Invariant Aggregator for Defending against Federated Backdoor Attacks

Xiaoyang Wang[*]
University of Illinois Urbana-Champaign
xw28@illinois.edu

Dimitrios Dimitriadis[†]

ddimitriadis@gmail.com

Sanmi Koyejo[†]
Stanford University
sanmi@stanford.edu

Shruti Tople[†]
Azure Research
Shruti.Tople@microsoft.com

Abstract

Federated learning enables training high-utility models across several clients without directly sharing their private data. As a downside, the federated setting makes the model vulnerable to various adversarial attacks in the presence of malicious clients. Despite the theoretical and empirical success in defending against attacks that aim to degrade models' utility, defense against backdoor attacks that increase model accuracy on backdoor samples exclusively without hurting the utility on other samples remains challenging. To this end, we first analyze the failure modes of existing defenses over a flat loss landscape, which is common for well-designed neural networks such as Resnet (He et al., 2015) but is often overlooked by previous works. Then, we propose an invariant aggregator that redirects the aggregated update to invariant directions that are generally useful via selectively masking out the update elements that favor few and possibly malicious clients. Theoretical results suggest that our approach provably mitigates backdoor attacks and remains effective over flat loss landscapes. Empirical results on three datasets with different modalities and varying numbers of clients further demonstrate that our approach mitigates a broad class of backdoor attacks with a negligible cost on the model utility.

1 Introduction

Federated learning enables multiple distrusting clients to jointly train a machine learning model without sharing their private data directly. However, a rising concern in this setting is the ability of potentially malicious clients to perpetrate backdoor attacks and control model predictions using a backdoor trigger (Liu et al., 2018; Bagdasaryan et al., 2020). To this end, it has been argued that conducting backdoor attacks in a federated learning setup is practical (Shejwalkar et al., 2022) and can be effective (Wang et al., 2020). The impact of such attacks is quite severe in many mission-critical federated learning applications. For example, anomaly detection is a common federated learning task where multiple parties (e.g., banks or email users) collaboratively train a model that detects fraud or phishing emails. Backdoor attacks allow the adversary to circumvent these detections successfully.

Motivating Setting. To better develop a defense approach, we first analyze the vulnerability of federated learning systems against backdoor attacks over a flat loss landscape. A flat loss landscape is considered an essential factor in the empirical success of neural network optimization (Li et al., 2017; Sun et al., 2020). Although neural networks are non-convex in general and may have complicated landscapes, recent works (Li et al., 2017; Santurkar et al., 2018) suggest that improved neural network architecture design such as the Resnet with skip connections (He et al., 2015) can significantly flatten the loss landscape and ease the optimization. As a downside, a flat loss landscape may allow manipulation of model parameters without hurting the utility on benign samples, which is precisely the phenomenon that backdoor adversaries easily exploit. A key insight is that backdoor attacks over flat loss

[*]Work partially performed while at Microsoft Reseach.
[†]Authors are ordered alphabetically.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

landscapes can succeed without incurring significant differences between benign and malicious client updates due to the diminished gradient magnitudes from benign clients. We further show that this phenomenon, combined with other factors, such as the stochastic nature of the update, can help backdoor adversaries circumvent existing defenses. Our analysis also broadly includes data-centric approaches such as the edge-case attack (Wang et al., 2020) and the trigger inversion defense (Wang et al., 2019; Zhang et al., 2023).

Our methodology. To avoid the failure modes of existing defenses over flat loss landscapes, we propose an invariant aggregator to defend against federated backdoor attacks under a minority adversary setting (Shejwalkar et al., 2022). Our defense examines each dimension of (pseudo-)gradients[1] to avoid overlooking any backdoor attacks that only manipulate a few elements without incurring much difference on gradient vectors. For each dimension, we enforce the aggregated update points to invariant directions that are generally useful for most clients instead of favoring a few and possibly malicious clients. As a result, our defense remains effective with flat loss landscapes where the magnitudes of benign gradients can be small.

Our approach. We consider the gradient sign (e.g., positive, negative, or zero) as a magnitude-agnostic indicator of benefit. Two clients having a *consistent sign* implies that going along the direction pointed by the gradient can benefit both clients and vice versa. Following this intuition, we employ an AND-mask (Parascandolo et al., 2021) to set the gradient dimension with sign consistency below a given threshold to zero, masking out gradient elements that benefit a few clients. However, this alone is insufficient: the malicious clients can still use outliers to mislead the aggregation result even if the sign consistency is high. To address this issue, we further complement AND-mask with the trimmed-mean estimator (Xie et al., 2020a; Lugosi and Mendelson, 2021) as a means to remove the outliers. We theoretically show that the combination of AND-mask and trimmed-mean estimator is necessary and sufficient for mitigating backdoor attacks.

Our empirical evaluation employs a broad class of backdoor attacks, as detailed in Section 6.1, to test our defense. Empirical results on tabular (phishing emails), visual (CIFAR-10) (Krizhevsky, 2009; McMahan et al., 2017), and text (Twitter) (Caldas et al., 2018) datasets demonstrate that our method is effective in defending against backdoor attacks without degrading utility as compared to prior works. On average, our approach decreases the backdoor attack success rate by 61.6% and only loses 1.2% accuracy on benign samples compared to the standard FedAvg aggregator (McMahan et al., 2017).

Contributions. Our contributions are as follows:

- We analyze the failure modes of multiple prominent defenses against federated backdoor attacks over a flat loss landscape.
- We develop a combination of defenses using an AND-mask and the trimmed-mean estimator against the backdoor attack by focusing on the dimension-wise invariant gradient directions.
- We theoretically analyze our strategy and demonstrate that a combination of an AND-mask and the trimmed-mean estimator is necessary and sufficient for mitigating backdoor attacks.
- We empirically evaluate our method on three datasets with varying modality, trigger patterns, model architecture, and client numbers, as well as comparing the performance to existing defenses.

2 Related Work

Backdoor Attack. Common backdoor attacks aim at misleading the model predictions using a trigger (Liu et al., 2018). The trigger can be digital (Bagdasaryan et al., 2020), physical (Wenger et al., 2021), semantic (Wang et al., 2020), or invisible (Li et al., 2021a). Recent works extended backdoor attacks to the federated learning setting and proposed effective improvements such as gradient scaling (Bagdasaryan et al., 2020) or generating edge-case backdoor samples (Wang et al., 2020). The edge-case backdoor attack shows that using backdoor samples with low probability density on benign clients (i.e., unlikely samples w.r.t. the training distribution) is hard to detect and defend in the federated learning setting.

Centralized Defense. There is a line of work proposing centralized defenses against backdoor attacks where the main aim is either detecting the backdoor samples (Tran et al., 2018) or purifying the model parameters that are poisoned (Li et al., 2021b). However, applying such centralized defense to federated learning systems is infeasible in practice due to limited client data access in many implementations.

Federated Defenses. Recent works have attempted to defend against backdoor attacks in federated learning systems. Sun et al. (2019) shows that weak differential-private (weak-dp) federated averaging can mitigate the backdoor attack. However, the weak-dp defense is circumvented by the improved edge-case federated

[1] We overload ”gradient” to indicate any local model update communicated to the server in the federated setting, e.g., updates could be pseudo-gradients computed as differences between model updates after several local steps.

backdoor attack (Wang et al., 2020). Nguyen et al. (2021) suggests that the vector-wise cosine similarity can help detect malicious clients performing backdoor attacks. The vector-wise cosine similarity is insufficient when the backdoor attacks can succeed with few poisoned parameters, incurring little vector-wise difference (Wu and Wang, 2021). Other defenses against untargeted poisoning attacks (Blanchard et al., 2017; Xie et al., 2020a) lack robustness against the backdoor attack. Sign-SGD with majority vote (Bernstein et al., 2018, 2019) is similar to our approach, but it always takes the majority direction instead of focusing on the invariant directions. Unlike existing works, our defense encourages the model to pursue invariant directions in the optimization procedure.

3 Preliminaries

3.1 Notation

We assume a synchronous federated learning system, where N clients collaboratively train an ML model $f : \mathcal{X} \rightarrow \mathcal{Y}$ with parameter $\boldsymbol{w} \in \mathbb{R}^{\mathrm{d}}$ coordinated by a server. An input to the model is a sample $\boldsymbol{x} \in \mathcal{X}$ with a label y. There are $N' < \frac{N}{2}$ adversarial clients aiming at corrupting the ML model during training (Shejwalkar et al., 2022). The i^{th}, $i \in [1, ..., N]$, client has n_i data samples, being benign for $i \in [1, ..., N-N']$ or being adversarial for $i \in [N-N'+1, ..., N]$. The synchronous federated learning is conducted in T rounds. In each round $t \in [1, ..., T]$, the server broadcasts a model parameterized by $\boldsymbol{w}_{t-1}$ to all the participating clients. We omit the subscript t while focusing on a single round. Then, the i^{th} client optimizes $\boldsymbol{w}_{t-1}$ on their local data samples indexed by j and reports the locally optimized $\boldsymbol{w}_{t,i}$ to the server. We define pseudo-gradient $\boldsymbol{g}_{t,i} = \boldsymbol{w}_{t-1} - \boldsymbol{w}_{t,i}$ being the difference between the locally optimized model and the broadcasted model from the previous round. For simplicity, we often use the term "gradient" to refer to the pseudo-gradient. Once all gradients are uploaded, the server aggregates them and produces a new model with parameters $\boldsymbol{w}_t$ using the following rule: $\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \sum_{i=1}^{N} \frac{n_i}{\sum_{i=1}^{N} n_i} \boldsymbol{g}_{t,i}$. The goal of federated learning is to minimize a weighted risk function over the N clients: $\mathcal{L}(\boldsymbol{w}) = \sum_{i=1}^{N} \frac{n_i}{\sum_{i=1}^{N} n_i} \mathcal{L}_i(\boldsymbol{w}) = \sum_{i=1}^{N} \frac{n_i}{\sum_{i=1}^{N} n_i} \mathbb{E}_{\mathcal{D}_i}[\ell(f(x; \boldsymbol{w}), y)]$, where $\ell : \mathbb{R} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a loss function. $\mathrm{sign}(\cdot)$ denotes an element-wise sign operator, $\odot$ denotes the Hadamard product operator, and $\mathrm{W}_1(\cdot, \cdot)$ denotes the Wasserstein-1 distance.

