EXPLOIT GRADIENT SKEWNESS TO CIRCUMVENT BYZANTINE DEFENSES FOR FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) is notorious for its vulnerability to Byzantine attacks. Most current Byzantine defenses share a common inductive bias: among all the gradients, the majorities are more likely to be honest. However, such a bias is a poison to Byzantine robustness due to a newly discovered phenomenon in this paper – gradient skew. We discover that the majority of honest gradients skew away from the optimal gradient (the average of honest gradients) as a result of heterogeneous data distribution. This gradient skew phenomenon allows Byzantine gradients to hide within the skewed majority of honest gradients and thus be recognized as the majority. As a result, Byzantine defenses are deceived into perceiving Byzantine gradients as honest. Motivated by this observation, we propose a novel skew-aware attack called STRIKE: first, we search for the skewed majority of honest gradients; then, we construct Byzantine gradients within the skewed majority. Experiments on three benchmark datasets validate the effectiveness of our attack.

1 INTRODUCTION

Federated Learning (FL) (McMahan et al., 2017; Li et al., 2020) emerged as a privacy-aware learning paradigm, in which data owners, i.e., clients, repeatedly use their private data to compute local gradients and upload them to a central server. The central server collects the uploaded gradients from clients and aggregates these gradients to update the global model. In this way, clients can collaborate to train a model without exposing their private data.

Unfortunately, FL is susceptible to Byzantine attacks due to its distributed nature (Blanchard et al., 2017; Guerraoui et al., 2018). A malicious party can control a small subset of clients, i.e., Byzantine clients, to degrade the utility of the global model. During the training phase, Byzantine clients can send arbitrary messages to the central server to bias the global model. A wealth of defenses (Blanchard et al., 2017; Pillutla et al., 2019; Shejwalkar & Houmansadr, 2021) have been proposed to defend against Byzantine attacks in FL. They aim to estimate the optimal gradient, i.e., the average of gradients from honest clients, in the presence of Byzantine clients.

Most existing defenses (Blanchard et al., 2017; Shejwalkar & Houmansadr, 2021; Karimireddy et al., 2022) share a common inductive bias: the majority gradients are more likely to be honest. Generally, they assign higher weights to the majority gradients. Then they compute the global gradient and use it to update the global model. As a result, the output global gradient of defenses is biased towards the majority of gradients.

However, this inductive bias of Byzantine defenses is harmful to Byzantine robustness in FL due to the presence of gradient skewness. In practical FL, data across different clients is non-independent and identically distributed (non-IID), which gives rise to heterogeneous honest gradients (McMahan et al., 2017; Li et al., 2020; Karimireddy et al., 2022). On closer inspection, we find that these heterogenous honest gradients are highly skewed. In Figure 1, we use Locally Linear Embedding (LLE) (Roweis & Saul, 2000) to visualize the honest gradients on CIFAR-10 dataset (Krizhevsky et al., 2009) when data is non-IID split. Detailed setups and more results are provided in Appendix A. As shown in Figure 1, the majority of honest gradients skew away from the optimal gradient. We term this phenomenon as ”gradient skew”.

When honest gradients are skewed, the defenses' bias towards majority gradients is a poison to Byzantine robustness. In fact, we can hide Byzantine gradients within the skewed majority of honest gradients as shown in Figure 1. In this case, the bias of defenses would drive the global gradient close to the skewed majority but far from the optimal gradient.

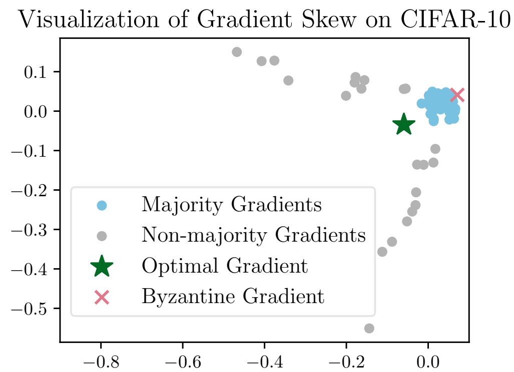


Figure 1: The LLE visualization of honest gradients in the non-IID setting on CIFAR-10. The majority of honest gradients (blue circles) are skewed away from the optimal gradient (green star). In this case, we can hide Byzantine gradients (pink crosses) within the skewed majority of honest gradients to circumvent defenses.

