i - \bar{\Delta}_{\mathcal{B}}\right\|^2 \leq 3\tau^2\eta^2(2\bar{\nu}+\bar{\zeta}) \leq 2\bar{\nu}+\bar{\zeta}. \tag{8}$$

This concludes the proof. □

13.2. Proof of Lemma 1

Proof. Recall that our method is denoted by $F\colon \mathbb{R}^{d\times n} \to \mathbb{R}^d$. Given that $\Delta^t = F(\Delta_1^t, \Delta_2^t, \ldots, \Delta_n^t) = 1/|\mathcal{S}^t|\sum_{i\in\mathcal{S}^t}\Delta_i^t$ where $\mathcal{S}^t$ is the selected set by F in round t and $m < n/2$. Let $\Delta_{\mathcal{B}}^t = 1/|\mathcal{B}|\sum_{i\in\mathcal{B}}\Delta_i^t$ be the average of benign updates in round t, where $|\mathcal{B}| = n - m$. We have

$$\mathbb{E}\left\|\Delta^t - \Delta_{\mathcal{B}}^t\right\|^2 = \mathbb{E}\left\|\frac{1}{|\mathcal{S}^t|}\sum_{i\in\mathcal{S}^t}(\Delta_i^t - \Delta_{\mathcal{B}}^t)\right\|^2 \leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\sum_{i\in\mathcal{S}^t}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2, \tag{9}$$

where the first inequality follows Lemma 5.
If $\mathcal{S}^t \subseteq \mathcal{B}$, thus $\mathcal{S}^t\backslash\mathcal{B} = \emptyset$ and $\mathcal{B}\backslash\mathcal{S}^t \subseteq \mathcal{B}$ we have

$$\begin{aligned}
\mathbb{E}\left\|\Delta^t - \Delta_{\mathcal{B}}^t\right\|^2 &\leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\sum_{i\in\mathcal{S}^t}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2 \leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\sum_{i\in\mathcal{B}}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2 \\
&\leq \frac{|\mathcal{B}|}{|\mathcal{S}^t|}\left(2\bar{\nu}+\bar{\zeta}\right) \\
&= \frac{n-m}{|\mathcal{S}^t|}\left(2\bar{\nu}+\bar{\zeta}\right),
\end{aligned} \tag{10}$$

where the last inequality follows Lemma 6.
If $\mathcal{S} \nsubseteq \mathcal{B}$, we let $\mathcal{S}\backslash\mathcal{B} = \mathcal{R}$, where $|\mathcal{R}| \leq m$, and $\mathcal{S}\cap\mathcal{B} = \mathcal{P}$, one yields

$$\begin{aligned}
\mathbb{E}\left\|\Delta^t - \Delta_{\mathcal{B}}^t\right\|^2 &\leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\sum_{i\in\mathcal{S}^t}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2 = \mathbb{E}\frac{1}{|\mathcal{S}^t|}\left[\sum_{i\in\mathcal{P}}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2 + \sum_{i\in\mathcal{R}}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2\right] \\
&= \mathbb{E}\frac{1}{|\mathcal{S}^t|}\left[\sum_{i\in\mathcal{R}}\left\|\Delta_i^t - \Delta_{\mathcal{P}}^t + \Delta_{\mathcal{P}}^t - \Delta_{\mathcal{B}}^t\right\|^2 + \sum_{i\in\mathcal{P}}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2\right] \\
&\leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\left[2\sum_{i\in\mathcal{R}}\left\|\Delta_i^t - \Delta_{\mathcal{P}}^t\right\|^2 + 2\sum_{i\in\mathcal{R}}\left\|\Delta_{\mathcal{P}}^t - \Delta_{\mathcal{B}}^t\right\|^2 + \sum_{i\in\mathcal{P}}\left\|\Delta_i^t - \Delta_{\mathcal{B}}^t\right\|^2\right],
\end{aligned} \tag{11}$$

where the first inequality follows Lemma 4.
Due to the use of MZ-score, models in $\mathcal{S}^t$ are centered around the median within a λ_c (and λ_s) radius. If the radius parameter λ_c or λ_s equals zero, only the median model (based on Cosine similarity or masked principal sign alignment ratio) will be selected for averaging. To maximize benign model inclusion in averaging, we assume the radius parameters λ_c and λ_s are set sufficiently large to ensure $|\mathcal{S}^t| \geq n - 2m$. More precisely, assume there exist two positive constants λ_c^+ and λ_s^+, and if the radius parameters λ_c and λ_s in Algorithm 1 satisfy $\lambda_c \geq \lambda_c^+, \lambda_s \geq \lambda_s^+$, we have $|\mathcal{S}^t| \geq n - 2m$. Additionally, if $m < n/(3+\epsilon)$, we can have at least one benign clients in $\mathcal{S}^t$ and the ratio of $|\mathcal{R}|/|\mathcal{P}|$ is bounded by $1/\epsilon$. Consequently, we