3.2 Threat Model

The adversary generates a backdoor data sample $\boldsymbol{x}'$ by embedding a trigger in a benign data sample $\boldsymbol{x}$ and correlating the trigger with a label y', which is different from the label y of the benign data sample. We use $\mathcal{D}'$ to denote the distribution of backdoor data samples. Then, the malicious clients connect to the federated learning system and insert backdoor data samples into the training set. Since federated learning aims to minimize the risk over all clients' datasets, the model can entangle the backdoor signals while trying to minimize the risk over all clients.

3.3 Assumptions

Bounded heterogeneity is a common assumption in federated learning literature (Wang et al., 2021). Let $\boldsymbol{w}_i^*$ be a minimum in client i's loss landscape. We assume the distance between the minimum of benign clients is bounded. Here, $\boldsymbol{w}_i^*$ is not necessarily a global minimum or a minimum of any global federated learning model but a parameter that a local model would converge to alone if client i has a sufficient amount of data.

Assumption 1. *(Bounded heterogeneity)* $\|\boldsymbol{w}_i^* - \boldsymbol{w}_j^*\| \leq \delta, \forall i \neq j, i \leq N - N', j \leq N - N'$.

Let $\mathcal{W}^*$ be a convex hull of $\{\boldsymbol{w}_i^* \mid i = 1, ..., N - N'\}$, we assume that malicious clients aim to converge to a model $\boldsymbol{w}'$ that is not in the convex hull $\mathcal{W}^*$ of benign clients' minima. However, we do not assume that all parameters in the convex hull $\mathcal{W}^*$ lead to zero backdoor success rate, especially since the convex hull may increase as the diameter δ of $\mathcal{W}^*$ increases. We empirically justify this separability assumption in Appendix D. Formally, this separability assumption is stated in the following.

Assumption 2. *(Separable minimum) Let $\mathcal{W}^*$ be a convex hull with diameter δ of benign minima $\{\boldsymbol{w}_i^* \mid i = 1, ..., N - N'\}$ and $\boldsymbol{w}'$ be a minimum of malicious client, we have $\boldsymbol{w}' \notin \mathcal{W}^*$.*

The estimated gradient often differs from the expected gradient in stochastic gradient descent. One of the most common models for estimated gradients is the additive noise model, which adds a noise term (e.g., Gaussian noise) to the expected gradient (Wu et al., 2019). For a given noise magnitude, the directional change of an estimated gradient may increase if its corresponding expected gradient shrinks. Formally, this noise assumption is stated in the following.

Assumption 3. *(Noisy gradient estimation) Let $\boldsymbol{g}_i$ be an estimated gradient vector from client i, we assume $\boldsymbol{g}_i = \mathbb{E}_{\mathcal{D}_i}[\boldsymbol{g}_i] + \epsilon_i$ where the noise term $\epsilon_i \sim \mathcal{N}(0, \boldsymbol{\sigma}_i)$ and $\mathcal{N}(0, \boldsymbol{\sigma}_i)$ is a Gaussian distribution with finite-norm covariance matrix $\boldsymbol{\sigma}_i$.*

4 Motivating Setting

Many recent works (Keskar et al., 2017; Li et al., 2017; Santurkar et al., 2018; yeh Chiang et al., 2023) suggest that the loss landscape of neural networks is "well-behaved" and has a flat region around the minimum (e.g., Figure 3 in (Li et al., 2017)). Following previous works, we discuss the difficulty of defending against federated backdoor attacks over flat loss landscapes and present concrete case studies where multiple prominent defenses can fail. Specifically, we consider a backdoor attack successful as long as the malicious clients can control the gradient direction and subsequently mislead model parameters toward the malicious minimum (Assumption 2, Figure 1).

To begin, we formally define a flat region around a minimum $\boldsymbol{w}_i^*$ as a path-connected set (i.e., there exists at least one path that connects two points in the set) where the gradient magnitude is small. Note that a flat region may not span over the entire space but exists within a subspace, and the flatness may depend on the weight norm $\|\boldsymbol{x}\|$ (Petzka et al., 2020a,b).

Definition 4. *(Flat region) Let $\mathcal{V}$ by a subspace of the parameter space $\mathbb{R}^{\mathrm{d}}$, we define a γ-flat region that spans over $\mathcal{V}$ around a minimum $\boldsymbol{w}^*$ as a path-connected set $\mathcal{B}^*$ that includes $\boldsymbol{w}^*$ where the magnitude of gradient within $\mathcal{V}$ is bounded by γ: $\|\mathbb{E}_{\mathcal{D}}[\boldsymbol{g}_{\mathcal{V}}]\| \leq \gamma$.*

4.1 Backdoor Attacks over a Flat Loss Landscape

The magnitude of benign gradients is small over flat loss landscapes, making it easier for the adversary to (1) mislead the aggregated gradient to the malicious minimum $\boldsymbol{w}'$ and (2) mimic benign clients to circumvent detection, e.g., by suffering a lower penalty for attack effectiveness. Figure 1 provides some intuitive examples. In addition, the adversary can intentionally exploit the flatness property.

Less dimensional perturbation requirements. Let $\boldsymbol{w}_t$ be the parameter of a global federated learning model in round t, where $\boldsymbol{w}_t$ is in regions of benign clients' loss landscapes with flatness at least γ. If a malicious client wants to guarantee that the parameter $\boldsymbol{w}_{t+1}$ is closer to the malicious minimum $\boldsymbol{w}'$ along dimension k, the magnitude of its gradient $\boldsymbol{g}'$ along dimension k needs to be at least $\frac{\sum_{i=1}^{N-1} n_i}{n'}\gamma$, which decreases as the flatness increases (i.e., smaller γ). Intuitively, the more flat the loss landscapes of benign clients are, the easier it is for the malicious client to "overwrite" the aggregation result (See the horizontal axis of Figure 1b for an illustration).

Further, backdoor adversaries do not necessarily need to "overwrite" the aggregation result along all dimensions. Instead, backdoor attacks may perturb only a few gradient elements to minimize the overall difference between malicious and benign gradients without losing effectiveness (See the red dashed gradient in Figure 1c for an illustration).

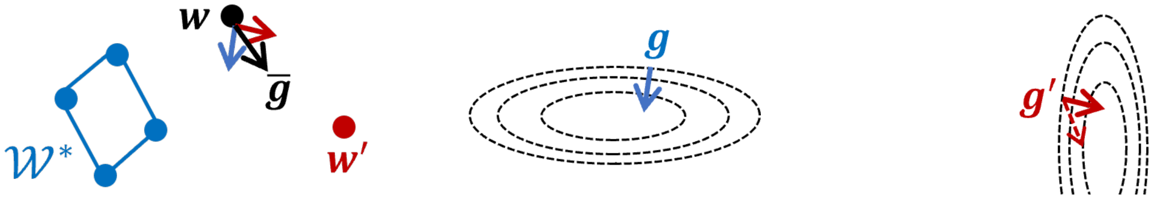
Less penalty for mimicking benign clients. Since the flat loss landscape is a general property of well-designed neural networks, the loss landscape of a malicious client in the unperturbed subspace can also be flat. Then, the malicious clients may partially mimic the behavior of benign clients to circumvent detection without significantly decreasing the attack success rate. Specifically, if the loss landscape of a malicious client within the unperturbed subspace is γ'-flat and the gradient magnitude of a benign client is upper bounded by γ, then it is easy to see that mimicking the benign client only decreases the effectiveness of backdoor attacks measured by the loss on backdoor samples by up to $\gamma\gamma'$, which decreases as the flatness increases (i.e., smaller γ'), via the Lagrange mean value theorem,

So far, we have seen how flat landscapes can reduce the gradient perturbation requirement of backdoor attacks and help attacks remain effective while malicious clients mimic benign clients. Even worse, an adversary may intentionally work to flatten the loss landscape, e.g., through edge-case backdoor attacks (Wang et al., 2020) to further increase the attack's effectiveness and circumvent defenses.

Edge-case attack flattens the loss landscape. The main idea of the edge-case backdoor attack is minimizing the marginal probability of backdoor samples in the benign data distribution (Wang et al., 2020). If a backdoor sample appears on both benign and malicious clients, it gets different label assignments on different types of clients. Thus, for such backdoor samples, the loss on benign clients would increase because at least one data sample in their datasets would be mispredicted. The more the loss increases, the less flat the loss landscape is, and vice versa. The edge-case backdoor attack intentionally prevents backdoor samples from appearing on benign clients and avoids the prediction error being observed to flatten loss landscapes of benign clients, as is empirically verified in Appendix D.

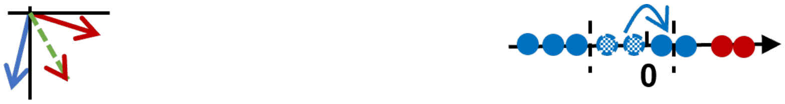
4.2 Limitation of Existing Defenses over a Flat Loss Landscape

Under the flat loss landscape setting, we discuss how existing defenses can fail to recover the correct gradient direction, including vector-wise, dimension-wise, and trigger inversion defenses. The following case study shows the failure mode of FLTrust, a vector-wise defense for federated learning systems.



(a) Overview (b) Landscape on a benign client (c) Landscape on a malicious client.