In this paper, we study how to exploit the gradient skew in the more practical non-IID setting to circumvent Byzantine defenses. We first formulate the definition of gradient skew and theoretically analyze the vulnerability of Byzantine defenses under the skew. Based on the above analysis, we design a novel two-Stage aTtack based on gRadIent sKEw called STRIKE. In particular, STRIKE hides Byzantine gradients within the skewed majority of the honest gradients as shown in Figure 1. STRIKE can take advantage of the gradient skew in FL to break Byzantine defenses.

In summary, our contributions are:

- To the best of our knowledge, we are the first to discover the gradient skew phenomenon in FL: the majority of honest gradients are skewed away from the optimal gradient. We theoretically analyze the vulnerability of Byzantine defenses under gradient skew. In particular, we can circumvent defenses by hiding Byzantine gradients within the skewed majority of honest gradients.
- Based on the theoretical analysis, we propose a two-stage Byzantine attack called STRIKE. In the first stage, STRIKE searches for the majority of the honest gradients under the guidance of Karl Pearson's formula. In the second stage, STRIKE constructs the Byzantine gradients within the skewed majority by solving a constrained optimization problem.
- Experiments on three benchmark datasets validate the effectiveness of the proposed attack. For instance, STRIKE attack improves upon the best baseline by 57.84% against DnC on FEMNIST dataset when there are 20% Byzantine clients.

2 Related Works

Byzantine attacks. Blanchard et al. (2017) first disclose the Byzantine vulnerability of FL. Baruch et al. (2019) observe that the variance of honest gradients is high enough for Byzantine clients to compromise Byzantine defenses. Based on this observation, they propose a LIE attack that hides Byzantine gradients within the variance. Xie et al. (2020) further utilize the high variance and propose an IPM attack. Particularly, they show that when the variance of honest gradients is large enough, IPM can make the inner product between the aggregated gradient and the honest average negative. However, this result is restricted to a few defenses, i.e., Median Yin et al. (2018), Trmean Yin et al. (2018), and Krum Blanchard et al. (2017). Fang et al. (2020) establish an omniscient attack called Fang. However, the Fang attack requires knowledge of the Byzantine defense, which is unrealistic in practice. Shejwalkar & Houmansadr (2021) propose Min-Max and Min-Sum attacks that solve a constrained optimization problem to determine Byzantine gradients. From a high level, both Min-Max and Min-Sum aim to maximize the perturbation to a reference benign gradient while ensuring the Byzantine gradients lie within the variance. Karimireddy et al. (2022) propose a Mimic attack that takes advantage of data heterogeneity in FL. In particular, Byzantine clients pick an honest client to mimic and copy its gradient. The above attacks take advantage of the large variance of honest gradients to break Byzantine defenses. However, they all ignore the skew nature of honest gradients in FL and fail to exploit this vulnerability.

Byzantine resilience. El-Mhamdi et al. (2021); Farhadkhani et al. (2022); Karimireddy et al. (2022) provide state-of-the-art theoretical analysis of Byzantine resilience under data heterogeneity. El-Mhamdi et al. (2021) discuss Byzantine resilience in a decentralized, asynchronous setting. Farhadkhani et al. (2022) provide a unified framework for Byzantine resilience analysis, which enables comparison among different defenses on a common theoretical ground. Karimireddy et al. (2022) improve the error bound of Byzantine resilience to be upper-bounded by the fraction of Byzantine clients, which recovers the standard convergence rate when there are no Byzantine clients. They all share a common bias: the majority of gradients are more likely to be honest. However, this bias is a poison to Byzantine robustness in the presence of gradient skew. In practical FL, the distribution of honest gradients is highly skewed due to data heterogeneity. Therefore, existing defenses are especially vulnerable to attacks that are aware of gradient skew.

3 Notations and Preliminary

3.1 Notations

$\|\cdot\|$ denotes the ℓ_2 norm of a vector. For vector $\boldsymbol{v}$, $(\boldsymbol{v})_k$ represents the k-th coordinate of $\boldsymbol{v}$. Model parameters are denoted by $\boldsymbol{w}$ and gradients are denoted by $\boldsymbol{g}$. We use $\bar{\boldsymbol{g}}$ to denote the optimal gradient, i.e., the average of honest gradients, and $\hat{\boldsymbol{g}}$ denotes the global gradients obtained by Byzantine defenses. We use subscript i to denote client i and use superscript t to denote communication round t.

3.2 Preliminary

Federated learning. Suppose that there are n clients and a central server. The goal is to optimize the global loss function $\mathcal{L}(\cdot)$:

$$\min_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}), \quad \text{where } \mathcal{L}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{w}). \tag{1}$$

Here $\boldsymbol{w}$ is the model parameter, and $\mathcal{L}_i(\cdot)$ is the local loss function on client i for $i = 1, \ldots, n$.