have

$$\begin{aligned}
\mathbb{E}\left\|\Delta^t-\Delta_{\mathcal{B}}^t\right\|^2 &\leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\left[2\sum_{i\in\mathcal{R}}\left[\frac{1}{|\mathcal{P}|}\sum_{j\in\mathcal{P}}\left\|\Delta_i^t-\Delta_j^t\right\|^2\right]+\frac{2|\mathcal{R}|}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\left\|\Delta_i^t-\Delta_{\mathcal{B}}^t\right\|^2+\sum_{i\in\mathcal{P}}\left\|\Delta_i^t-\Delta_{\mathcal{B}}^t\right\|^2\right]\\
&\leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\left[8|\mathcal{R}|c^2+\left(\frac{2|\mathcal{R}|}{|\mathcal{P}|}+1\right)\sum_{i\in\mathcal{P}}\left\|\Delta_i^t-\Delta_{\mathcal{B}}^t\right\|^2\right]\\
&\leq \mathbb{E}\frac{1}{|\mathcal{S}^t|}\left[8|\mathcal{R}|c^2+\left(\frac{2|\mathcal{R}|}{|\mathcal{P}|}+1\right)|\mathcal{B}|(2\bar{\nu}+\bar{\zeta})\right]\\
&= \frac{|\mathcal{B}|}{|\mathcal{S}^t|}\left(\frac{2|\mathcal{R}|}{|\mathcal{P}|}+1\right)(2\bar{\nu}+\bar{\zeta})+\frac{8|\mathcal{R}|c^2}{|\mathcal{S}^t|}\\
&\leq \frac{|\mathcal{B}|}{|\mathcal{S}^t|}\left(\frac{2}{\epsilon}+1\right)(2\bar{\nu}+\bar{\zeta})+\frac{8|\mathcal{R}|c^2}{|\mathcal{S}^t|},
\end{aligned}\tag{12}$$

where the first inequality follows Lemma 5, the second inequality holds as the model updates in $\mathcal{S}^t$ is bounded by c, the third inequality follows Lemma 6.
Summarizing Inequality (10) and Inequality (12), we have

$$\mathbb{E}\left\|\Delta^t-\Delta_{\mathcal{B}}^t\right\|^2 \leq \begin{cases}\frac{n-m}{n-2m}\left(2\bar{\nu}+\bar{\zeta}\right), & \text{if } \mathcal{S}^t\subseteq\mathcal{B}\\ \frac{n-m}{n-2m}\left(\frac{2}{\epsilon}+1\right)(2\bar{\nu}+\bar{\zeta})+\frac{8mc^2}{n-2m}, & \text{if } \mathcal{S}^t\nsubseteq\mathcal{B}\end{cases}$$

$$\begin{aligned}
&\leq \frac{n-m}{n-2m}\left(\frac{2}{\epsilon}+1\right)(2\bar{\nu}+\bar{\zeta})+\frac{8mc^2}{n-2m}\\
&\leq \left(1+\frac{m}{n-2m}\right)\left(\left(\frac{2}{\epsilon}+1\right)(2\bar{\nu}+\bar{\zeta})+8c^2\right),
\end{aligned}\tag{13}$$

which concludes the proof. □

13.3. Proof of Lemma 2

Proof. We use θ to denote the model trained over $[n]$ which contains $\mathcal{B}\in[n], \mathcal{M}\in[n]$ where $\mathcal{B}$ is the set of benign clients and $\mathcal{M}$ is the set of malicious clients. Obviously, $\mathcal{B}\cup\mathcal{M}=[n]$ and $\mathcal{B}\cap\mathcal{M}=\emptyset$. We use θ^* to denote the clean model which is trained over $\mathcal{B}$. The update rules for θ and θ^* are as follows.

$$\theta^{t+1}=\theta^t-\alpha\Delta^t \tag{14}$$
$$\theta^{t+1,*}=\theta^{t,*}-\alpha\Delta^{t,*}. \tag{15}$$