Figure 1: (a) Overview of our motivating setting where benign minima (with convex hull $\mathcal{W}^*$) and the malicious minimum $\boldsymbol{w}'$ are separable. $\boldsymbol{w}$ is the parameter in the previous round, and the dashed circles in (b) and (c) are loss contours. (b) The flat landscape of a benign client (blue) along the horizontal axis reduces the horizontal gradient magnitude, allowing a malicious client (red) to easily mislead the aggregated gradient $\bar{\boldsymbol{g}}$ toward the malicious minimum $\boldsymbol{w}'$. (c) The malicious client can mimic the benign client (red dashed arrow) along the vertical dimension with less penalty due to its flat loss landscape along the vertical axis.



(a) FLTrust failure (b) Median failure

Figure 2: Failure mode examples of existing approaches. (a) FLTrust can fail to recover the benign direction (blue) along the horizontal axis, which may subsequently converge model parameters to a malicious minimum (Figure 1). This is because a malicious client (red) can mimic the benign client along the vertical axis to avoid being detected as an anomaly, and misleading the aggregation result along the horizontal axis is easier due to the small benign gradient magnitude caused by flat loss landscape. (b) Median can fail to recover the benign direction (blue) even if the estimation error is small when a few benign gradients flip (blue arrow) their sign due to gradient estimation noise. Gradients with smaller magnitudes may be easier to flip for a given noise level.

FLTrust. FLTrust (Cao et al., 2020) uses a trusted root dataset (Xie et al., 2019) to generate a reference gradient vector $\boldsymbol{g}^*$. Then, FLTrust weights each reported gradient vector using its cosine similarity to the reference before normalization and aggregation. To simplify the discussion, we consider an example with one benign client whose gradient aligns with the reference gradient (i.e., cosine similarity is 1), one malicious client, a uniform sample number across clients, and two-dimensional gradient vectors. Suppose the benign client has a flat loss landscape along the first dimension (i.e., $|\boldsymbol{g}_1| \gg \gamma_0 \geq |\boldsymbol{g}_0|$). The following proposition suggests that the aggregation result always points to the direction specified by the malicious client along the first dimension (Figures 2a) due to the reduced perturbation requirement in a flat loss landscape setting (i.e., a small γ_0) as is discussed in Section 4.1.

Proposition 5. *Let $\boldsymbol{g}$ be a 2-dimensional (2-d) benign gradient, $\boldsymbol{g}'$ be a 2-d malicious gradient, and $\boldsymbol{g}^*$ be a 2-d reference gradient estimated over the trust root dataset, suppose $\boldsymbol{g}_0\boldsymbol{g}'_0 < 0$ and $\boldsymbol{g}_1\boldsymbol{g}'_1 > 0$, under the aggregation rule of FLTrust which enforces $\|\boldsymbol{g}\| = \|\boldsymbol{g}'\| = \|\boldsymbol{g}^*\|$, if $\|\boldsymbol{g}_0\| \leq \gamma_0 \leq \|\boldsymbol{g}\| \cdot \cos(0.4\pi)$, there exists a malicious gradient $\boldsymbol{g}'$ such that $\|\boldsymbol{g}'_0\|\frac{\boldsymbol{g}'\cdot\boldsymbol{g}^*}{\|\boldsymbol{g}'\|\|\boldsymbol{g}^*\|} > \|\boldsymbol{g}_0\|$ and $\boldsymbol{g}'_0(\boldsymbol{g}'_0\frac{\boldsymbol{g}'\cdot\boldsymbol{g}^*}{\|\boldsymbol{g}'\|\|\boldsymbol{g}^*\|} + \boldsymbol{g}_0) > 0$*

Meanwhile, directly applying dimension-wise defenses may not be effective either. The following example focuses on the median estimator (Xie et al., 2020a) from robust statistics.

Median. Robust statistics provides valid estimation results with small errors in the presence of outliers (Lugosi and Mendelson, 2021) and gradient estimator noise. However, they are not aware of the ”location” of the ground-truth value. Therefore, for a median $\tilde{\boldsymbol{g}}_k$ and an estimation error $\|\tilde{\boldsymbol{g}}_k - \mathbb{E}[\boldsymbol{g}_k]\|$, if the median $\tilde{\boldsymbol{g}}_k$ has a small magnitude over a flat landscape such that $\|\tilde{\boldsymbol{g}}_k\| < \|\tilde{\boldsymbol{g}}_k - \mathbb{E}[\boldsymbol{g}_k]\|$, the adversary may mislead the aggregation result pointing to the malicious minimum. Figure 2b shows such a failure mode. Later, we will theoretically show that our invariant aggregator can reduce the probability of such a failure mode.

In addition to the vector-wise and dimension-wise defenses, trigger inversion (Wang et al., 2019; Zhang et al., 2023) is a more directed defense against backdoor attacks that aim to reverse engineer backdoor samples and assign them correct labels. However, we find that trigger inversion can be less effective against semantic backdoor attacks (Wang et al., 2020), whose trigger does not have a fixed shape or location and is, therefore, difficult to reverse engineer precisely. The following discussion further details the limitation of trigger inversion approaches from a statistical distance perspective, showing why imprecise trigger inversion has reduced effectiveness.

Trigger inversion. Trigger inversion approaches aim to generate a data distribution $\mathcal{D}$ that is close to the backdoor data distribution $\mathcal{D}'$ but has correct label assignments. Suppose a trigger inversion defense recovers the backdoor data distribution. In that case, any backdoor attack that misleads model predictions will be observed via evaluating models over $\mathcal{D}$, resulting in a non-flat loss landscape. However, the useful gradients may get redirected and fail to erase the backdoor as the inversed distribution $\mathcal{D}$ shifts away from the backdoor data distribution $\mathcal{D}'$, hurting the non-flatness guarantee from trigger inversion defense. The following formalizes this observation.

Proposition 6. *For a model with λ-Lipschitz gradient, the difference $\|\mathbb{E}_{\mathcal{D}}[\boldsymbol{g}] - \mathbb{E}_{\mathcal{D}'}[\boldsymbol{g}']\|$ between the gradient $\boldsymbol{g}$ over $\mathcal{D}$ and $\boldsymbol{g}'$ over $\mathcal{D}'$ can go up to $\lambda \mathrm{W}_1(\mathcal{D}, \mathcal{D}')$, meaning that $\gamma \geq \max\Big(0, \gamma' - \lambda \mathrm{W}_1(\mathcal{D}, \mathcal{D}')\Big)$.*

5 Method

We propose an invariant aggregator to defend against backdoor attacks by avoiding the updates where the aggregated gradient benefits a few and possibly malicious clients regardless of their gradient magnitudes. Specifically, our approach considers the gradient sign as a magnitude-agnostic indicator of benefit and avoids taking any optimization steps that can not benefit sufficiently many clients. The magnitude-agnostic property maintains the robustness of our approach over flat loss landscapes where the gradient magnitude shrinks. Thus, our approach is distinct from existing approaches that consider malicious clients as anomalies or outliers (Cao et al., 2020; Pillutla et al., 2022). In what follows, we introduce the technical details of our invariant aggregator and discuss how it provably mitigates backdoor attacks by decreasing the attack success rate.

5.1 Invariant Aggregator

AND-Mask. The AND-mask (Parascandolo et al., 2021) computes a dimension-wise mask by inspecting the sign consistency of each dimension across samples. Here, we apply it to the sign consistency across clients. For dimension k, the sign consistency is: $|\frac{1}{N}\sum_{i=1}^{N}\mathrm{sign}(\boldsymbol{g}_{i,k})|$. If the sign consistency is below a given threshold τ, the mask element m_k is set to 0; otherwise, m_k is set to 1. The mask along dimension k is defined as:

Definition 7. *(AND-Mask) For the k^{th} dimension in the gradient vector, the corresponding mask m_k is defined as: $m_k := \frac{1}{N} \cdot \mathbf{1}\Big[|\sum_{i=1}^{N}\mathrm{sign}(\boldsymbol{g}_{i,k})| > \tau\Big]$.*

Our defense then multiplies the mask m with the aggregated gradient $\tilde{\boldsymbol{g}}$ element-wise, setting the inconsistent dimension to zero to avoid benefiting few malicious clients. Since the AND-mask focuses on the gradient direction, its effectiveness does not diminish due to the flatness of the landscape, which only affects the gradient magnitudes. For a consistent dimension, we call the majority gradient direction "invariant direction". However, if we naively average the gradient elements in consistent dimensions, malicious clients could use outliers to mislead the averaging result away from the invariant direction.

Trimmed-mean. To complement the AND-mask and ensure that the aggregation result does follow invariant directions in consistent dimensions, our defense broadcasts the trimmed-mean estimator to each gradient dimension. The trimmed-mean estimator alleviates the outlier issue by removing the subset of the largest and smallest elements before computing the mean. The largest and smallest elements appear on the two tails of a sorted sequence. Next, we define order statistics and the trimmed mean estimator.

Definition 8. *(Order Statistics) (Xie et al., 2020a) By sorting the scalar sequence $\{x_i : i \in \{1, ..., N\}, x_i \in \mathbb{R}\}$, we get $x_{1:N} \leq x_{2:N} \leq ... \leq x_{N:N}$, where $x_{i:N}$ is the i^{th} smallest element in $\{x_i : i \in \{1, ..., N\}\}$.*

Then, the trimmed-mean estimator removes $\alpha \times N$ elements from each tail of the sorted sequence.

Definition 9. *(Trimmed Mean Estimator) (Xie et al., 2020a) For $\alpha \in [0, 1]$, the α-trimmed mean of the set of scalars $x_{i:N} \in \{1, ..., N\}$ is defined as: $\mathrm{TrMean}(\{x_1, ..., x_N\}; \alpha) = \frac{1}{N - 2\cdot\lceil\alpha\cdot N\rceil}\sum_{i=\lceil\alpha\cdot N\rceil+1}^{N-\lceil\alpha\cdot N\rceil} x_{i:N}$, where $\lceil . \rceil$ denotes the ceiling function.*

Algorithm 1 outlines the steps of our server-side defense that implements invariant aggregation of gradients from the clients. The solution is composed of the AND-mask and trimmed-mean estimator. Our defense applies the two components separately based on the sign consistency of each dimension with a threshold τ.

5.2 Provable Mitigation

We demonstrate the provable backdoor mitigation of our invariant aggregator by showing that it maintains the progress of converging a model parameter $\boldsymbol{w}$ toward benign minima $\mathcal{W}^*$ and reduces the probability of moving toward the malicious minimum $\boldsymbol{w}'$(Assumption

Algorithm 1 Server-side Defense

```
Input:
    A set of reported gradients, $\{\boldsymbol{g}_i \mid i \in \{1, ..., N\}\}$;
    Hyper-parameters $\tau$, $\alpha$;
Aggregator:
Compute $m := \frac{1}{N} \cdot \mathbf{1}\left[|\sum_{i=1}^{N} \mathrm{sign}(\boldsymbol{g}_i)| > \tau\right]$;
Compute $\tilde{\boldsymbol{g}} := \mathrm{TrMean}(\{\boldsymbol{g}_1, ..., \boldsymbol{g}_N\}; \alpha)$;
return $\bar{\boldsymbol{g}} := m \odot \tilde{\boldsymbol{g}}$;
```

2). Our results also suggest that (1) the invariant aggregator is more effective than baselines and (2) the flat loss landscape is less likely to break our invariant aggregation, differing from existing approaches (Section 2). We start with a single-dimension analysis where the benign minima $\mathcal{W}^*$ and the malicious minimum $\boldsymbol{w}'$ are on different sides of the current parameter (e.g., horizontal dimension in Figure 1a).

Theorem 10. *(Single-dimension) Under Assumption 3, for a parameter $\boldsymbol{w} \notin \mathcal{W}^*$ where $(\boldsymbol{w}_k - \boldsymbol{w}^*_{i,k})(\boldsymbol{w}_k - \boldsymbol{w}'_k) \leq 0, \forall i \in \{1, ..., N - N'\}$ along the k^{th} dimension, let the sign-flipping probability $p_k = \max_{i \in \{1,...,N-N'\}} \mathbb{P}[\mathbb{E}[\boldsymbol{g}_{i,k}]\boldsymbol{g}_{i,k} < 0]$ and $\bar{\boldsymbol{g}}$ be the aggregated gradient, using the invariant aggregator with $\frac{N'}{N} \leq \alpha < \frac{1}{2}$ and $\tau = 1 - 2\alpha$, with probability at least $p_- = \sum_{i=N-\alpha N}^{N-N'} (1-p)^i$, we have the aggregated $\bar{\boldsymbol{g}}_k$ points to the benign $\mathcal{W}^*$ and with probability at most $p_+ = \sum_{i=\frac{1+\tau}{2}N-N'+1}^{N-N'} p^i$ we have the aggregated $\bar{\boldsymbol{g}}_k$ points to the malicious $\boldsymbol{w}'$. In contrast, we have $p_- = \sum_{i=N-\alpha N}^{N-N'} (1-p)^i$ and $p_+ = \sum_{i=\alpha N-N'+1}^{N-N'} p^i$ if we use the trimmed-mean estimator alone and have $p_- = 0$ and $p_+ = 1$ if using the arithmetic mean.*

The high-level idea of Theorem 10 is to preserve the stable progress toward benign minima (i.e., p_-) and cut off potential progress toward the malicious minimum (i.e., p_+). For example, in the case of Figure 2b, our invariant aggregator will not take any steps along malicious gradients due to reduced sign consistency if $\tau \geq \frac{1}{7}$. In general, Theorem 10 suggests our approach is more effective than the trimmed-mean estimator in reducing the attack success rate because $\frac{1+\tau}{2} > \alpha$. Then, we analyze the convergence guarantee of our aggregator to a neighborhood of benign minima $\mathcal{W}^*$ because our single-dimension result holds when the current parameter $\boldsymbol{w} \notin \mathcal{W}^*$ and malicious clients may pull $\boldsymbol{w}$ out of $\mathcal{W}^*$ by up to one step.

Theorem 11. *(Convergence) Under Assumptions 1 - 3 and Theorem 10, let $\boldsymbol{w}$ be a initial parameter, suppose $|\boldsymbol{w}_k - \boldsymbol{w}^*_{i,k}| \leq c$ and $\eta_- \leq |\bar{\boldsymbol{g}}_{k,t}| \leq \eta^-$, $\forall i \in \{1, ..., N - N'\}, k \in \{1, ..., \mathrm{d}\}, t \in \{1, ...\}$, $|\bar{\boldsymbol{g}}_{k,t}| > 0$, if the number of round $T \geq \frac{c}{\eta_-}$, with a probability at least $\left[\sum_{i=\frac{c}{\eta_-}}^{T} \mathcal{F}(T, i, p_-) \cdot \sum_{j=\frac{i \cdot n_- - c}{n^-}}^{T-i} \mathcal{F}(T-i, j, \frac{1-p_- - p_+}{1-p_-})\right]^{\mathrm{d}}$ where $\mathcal{F}(T, i, p_-)$ denotes a binomial density function with T trails, i success, and probability p_-, we have $\boldsymbol{w}_T \in \{\boldsymbol{w} \mid \exists \boldsymbol{w}^* \in \mathcal{W}^*, \|\boldsymbol{w} - \boldsymbol{w}^*\| \leq \sqrt{\mathrm{d}}\eta^-\}$.*

The proof of Theorem 11 is straightforward: we compute the minimum number of steps that a model needs to converge the benign minima and count how many wrong steps toward the malicious minimum we can tolerate. Since the binomial cumulative density function (i.e., the sum of $\mathcal{F}$) is monotonic, we can see that the convergence probability increases with a larger p_- and a smaller p_+. In Theorem 10, we have already shown that our approach reduces p_+ without hurting p_-.

Connection to Flatness The loss landscape flatness still plays a role because the sign-flipping probability p in Theorem 10 can increase with more flatness (i.e., a smaller γ) and larger noise magnitudes $\boldsymbol{\sigma}$ (Section 3.3) via Chebyschev's inequality: $\mathbb{P}(|\mathbb{E}[\boldsymbol{g}_{i,k}] - \boldsymbol{g}_{i,k}| \geq \gamma_k) \leq \frac{\sigma^2_{i,k}}{\gamma^2_k}$. An increased sign-flipping probability can subsequently increase the failure probability p_+ and reduce the effectiveness of our approach. However, this does not imply that our approach suffers from flat landscapes similar to existing approaches (Section 4) because the sign-flipping probability p depends on the interaction between the flatness and the noise. In other words, obtaining a more accurate gradient estimation (Johnson and Zhang, 2013) can improve our approach over flat loss landscapes.

Connection to Backdoor Success. We further connect our convergence guarantee in Theorem 11 to the success of backdoor attacks, which is defined via classification results. With a distribution $\mathcal{D}'$ of backdoor samples, the attack success loss $\mathcal{L}_{\mathcal{D}'}(\mathbf{w}^*)$ over a benign minimum $\mathbf{w}^* \in \mathcal{W}^*$ is expected to be high because an unperturbed benign model does not entangle backdoor triggers. With our theoretical guarantees, it is easy to lower bound the attack success loss $\mathcal{L}_{\mathcal{D}'}(\mathbf{w}) \geq \max\left(0, \mathcal{L}_{\mathcal{D}'}(\mathbf{w}^*) - \gamma(\delta + \sqrt{d}\eta^-)\right)$ via Lagrange mean value theorem. Here, δ is the diameter of $\mathcal{W}^*$, d is the dimension of the parameter space, and η^- is the maximum step size defined in Theorem 11.

Comparison. Our approach provides a high probability bound of mitigating backdoor attacks in Theorem 11. In contrast, FLTrust suffers from the failure mode in Proposition 5 and is insufficient for limiting the progress towards the malicious minimum $\boldsymbol{w}'$. In addition, the trimmed-mean defense, including the median, does not meet the same probability as our approach.

Table 1: Accuracy of Aggregators under Continuous Edge-case Backdoor Attack. Our approach reduces the model accuracy on backdoor samples by 61.7% on average, mitigating the backdoor attack, and achieves a comparable utility on benign samples as the standard FedAvg aggregator.

Method	CIFAR-10		Twitter		Phishing	
	Acc	ASR	Acc	ASR	Acc	ASR
FedAvg	.679 ± .001	.717 ± .001	.722 ± .001	.440 ± .001	.999 ± .001	.999 ± .001
Krum (Blanchard et al., 2017)	.140 ± .001	.275 ± .012	.579 ± .001	.766 ± .002	.999 ± .001	.999 ± .001
Multi-Krum	.541± .002	.923 ± .021	.727 ± .001	.656 ± .008	.999 ± .001	.999 ± .001
Multi-Krum$_C$	.681 ± .002	.821 ± .001	.594 ± .002	.701 ± .001	.999 ± .001	.333 ± .333
Trimmed-Mean (Xie et al., 2020a)	.687 ± .001	.512 ± .002	.728 ± .001	.640 ± .016	.999 ± .001	.999 ± .001
Krum Trimmed-Mean	.682 ± .001	.607 ± .002	.727 ± .001	.641 ± .001	.999 ± .001	.999 ± .001
Sign-SGD (Bernstein et al., 2019)	.301 ± .005	**.000 ± .001**	.610 ± .003	.751 ± .076	.999 ± .000	.667 ± .333
Weak-DP (Sun et al., 2019)	.454 ± .003	.828 ± .003	.667 ± .001	.374 ± .002	.999 ± .001	.999 ± .001
RLR (Ozdayi et al., 2021)	.659 ± .001	.450 ± .001	.639 ± .002	.447 ± .001	.999 ± .001	.999 ± .001
RFA (Pillutla et al., 2022)	.685 ± .001	.853 ± .002	.718 ± .001	.704 ± .002	.999 ± .001	.999 ± .001
SparseFed (Panda et al., 2022)	.662 ± .001	.984 ± .001	.667 ± .001	.608 ± .002	.999 ± .001	.999 ± .001
FLTrust (Cao et al., 2020)	.671 ± .001	.574 ± .002	.691 ± .001	.473 ± .002	.999 ± .001	.333 ± .001
FLIP (Zhang et al., 2023)	.667 ± .002	.250 ± .001	N\A	N\A	.999 ± .001	**.000 ± .001**
No Attack	.718 ± .001	.000 ± .001	.731 ± .001	.095 ± .001	.999 ± .001	.000 ± .001
Ours	.677 ± .001	**.001 ± .001**	.687 ± .001	**.296 ± .003**	.999 ± .001	**.000 ± .001**

Note: The numbers are average accuracy over three runs. Variance is rounded up.