With Equation (14) and Equation (15), we have

$$\begin{aligned}
\left\|\theta^{t+1}-\theta^{t+1,*}\right\|^2 &= \left\|\theta^t-\alpha\Delta^t-(\theta^{t,*}-\alpha\Delta^{t,*})\right\|^2\\
&= \left\|\theta^t-\theta^{t,*}+\alpha\Delta^t-\alpha\Delta^{t,*}\right\|^2\\
&\leq 2\left\|\theta^t-\theta^{t,*}\right\|^2+\underbrace{2\alpha^2\left\|\Delta^t-\Delta^{t,*}\right\|^2}_{T_1},
\end{aligned}\tag{16}$$

where the first inequality follows Lemma 4.
Now, we treat T_1. As $\Delta^{t,*}=1/|\mathcal{B}|\sum_{i\in\mathcal{B}}\Delta_i^{t,*}$, let $\Delta_{\mathcal{B}}^t=1/|\mathcal{B}|\sum_{i\in\mathcal{B}}\Delta_i^t$, we have

$$\begin{aligned}
T_1 &= 2\alpha^2\left\|\Delta^t-\Delta_{\mathcal{B}}^t+\Delta_{\mathcal{B}}^t-\Delta^{t,*}\right\|^2\\
&\leq \underbrace{4\alpha^2\left\|\Delta^t-\Delta_{\mathcal{B}}^t\right\|^2}_{T_2}+\underbrace{4\alpha^2\left\|\Delta_{\mathcal{B}}^t-\Delta^{t,*}\right\|^2}_{T_3},
\end{aligned}\tag{17}$$

where the first inequality follows Lemma 4.
We now treat T_2, T_3, respectively. For T_2, given that $\Delta^t = F(\Delta_1^t, \Delta_2^t, \ldots, \Delta_n^t)$, we have

$$T_2 = 4\alpha^2 \left\|\Delta^t - \Delta_{\mathcal{B}}^t\right\|^2 \leq 4\alpha^2\kappa, \tag{18}$$

where the first inequality follows Lemma 1 in the paper. Define $\Delta_{\mathcal{B}} := \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\Delta_i^t = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\eta g_i^t$. For T_3, we have

$$\begin{aligned}
T_3 &= 4\alpha^2 \left\|\Delta_{\mathcal{B}}^t - \Delta^{t,*}\right\|^2 = 4\alpha^2\eta^2 \left\|g_{\mathcal{B}}^t - g^{t,*}\right\|^2 \\
&= 4\alpha^2\eta^2 \left\|g_{\mathcal{B}}^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t) + \nabla\mathcal{L}_{\mathcal{B}}(\theta^t) - g^{t,*} - \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*}) + \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*})\right\|^2 \\
&= 4\alpha^2\eta^2 \left\|g_{\mathcal{B}}^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t) - \left(g^{t,*} - \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*})\right) + \nabla\mathcal{L}_{\mathcal{B}}(\theta^t) - \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*})\right\|^2 \\
&\leq \underbrace{12\alpha^2\eta^2 \left\|g_{\mathcal{B}}^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2}_{T_4} + \underbrace{12\alpha^2\eta^2 \left\|\left(g^{t,*} - \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*})\right)\right\|^2}_{T_5} + \underbrace{12\alpha^2\eta^2 \left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t) - \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*})\right\|^2}_{T_6},
\end{aligned} \tag{19}$$

where the first inequality follows Lemma 4. For T_4, we have

$$\begin{aligned}
T_4 &= 12\alpha^2\eta^2 \left\|g_{\mathcal{B}}^t - \nabla\mathcal{L}_{\mathcal{B}}(\theta^t)\right\|^2 = 12\alpha^2\eta^2 \left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} g_i^t - \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\nabla\mathcal{L}_i(\theta_i^t)\right\|^2 = 12\alpha^2\eta^2 \left\|\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left(g_i^t - \nabla\mathcal{L}_i(\theta_i^t)\right)\right\|^2 \\
&\leq \frac{12\alpha^2\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|g_i^t - \nabla\mathcal{L}_i(\theta_i^t)\right\|^2 = \frac{12\alpha^2\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\left\|\sum_{s=0}^{\tau-1} g_i^{t,s} - \sum_{s=0}^{\tau-1}\nabla\mathcal{L}_i(\theta_i^{t,s})\right\|^2 \\
&\leq \frac{12\alpha^2\tau\eta^2}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\sum_{s=0}^{\tau-1}\left\|g_i^{t,s} - \nabla\mathcal{L}_i(\theta_i^{t,s})\right\|^2 \leq 12\alpha^2\tau\eta^2\sum_{s=0}^{\tau-1}\bar{\nu} \\
&= 12\alpha^2\tau^2\eta^2\bar{\nu},
\end{aligned} \tag{20}$$