6 Experiments

We extensively evaluate our defense on three realistic tasks with diverse data modalities and against multiple state-of-the-art backdoor attacks (Wang et al., 2020; Xie et al., 2020b; Panda et al., 2022) with pixel, semantic visual, text, and value backdoor triggers.

Additional Results. Due to the limited space, (1) the loss landscape visualization, (2) an ablation study, (3) an evaluation of the hyper-parameter sensitivity, (4) empirical verifications of the separability assumption (Assumption 2), and (5) empirical evaluations under multiple state-of-the-art backdoor attack strategies (Wang et al., 2020; Xie et al., 2020b; Panda et al., 2022) are in Appendix D.

6.1 Experimental Setup

We briefly summarize our setup and report more details in Appendix C.

Metrics. We employ two metrics: the main task accuracy (Acc) estimated on the benign samples and the backdoor attack success rate (ASR) measured by the accuracy over backdoor samples.

Datasets. The visual data of the object detection task and text data of the sentiment analysis task are from CIFAR-10 (Krizhevsky, 2009; McMahan et al., 2017) and Twitter (Caldas et al., 2018), respectively. Each phishing email data sample has 45 standardized numerical features and binary labels (Appendix C.4)

Federated Learning Setup. We consider horizontal federated learning (Kairouz et al., 2021) where the clients share the same feature and label spaces but non-i.i.d. data distributions. The number of clients is 100 for the three tasks. The server samples 20 clients at each round on the CIFAR-10 and phishing email experiments. Due to hardware memory limitations, we reduce the sampled client number to 15 with Twitter.

Backdoor Attack Setup. We adopt edge-case (Wang et al., 2020), distributed (Xie et al., 2020b), and collude (Panda et al., 2022) backdoor strategies. The pixel, semantic visual, and text backdoor trigger follow the previous work (Wang et al., 2020; Xie et al., 2020b). For the tabular data, we select the 38th feature (reputation), whose value is 0 on most of the data samples. Then, we let the adversary manipulate the 38th feature to 0.2, which has a low probability density on phishing emails, and flip the label to non-phishing.

Adversary Setup. We employ strong adversaries that *continuously* participate in the training (Shejwalkar et al., 2022; Zhang et al., 2023) and can control 20% clients on the CIFAR-10 and phishing email experiments and 10% clients on the Twitter experiment.

6.2 Result and Comparison to Prior Works

Table 1 summarizes the performance of each defense on three tasks against edge-case backdoor attacks that can intentionally flatten loss landscapes (Section 4). Our approach decreases the ASR by 61.6% on average, outperforming all competitors. The edge-case backdoor attack on the text sentiment analysis task (Twitter) is more difficult to defend, and our approach reduces the ASR by 41.7%. We hypothesize that the text sentiment analysis task has few invariances among clients. For example, the shape features (Sun et al., 2021) in object classification tasks can be invariant across objects. In contrast, the sentiment largely depends on the entire sentence instead of a few symbols or features. We further discuss the limitations of prior defenses and evaluate our defense against various attacks in Appendix D.

7 Conclusion and Future Work

This paper shows a defense against backdoor attacks by focusing on the invariant directions in the model optimization trajectory. Enforcing the model to follow the invariant direction requires AND-mask to compute the sign-consistency of each gradient dimension, which estimates how invariant a dimension-wise direction can be, and use the trimmed-mean estimator to guarantee the model follows the invariant direction within each dimension. Both theoretical and empirical results demonstrate the combination of AND-mask and the trimmed-mean estimator is necessary and effective.

In addition, our work reveals a potential downside of flat loss landscapes, which are considered beneficial in prior work (Keskar et al., 2017; Foret et al., 2021). Specifically, few existing works consider the invariance of the flatness property, i.e., whether every data sample sees a flat landscape. Therefore, further combining distributional robust optimization (Duchi and Namkoong, 2021; Sagawa* et al., 2020) and sharpness-aware minimization (Foret et al., 2021) to discover invariant flat minima can be interesting for future work.

Acknowledgements

XW is partially supported by Microsoft Research and a research grant from C3.ai digital transformation institute. This work of SK is partially supported by NSF III 2046795, IIS 1909577, CCF 1934986, NIH 1R01MH116226-01A, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc.

References

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and X. Zhang. Trojaning attack on neural networks. In *NDSS*, 2018.

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2938–2948. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/bagdasaryan20a.html.

Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel 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)*, pages 1354–1371, 2022. doi: 10.1109/SP46214.2022.9833647.

Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy-yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16070–16084. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b8ffa41d4e492f0fad2f13e29e1762eb-Paper.pdf.

Hao Li, Zheng Xu, Gavin Taylor, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2017.

Ruoyu Sun, Dawei Li, Shiyu Liang, Tian Ding, and Rayadurgam Srikant. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37:95–108, 2020.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, 2019.

Kaiyuan Zhang, Guanhong Tao, Qiuling Xu, Siyuan Cheng, Shengwei An, Yingqi Liu, Shiwei Feng, Guangyu Shen, Pin-Yu Chen, Shiqing Ma, and Xiangyu Zhang. FLIP: A provable defense framework for backdoor mitigation in federated learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Xo2E217_M4n.

Giambattista Parascandolo, Alexander Neitz, ANTONIO ORVIETO, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hb1sDDSLbV.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Slsgd: Secure and efficient distributed on-device machine learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 213–228, Cham, 2020a. Springer International Publishing. ISBN 978-3-030-46147-8.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, 49(1):393 – 410, 2021. doi: 10.1214/20-AOS1961. URL https://doi.org/10.1214/20-AOS1961.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konecný, H. B. McMahan, Virginia Smith, and Ameet S. Talwalkar. Leaf: A benchmark for federated settings. *ArXiv*, abs/1812.01097, 2018.

Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y. Zhao. Backdoor attacks against deep learning systems in the physical world. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6202–6211, 2021.

Yuezun Li, Y. Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16443–16452, 2021a.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/280cf18baf4311c92aa5a042336587d3-Paper.pdf.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *Advances in Neural Information Processing Systems*, 2021b.

Ziteng Sun, Peter Kairouz, Ananda Theertha Suresh, and H. B. McMahan. Can you really backdoor federated learning? *ArXiv*, abs/1911.07963, 2019.

Thien Duc Nguyen, Phillip Rieger, Huili Chen, Hossein Yalame, Helen Mollering, Hossein Fereidooni, Samuel Marchal, Markus Miettinen, Azalia Mirhoseini, Shaza Zeitouni, Farinaz Koushanfar, Ahmad-Reza Sadeghi, and T. Schneider. Flame: Taming backdoors in federated learning. 2021.

Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4cEapqXfP30.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NIPS*, 2017.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Anima Anandkumar. signsgd with majority vote is communication efficient and fault tolerant. In *ICLR*, 2019.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Ping yeh Chiang, Renkun Ni, David Yu Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, and Tom Goldstein. Loss landscapes are all you need: Neural network generalization can be explained without the implicit bias of gradient descent. In *The Eleventh International Conference on Learning Representations*,

2023. URL https://openreview.net/forum?id=QC10RmRbZy9.

Henning Petzka, Martin Trimmel, and Cristian Sminchisescu. Notes on the symmetries of 2-layer relu-networks. In *NLDL*, 2020a.

Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Neural Information Processing Systems*, 2020b.

Xiaoyu Cao, Minghong Fang, Jia Liu, and Neil Zhenqiang Gong. Fltrust: Byzantine-robust federated learning via trust bootstrapping. *ArXiv*, abs/2012.13995, 2020.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Zeno: Distributed stochastic gradient descent with suspicion-based fault-tolerance. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6893–6901. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/xie19b.html.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/ac1dd209cbcc5e5d1c6e28598e8cbbe8-Paper.pdf.

Mustafa Safa Ozdayi, Murat Kantarcioglu, and Yulia R Gel. Defending against backdoors in federated learning with robust learning rate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9268–9276, 2021.

Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, and Prateek Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7587–7624. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/panda22a.html.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=rkgyS0VFvr.

Peter Kairouz, H. B. McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary B. Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim Y. El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Oluwasanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, R. Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Xiaodong Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *ArXiv*, abs/1912.04977, 2021.

Mingjie Sun, Zichao Li, Chaowei Xiao, Haonan Qiu, Bhavya Kailkhura, Mingyan Liu, and Bo Li. Can shape structure features improve model robustness under diverse adversarial settings? *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7506–7515, 2021.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3): 1378 – 1406, 2021. doi: 10.1214/20-AOS2004. URL https://doi.org/10.1214/20-AOS2004.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://aclanthology.org/D14-1162`.

Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decis. Support Syst.*, 43:618–644, 2007.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BylIciRcYQ`.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2698–2707. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kleinberg18a.html`.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=jXKKDEi5vJt`.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Table of Notations

Table 2: Table of Notations

Symbol	Description
$\boldsymbol{x}, y$	A pair of data sample and label
$\text{sign}(\cdot)$	An element-wise sign operator
$\boldsymbol{w}$	Parameters of the global federated learning model
$\boldsymbol{w}_i$	Parameters of the i^{th} local federated learning model
$\boldsymbol{g}$	Client update (i.e., pseudo-gradient or gradient)
N	The number of clients
N'	The number of malicious clients
T	The number of training rounds
$\text{W}_1(\cdot,\cdot)$	Wasserstain-1 distance between two distributions

B Proofs

Proposition 5. *Let $\boldsymbol{g}$ be a 2-dimensional (2-d) benign gradient, $\boldsymbol{g}'$ be a 2-d malicious gradient, and $\boldsymbol{g}^*$ be a 2-d reference gradient estimated over the trust root dataset, suppose $\boldsymbol{g}_0\boldsymbol{g}_0' < 0$ and $\boldsymbol{g}_1\boldsymbol{g}_1' > 0$, under the aggregation rule of FLTrust which enforces $\|\boldsymbol{g}\| = \|\boldsymbol{g}'\| = \|\boldsymbol{g}^*\|$, if $|\boldsymbol{g}_0| \leq \|\boldsymbol{g}\| \cdot \cos(0.4\pi)$, there exists a malicious gradient $\boldsymbol{g}'$ such that $|\boldsymbol{g}_0'|\frac{\boldsymbol{g}'\cdot\boldsymbol{g}^*}{\|\boldsymbol{g}'\|\|\boldsymbol{g}^*\|} > |\boldsymbol{g}_0|$ and $\boldsymbol{g}_0'(\boldsymbol{g}_0'\frac{\boldsymbol{g}'\cdot\boldsymbol{g}^*}{\|\boldsymbol{g}'\|\|\boldsymbol{g}^*\|} + \boldsymbol{g}_0) > 0$*

Proof. Let $\theta \in [0, \frac{\pi}{2}]$ be the angle between $\boldsymbol{g}$ and $[\boldsymbol{g}_0, 0]$ and $\theta' \in [0, \frac{\pi}{2}]$ be the angle between $\boldsymbol{g}'$ and $[\boldsymbol{g}_0', 0]$, with $\|\boldsymbol{g}\| = \|\boldsymbol{g}'\| = \|\boldsymbol{g}^*\|$, we have $|\boldsymbol{g}_0'|\frac{\boldsymbol{g}'\cdot\boldsymbol{g}^*}{\|\boldsymbol{g}'\|\|\boldsymbol{g}^*\|} = \|\boldsymbol{g}^*\|\text{cos}(\theta')\text{cos}(\pi - \theta - \theta')$ and $|\boldsymbol{g}_0| = \|\boldsymbol{g}^*\|\text{cos}\theta$. Then, $|\boldsymbol{g}_0'|\frac{\boldsymbol{g}'\cdot\boldsymbol{g}^*}{\|\boldsymbol{g}'\|\|\boldsymbol{g}^*\|} > |\boldsymbol{g}_0|$ is equivalent to

$$\text{cos}(\theta')\text{cos}(\pi - \theta - \theta') - \text{cos}\theta > 0. \tag{1}$$

We are interested in figuring out, for a given benign gradient with a fixed θ, whether there exists a malicious gradient with θ' such that Equation 1 holds. Since an analytical solution for Equation 1 may not be tractable, we resort to the numerical simulation. Figure 3 shows the plot of $\max_{\theta' \in [0, \frac{\pi}{2}]} \text{cos}(\theta')\text{cos}(\pi - \theta - \theta') - \text{cos}\theta$ for each given θ in $[0, \frac{\pi}{2}]$. It is easy to see that as long as $\theta \geq 0.4\pi$, there exists a $\theta' \in [0, \frac{\pi}{2}]$ such that Equation 1 holds. □

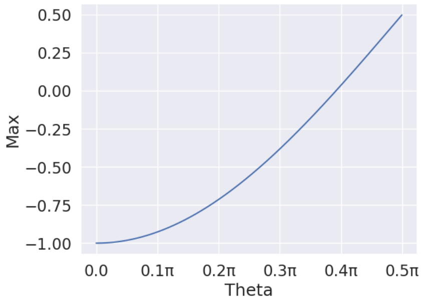


Figure 3: Numerical simulation for Equation 1.

Proposition 6. *For a model with λ-Lipschitz gradient, the difference $\|\mathbb{E}_{\mathcal{D}}[\boldsymbol{g}] - \mathbb{E}_{\mathcal{D}'}[\boldsymbol{g}']\|$ between the gradient $\boldsymbol{g}$ over $\mathcal{D}$ and $\boldsymbol{g}'$ over $\mathcal{D}'$ can go up to $\lambda\text{W}_1(\mathcal{D}, \mathcal{D}')$.*

Proof. Let $\boldsymbol{g}(\boldsymbol{x})$ denotes a gradient at $\boldsymbol{x}$ and $p(\boldsymbol{x}, \boldsymbol{x}')$ be the joint probability of $\boldsymbol{x}$ and $\boldsymbol{x}'$, we have

$$\begin{aligned} \|\mathbb{E}_{\mathcal{D}}[\boldsymbol{g}] - \mathbb{E}_{\mathcal{D}'}[\boldsymbol{g}']\| &= \|\mathbb{E}_{\mathcal{D}}[\boldsymbol{g}(\boldsymbol{x})] - \mathbb{E}_{\mathcal{D}'}[\boldsymbol{g}'(\boldsymbol{x}')]\| \\ &\leq \lambda \int \|\boldsymbol{x} - \boldsymbol{x}'\| p(\boldsymbol{x}, \boldsymbol{x}') \\ &\leq \lambda \mathrm{W}_1(\mathcal{D}, \mathcal{D}') \end{aligned} \tag{2}$$

□

Theorem 10. *(Single-dimension) Under Assumption 3, for a parameter $\boldsymbol{w} \notin \mathcal{W}^*$ where $(\boldsymbol{w}_k - \boldsymbol{w}^*_{i,k})(\boldsymbol{w}_k - \boldsymbol{w}'_k) \leq 0, \forall i \in \{1, ..., N - N'\}$ along the k^{th} dimension, let the sign-flipping probability $p = \max_{i \in \{1,...,N-N'\}} \mathbb{P}[\mathbb{E}[\boldsymbol{g}_{i,k}]\boldsymbol{g}_{i,k} < 0]$ and $\bar{\boldsymbol{g}}$ be the aggregated gradient, using the invariant aggregator with $\frac{N'}{N} \leq \alpha < \frac{1}{2}$ and $\tau = 1 - 2\alpha$, with probability at least $p_- = \sum_{i=N-\alpha N}^{N-N'} (1-p)^i$, we have the aggregated $\bar{\boldsymbol{g}}_k$ points to the benign $\mathcal{W}^*$ and with probability at most $p_+ = \sum_{i=\frac{1+\tau}{2}N-N'+1}^{N-N'} p^i$ we have the aggregated $\bar{\boldsymbol{g}}_k$ points to the malicious $\boldsymbol{w}'$. In contrast, we have $p_- = \sum_{i=N-\alpha N}^{N-N'} (1-p)^i$ and $p_+ = \sum_{i=\alpha N - N' + 1}^{N-N'} p^i$ if we use the trimmed-mean estimator alone and have $p_- = 0$ and $p_+ = 1$ if using the arithmetic mean estimator with or without the AND-mask.*

Proof. Under the given setup, a sufficient condition for the trimmed-mean pointing toward the benign direction is that the remaining elements that are not trimmed point toward the benign direction. Since the trim threshold α is greater or equal to the corruption level $\frac{N'}{N}$, guaranteeing at least $N - \alpha N$ benign elements point to the benign direction is sufficient for guaranteeing the direction of the remaining elements after trimming. In addition, setting the masking threshold $\tau = 1 - 2\alpha$ guarantees that a dimension with at least $N - \alpha N$ benign elements pointing to the benign direction will not be masked out. With the definition of sign flipping probability p, we have $p_- = \sum_{i=N-\alpha N}^{N-N'} (1-p)^i$ for our invariant aggregator and the trimmed-mean estimator. In contrast, for the arithmetic mean estimator, malicious clients can always use malicious gradient elements that are greater than the sum of benign gradient elements to control the aggregation result. Therefore, we have $p_- = 0$ for the arithmetic mean estimator.

Similarly, a dimension-wise aggregation result may align with the malicious gradient if there exists a remaining element that points to the malicious direction after trimming. Therefore, for the trimmed mean estimator, its failure rate is at most $p_+ = \sum_{i=\alpha N - N' + 1}^{N-N'} p^i$.

However, with AND-mask, only one remaining element pointing to the malicious direction after trimming may not be sufficient for reaching a high sign consistency to pass the mask. For an invariant aggregator with a masking threshold τ, there needs at least $\frac{1+\tau}{2}N - N' + 1$ benign gradient elements flipping their directions and align with the malicious gradient elements, resulting in a smaller failure rate upper bound $p_+ = \sum_{i=\frac{1+\tau}{2}N-N'+1}^{N-N'} p^i$.

This is because if there is less than $\frac{1+\tau}{2}N - N' + 1$ benign update elements flipping their directions and aligning with the malicious update elements, such a dimension (1) will be set to 0 due to low sign consistency or (2) passes the AND-mask and the trimmed-mean estimator recovers the benign gradient direction. In contrast, without masking, there only needs $\alpha N {-} N' + 1$ flipped benign update to cause a failure mode of progressing toward malicious minimum (Figure 2b). Note that $\alpha < \frac{1+\tau}{2}$. Without trimming, a single malicious update may manipulate the arithmetic mean to an arbitrary value in a consistent dimension. □

Theorem 11. *(Convergence) Under Assumptions 1 - 3 and Theorem 10, let $\boldsymbol{w}$ be a initial parameter, suppose $|\boldsymbol{w}_k - \boldsymbol{w}^*_{i,k}| \leq c$ and $\eta_- \leq |\bar{\boldsymbol{g}}_{k,t}| \leq \eta^-$, $\forall i \in \{1, ..., N - N'\}, k \in \{1, ..., \mathrm{d}\}, t \in \{1, ...\}$, $|\bar{\boldsymbol{g}}_{k,t}| > 0$, if the number of round $T \geq \frac{c}{\eta_-}$, with a probability at least $\left[\sum_{i=\frac{c}{\eta_-}}^{T} \mathcal{F}(T, i, p_-) \cdot \sum_{j=\frac{i \cdot n_- - c}{n^-}}^{T-i} \mathcal{F}(T - i, j, \frac{1-p_- - p_+}{1-p_-})\right]^{\mathrm{d}}$ where $\mathcal{F}(T, i, p_-)$ denotes a binomial density function with T trails, i success, and probability p_-, we have $\boldsymbol{w}_T \in \{\boldsymbol{w} \mid \exists \boldsymbol{w}^* \in \mathcal{W}^*, \|\boldsymbol{w} - \boldsymbol{w}^*\| \leq \sqrt{\mathrm{d}}\eta^-\}$.*

Proof. Under the given conditions, an optimization strategy needs at least $\frac{c}{\eta_-}$ steps to converge an initial parameter $\boldsymbol{w}$ to a minimum $\boldsymbol{w}^*_i$ for any given $i \in \{1, ..., N - N'\}$. Therefore, we need to number of rounds $T \geq \frac{c}{\eta_-}$. Then, using Theorem 10, for a given T, within a certain dimension, we have a probability at least $\sum_{i=\frac{c}{\eta_-}}^{T} \mathcal{F}(T, i, p_-)$ the parameter $\boldsymbol{w}_T$ can possibly reach the benign minimum, where $\mathcal{F}(T, i, p_-)$ denotes a binomial density function with T trails, i success, and probability p_-. Here, we are not absolutely certain about the convergence because malicious clients may also mislead the parameter and move away from the benign minima.

To further incorporate malicious clients into consideration, we analyze how many steps in an optimization trajectory can be misled by malicious clients without hurting the convergence to benign minima. Suppose the parameter steps toward benign minima for i steps. We need at least $\frac{i \cdot n_- - c}{n^-}$ steps among the remaining $T - i$ steps to not direct to the malicious minimum, whose probability is at least $\sum_{j=\frac{i \cdot n_- - c}{n^-}}^{T-i} \mathcal{F}(T-i, j, \frac{1-p_- - p_+}{1-p_-})$.

In addition, if malicious clients mislead the aggregation result in the last round, we will not be able to converge the parameter to the benign minima. Therefore, we relax the convergence to the benign minima to the convergence to a neighborhood around the benign minima, whose radius is bounded by the gradient norm $\sqrt{\mathrm{d}}\eta^-$.

Combining $\sum_{i=\frac{c}{\eta_-}}^{T} \mathcal{F}(T, i, p_-)$ and $\sum_{j=\frac{i \cdot n_- - c}{n^-}}^{T-i} \mathcal{F}(T-i, j, \frac{1-p_- - p_+}{1-p_-})$ and taking the product across all dimensions complete the proof. □

C Experimental Setup

C.1 Backdoor Attacks

The edge-case backdoor attack (Wang et al., 2020), distributed backdoor attack (Xie et al., 2020b), and colluding attack (Panda et al., 2022) follow previous works. For the adaptive adversary, we let it intentionally scale up the gradient elements in dimensions that are not masked out such that the gradient norms of the masked malicious gradients remain the same.

C.2 Model Architectures

We use a Resnet-18 (He et al., 2016) on the CIFAR-10 dataset. The GloVe-6B (Pennington et al., 2014) provides word embedding for the Twitter dataset, and a two-layer LSTM model with 200 hidden units further uses the embedding for sentiment prediction. A three-layer 128-256-2 fully connected network is employed to detect phishing emails.

C.3 Hyper-parameters

Table 3 lists the hyper-parameters. We start with $\tau = \frac{N'}{N}$ and increase τ if there needs more robustness.

Table 3: Hyper-parameters

Hyper-parameters	CIFAR-10	Twitter	Phishing
Optimizer	SGD	Adam	SGD
Learning Rate	0.01	0.0001	0.1
Batch Size	64	64	64
Local Epoch	1	0.1	1
Communication Round	300	300	20
Number of Clients	100	100	100
Number of Malicious Clients	20	10	20
Number of Clients per Round	20	15	20
τ	0.2	0.6	0.6
α	0.25	0.25	0.25

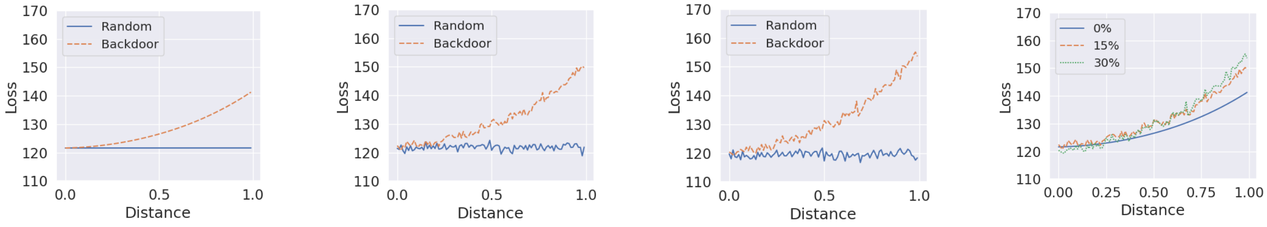
C.4 Phishing Dataset

Each phishing email data sample has 45 standardized numerical features of the sender that represent the sender's reputation scores. A large reputation score may indicate a phishing email. The reputation scores come from peer-reviewers in a reputation system (Jøsang et al., 2007). The adversary may use malicious clients to manipulate the reputation.

D Experimental Results

D.1 Loss Landscape Visualization

Due to the high dimensionality of model parameters, we shall focus on the loss landscape along the gradient directions. Specifically, we consider backdoor gradient directions and random gradient directions as comparisons. Figure 4 shows two loss landscapes with different edge-case backdoor attack configurations, whose backdoor samples have probabilities of 0%, 15%, and 30% to appear on benign data distributions, respectively. Here, the edge-case 0% attack has a most flat loss landscape, complementing our analysis in Section 4.1. We kindly refer to related studies (Zhou et al., 2019; Kleinberg et al., 2018) on epoch-wise convexity for readers that wonder why SGD can escape from the bad minima in the edge-case 15% and 30% loss landscapes.



(a) Edge-case 0%

(b) Edge-case 15%

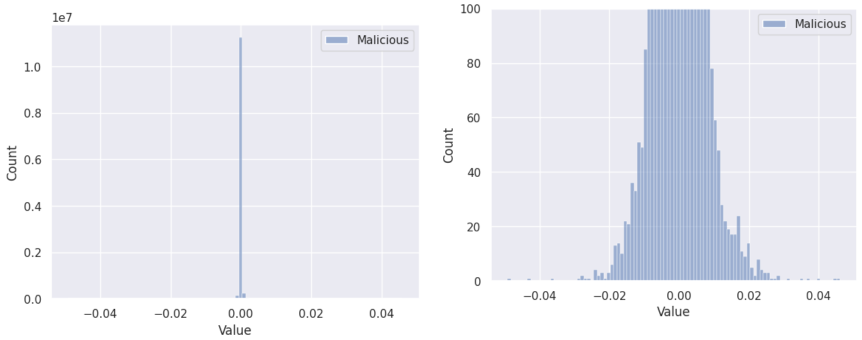
(c) Edge-case 30%

(d) Comparison

Figure 4: Loss landscapes along the direction of malicious and random gradients. Distance is defined using the gradient norm. Distance 0.0 means the parameter stays at a minimum.

D.2 Mimicking Benign Clients

We further empirically verify our analysis in Section 4.1, showing that mimicking benign clients has a low penalty. In this experiment, we consider two gradient vectors, one from a benign client and the other one from a malicious client. Figure 5 plots the histogram of malicious gradient elements and suggests that a majority of them are small, indicating a flat loss landscape. For example, there are only 747 out of 11689512 gradient elements whose absolute value is greater than 0.01. Then, we let the malicious gradient mimic a benign gradient by setting the value of all malicious gradient elements that are not among the 1000 largest ones to the corresponding benign gradient elements. Such a mimicking strategy increases the cosine similarity between the benign and the malicious gradient to 0.983 while only slowing down the *increase* of the attack success loss by 3.85%.



(a) Gradient element value histogram

(b) Gradient element value histogram (zoom-in)

Figure 5: Gradient element value histogram of a malicious gradient.

D.3 Separable Minima

We employ 5 federated learning clients with benign data samples and train a global federated learning model. Then, we let each client do one round of local fine-tuning (a.k.a. personalization) and find that $\boldsymbol{w}' \notin \mathcal{W}^*$.

D.4 Ablation Study

We include ablation studies to demonstrate that the AND-mask and trimmed-mean estimator are necessary for defending against backdoor attacks. Table 4 summarizes the results and shows that neither AND-mask nor trimmed-mean estimator defends against backdoor attacks alone. We also replace the sign consistency in AND-mask using the sample mean-variance ratio.