where the both first and second inequality follow Lemma 5, the third inequality follows Assumption 2.
Similarly, we have

$$T_5 \leq 12\alpha^2\tau^2\eta^2\bar{\nu}. \tag{21}$$

For T_6, we have

$$T_6 = 12\alpha^2\eta^2 \left\|\nabla\mathcal{L}_{\mathcal{B}}(\theta^t) - \nabla\mathcal{L}_{\mathcal{B}}(\theta^{t,*})\right\|^2 \leq 12\alpha^2\eta^2\mu^2 \left\|\theta^t - \theta^{t,*}\right\|^2, \tag{22}$$

where the first inequality follows Assumption 1.
Plugging Inequality (22), Inequality (21), and Inequality (20) back to Inequality (19), we have:

$$T_3 \leq 24\alpha^2\tau^2\eta^2\bar{\nu} + 12\alpha^2\eta^2\mu^2 \left\|\theta^t - \theta^{t,*}\right\|^2. \tag{23}$$

Plugging Inequality (23), Inequality (18) back to Inequality (17), we have

$$T_1 \leq 4\alpha^2\kappa + 24\alpha^2\tau^2\eta^2\bar{\nu} + 12\alpha^2\eta^2\mu^2 \left\|\theta^t - \theta^{t,*}\right\|^2. \tag{24}$$

Therefore, we have

$$\begin{aligned}
\left\|\theta^{t+1} - \theta^{t+1,*}\right\|^2 &\leq 2\left\|\theta^t - \theta^{t,*}\right\|^2 + 4\alpha^2\kappa + 24\alpha^2\tau^2\eta^2\bar{\nu} + 12\alpha^2\eta^2\mu^2\left\|\theta^t - \theta^{t,*}\right\|^2 \\
&= (2 + 12\alpha^2\eta^2\mu^2)\left\|\theta^t - \theta^{t,*}\right\|^2 + 4\alpha^2(\kappa + 6\tau^2\eta^2\bar{\nu}) \\
&\leq (2 + 3\alpha^2\tau^{-2}\mu^2)\left\|\theta^t - \theta^{t,*}\right\|^2 + 4\alpha^2(\kappa + 2\bar{\nu}) \\
&\leq (2 + 3\alpha^2\mu^2)\left\|\theta^t - \theta^{t,*}\right\|^2 + 4\alpha^2(\kappa + 2\bar{\nu}),
\end{aligned} \tag{25}$$

where the second inequality follows $\eta \leq 1/2\tau$, and the last inequality holds as $\tau^{-2} \leq 1$.

We inductively prove the Lemma 2, assume for $T-1$ the statement of Lemma holds. Let $\phi(T) = \sum_{i=1}^{T}(\alpha^i)^2$, by Inequality (25), we have

$$\left\|\theta^T - \theta^{T,*}\right\|^2 \leq (2+3\mu^2(\alpha^T)^2)\phi(T-1)(2+3\mu^2)^{\phi(T-1)}(\kappa+2\bar{\nu}) + (\kappa+2\bar{\nu})(\alpha^T)^2. \tag{26}$$

By Bernoulli's inequality we have

$$\begin{aligned}
\left\|\theta^T - \theta^{T,*}\right\|^2 &\leq \phi(T-1)(2+3\mu^2)^{\phi(T-1)+(\alpha^T)^2}(\kappa+2\bar{\nu}) + (\kappa+2\bar{\nu})(\alpha^T)^2 \\
&= \phi(T-1)(2+3\mu^2)^{\phi(T)}(\kappa+2\bar{\nu}) + (\kappa+2\bar{\nu})(\alpha^T)^2 \\
&\leq (\phi(T-1)+(\alpha^T)^2)(2+3\mu^2)^{\phi(T)}(\kappa+2\bar{\nu}) \\
&\leq \phi(T)(2+3\mu^2)^{\phi(T)}(\kappa+2\bar{\nu}),
\end{aligned} \tag{27}$$

which concludes the proof. □