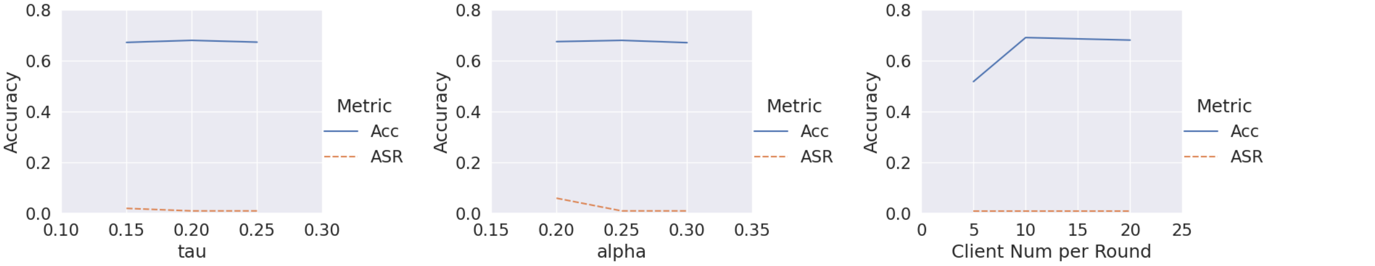
Table 4: Accuracy of Aggregators Under Edge-case Backdoor Attack.

Method	CIFAR-10		Twitter		Phishing	
	Acc	ASR	Acc	ASR	Acc	ASR
Ours	.677 ± .001	.001 ± .001	.687 ± .001	.296 ± .003	.999 ± .001	.000 ± .001
AND-mask	.672 ± .001	.655 ± .001	.652 ± .001	.493 ± .017	.999 ± .001	.999 ± .001
Trimmed-mean	.687 ± .001	.512 ± .001	.728 ± .001	.640 ± .016	.999 ± .001	.999 ± .001
Mean-Variance Ratio	.554 ± .001	.000 ± .001	.613 ± .001	.603 ± .001	.999 ± .000	.333 ± .333

Note: The numbers are average accuracy over three runs. Variance is rounded up.

D.5 Hyper-parameter Sensitivity

This section provides additional experiments using the CIFAR-10 dataset to show that our approach is easy to apply and does not require expensive hyperparameter tuning. Our experiments measure the impact of the masking threshold τ and trim level α on our defense and evaluate our approach with fewer clients per round. Experimental results in Figure 6 suggest that there exists a wide range of threshold configurations that are effective against backdoor attacks. Moreover, our approach can work with as few as 10 clients per round.



(a) Varying τ

(b) Varying α

(c) Varying client number

Figure 6: Performance of our defense with varying hyper-parameters.

D.6 Result Discussion

Vector-wise. Common vector-wise defenses such as Krum estimate pair-wise similarities in terms of Euclidean distance (Blanchard et al., 2017) (Krum and multi-Krum) of cosine similarity (Nguyen et al., 2021) (multi-Krum$_C$) between each gradient and others. The gradients that are dissimilar to others are removed. The vector-wise view is insufficient for defending against backdoor attacks because backdoor attacks can succeed by manipulating a tiny subset of parameters (e.g. 5%) (Wu and Wang, 2021) without incurring much vector-wise difference. In practice, we observe that the malicious gradients can get high similarity scores and circumvent vector-wise defenses. We also find that Krum achieves much lower accuracies than the results from previous works (Wang et al., 2020). We attribute such a difference to the capability of adversaries. We employ a strong adversary that can compromise 20% clients (Karimireddy et al., 2022) at every round, while prior work (Wang et al., 2020) lets an adversary compromise 1 client every 10 rounds.

Dimension-wise. Figure 2b in Section 4 shows the limitation of the trimmed-mean estimator, which was the most effective defense against the edge-case backdoor attack. We also include Sign-SGD with majority vote (Bernstein et al., 2019) as a defense, which binarizes the gradient and takes the majority vote as the aggregation result. However, Sign-SGD struggles to train a large federated model (e.g., Resnet-18 on CIFAR-10) and can suffer from the same failure mode as the median estimator where the clients have diverse signs. Then, the adversary can put more weight on one side and mislead the voting result.

Combination. A naive combination of multi-Krum and the trimmed-mean estimator fails to defend against the backdoor attack because neither multi-Krum nor the trimmed-mean estimator avoids the failure mode of the other.

Weak-DP. The weak-DP defense (Sun et al., 2019) first bounds the gradient norms, then adds additive noise (e.g., Gaussian noise) to the gradient vector. The edge-case backdoor attack can work without scaling up the gradients, circumventing the norm bounding. For the additive noise, we hypothesize that in some dimensions, the difference between malicious and benign gradients can be too large for the Gaussian noise to blur their boundary.

Advanced Defenses. RFA (Pillutla et al., 2022) computes geometric medians as the aggregation result, which is shown to be ineffective (Wang et al., 2020). SparseFed (Panda et al., 2022) only accepts elements with large magnitude in the aggregation results. However, both benign and malicious updates can contribute to large magnitudes. The limitations of FLTrust and trigger inversion are discussed in Section 4.

Robust Learning Rate (RLR). The RLR (Ozdayi et al., 2021) approach is similar to ours, but its design lacks a robustness guarantee. The RLR approach flips the gradient sign if a dimension's sign consistency is low. Although such a strategy can recover the correct gradient direction if the aggregation result is misled and has the same sign as the malicious gradient, it can proactively mislead the aggregation result if the adversary fails. Suppose the adversary uploads a gradient -0.1 in a low-consistency dimension but does not successfully mislead the aggregation result (e.g., 0.5). Then, using the RLR strategy leads to an aggregation result −0.5 and helps the adversary. In contrast, our approach directly sets the aggregation result to 0 in dimensions with low consistency and applies a trimmed-mean estimator for other dimensions.

Table 5: Accuracy of Invariant Aggregators Under Various Continuous Attack.

Method	CIFAR-10		Twitter		Phishing	
	Acc	ASR	Acc	ASR	Acc	ASR
No Attack	.685 ± .001	.000 ± .001	.691 ± .001	.095 ± .001	.999 ± .001	.000 ± .001
Edge-case (Wang et al., 2020)	.677 ± .001	.001 ± .001	.687 ± .001	.296 ± .003	.999 ± .001	.000 ± .001
DBA (Xie et al., 2020b)	.679 ± .001	.001 ± .001	N\A	N\A	.999 ± .001	.000 ± .001
Edge+Collude (Panda et al., 2022)	.675 ± .001	.021 ± .001	.685 ± .002	.324 ± .001	.999 ± .001	.000 ± .001
Edge+Adaptive	.668 ± .001	.013 ± .001	.682 ± .001	.331 ± .002	.999 ± .001	.000 ± .001
Edge+Collude+Adaptive	.667 ± .001	.049 ± .001	.683 ± .001	.335 ± .001	.999 ± .001	.000 ± .001

Note: The numbers are average accuracy over three runs. Variance is rounded up.

D.7 Results under Various Backdoor Attacks

Strategies. We extensively evaluate our approach against multiple state-of-the-art attack strategies (Table 5), including distributed backdoor attack (DBA) (Xie et al., 2020b), colluding attack (Panda et al., 2022), and an adaptive attack that aims to leverage the unmasked gradient elements (Appendix C). Our invariant aggregator remains effective with minor ASR increase under all strategies and their compositions. We further compare our approach against strong baselines under the collude and adaptive strategies on the CIFAR-10 dataset(Tables 6 and 7).

Table 6: Performance of Aggregators Under Collude Adversary on the CIFAR-10 Dataset.

Collude Adversary	Ours	FLTrust	RFA	FLIP
Acc	.675 ± .002	.671 ± .001	.681 ± .002	.665 ± .001
ASR	.021 ± .001	.580 ± .003	.878 ± .002	.271 ± .001

Table 7: Performance of Aggregators Under Adaptive Adversary on the CIFAR-10 Dataset.

Adaptive Adversary	Ours	FLTrust	RFA	FLIP
Acc	.668 ± .001	.672 ± .002	.683 ± .001	.668 ± .001
ASR	.013 ± .001	.543 ± .003	.837 ± .004	.236 ± .001

Poison Ratio. We evaluate our approach with various poison ratios (Table 8), which means the percentage of malicious clients in a federated learning system.

Table 8: Performance of Invariant Aggregator under Various Poison Ratio on the CIFAR-10 Dataset.

Poison Ratio	0%	5%	10%	15%	20%
Acc	.687 ± .002	.683 ± .003	.681 ± .001	.681 ± .002	.677 ± .001
ASR	.000 ± .001	.000 ± .001	.000 ± .001	.001 ± .001	.001 ± .001

Trigger Size. We evaluate our approach with various pixel trigger sizes on the CIFAR-10 dataset (Table 9).

Table 9: Performance of Invariant Aggregator under Various Trigger Sizes on the CIFAR-10 Dataset.

Trigger Size	×1	×2	×4
Acc	.679 ± .001	.677 ± .001	.675 ± .002
ASR	.001 ± .001	.003 ± .001	.003 ± .001

D.8 Results under Different Data Heterogeneity Level.

We evaluate our approach under different data heterogeneity levels on the CIFAR-10 dataset (Table 10). The parameter α in the Dirichlet distribution-based non-i.i.d. data partition controls the heterogeneity level.

Table 10: Performance of Invariant Aggregator under Various Data Heterogeneity Level on the CIFAR-10 Dataset. The parameter α in the Dirichlet distribution-based non-i.i.d. data partition controls the heterogeneity level. The i.i.d.(full) means every client has access to the full dataset locally.

Heterogeneity	i.i.d. (full)	i.i.d. ($\alpha = 1.0$)	non-i.i.d. ($\alpha = 0.5$)	non-i.i.d. ($\alpha = 0.2$)
Acc	.857 ± .003	.696 ± .002	.677 ± .003	.632 ± .001
ASR	.000 ± .001	.000 ± .001	.001 ± .001	.001 ± .001