In communication round t, the central server distributes global parameter $\boldsymbol{w}^t$ to the clients. Each client i performs several epochs of SGD to minimize its local loss function $\mathcal{L}_i(\cdot)$ and update its local parameter to $\boldsymbol{w}_i^{t+1}$. Then, each client i computes its local gradient $\boldsymbol{g}_i^t$ and sends it to the server.

$$\boldsymbol{g}_i^t = \boldsymbol{w}_i^t - \boldsymbol{w}_i^{t+1}, \quad i = 1, \ldots, n. \tag{2}$$

After receiving the uploaded local gradients, the server aggregates the local gradients and updates the global model to $\boldsymbol{w}^{t+1}$.

$$\bar{\boldsymbol{g}}^t = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^t, \quad \boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \bar{\boldsymbol{g}}^t. \tag{3}$$

Byzantine attack model. Assume that among the total n clients, f fixed clients are Byzantine clients. Let $\mathcal{B} \subseteq \{1, \ldots, n\}$ denote the set of Byzantine clients and $\mathcal{H} = \{1, \ldots, n\} \setminus \mathcal{B}$ denote the set of honest clients. In each communication round, Byzantine clients can send arbitrary messages to bias the global model. The local gradients that the server receives in the t-th communication round are

$$\boldsymbol{g}_i^t = \begin{cases} *, & i \in \mathcal{B}, \\ \boldsymbol{w}^t - \boldsymbol{w}_i^{t+1}, & i \in \mathcal{H}, \end{cases} \tag{4}$$

where $*$ represents an arbitrary message. Following Baruch et al. (2019); Xie et al. (2020), we consider the setting where the attacker only has the knowledge of honest gradients.

Byzantine resilience. Blanchard et al. (2017) show that the popular mean aggregation rule is not resilient to Byzantine attacks. Thus, the server replaces the mean aggregation rule in Equation (3) with a robust AGgregation Rules (AGR) $\mathcal{A}$, e.g., Krum (Blanchard et al., 2017), Median (Yin et al., 2018), to compute the global gradient $\hat{\boldsymbol{g}}^t$ and update the global model to $\boldsymbol{w}^{t+1}$.

$$\hat{\boldsymbol{g}}^t = \mathcal{A}(\boldsymbol{g}_1^t, \ldots, \boldsymbol{g}_n^t), \quad \boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \hat{\boldsymbol{g}}^t. \tag{5}$$

A body of recent works (Farhadkhani et al., 2022; Karimireddy et al., 2022; Allouah et al., 2023) have theoretically defined Byzantine resilience for general robust AGRs. Particularly, we adopt the definition from Farhadkhani et al. (2022) in this work for analysis. We also discuss how our analysis can apply to other definitions of Byzantine resilience in Appendix B.2.

Definition 1 ((f, λ)-resilient)**.** Given $f < n$ and $\lambda \geq 0$, an AGR $\mathcal{A}$ is (f, λ)-resilient if for any collection of n vectors $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$ and any set $\mathcal{G} \subseteq \{1, \ldots, n\}$ of size $n - f$,

$$\|\mathcal{A}(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n) - \bar{\boldsymbol{g}}_{\mathcal{G}}\| \leq \lambda \max_{i,j \in \mathcal{G}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|, \tag{6}$$

where $\bar{\boldsymbol{g}}_{\mathcal{G}} = \sum_{i \in \mathcal{G}} \boldsymbol{g}_i/(n - f)$ is the average of gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{G}\}$.

Essentially, a smaller λ means better resilience (Farhadkhani et al., 2022).

4 Vulnerability of Robust AGRs under Gradient Skew

In this section, we show that when honest gradients are skewed, we can establish Byzantine attacks to circumvent robust AGgregation Rules (AGRs). First, we verify the existence of gradient skew in FL and formally define gradient skew. Then, we show how to exploit the gradient skew to launch Byzantine attacks and circumvent robust AGRs.

4.1 Gradient Skew in FL Due to Non-IID data

Plenty of works (Baruch et al., 2019; Xie et al., 2020; Karimireddy et al., 2022) have explored how large variance can be harmful to Byzantine robustness. However, to the best of our knowledge, none of the existing works is aware of the skewed nature of honest gradients in the non-IID setting and how gradient skew can threaten Byzantine robustness.

We take a close look at the distribution of honest gradients in the non-IID setting (without attack). To construct our FL setup, we split CIFAR-10 (Krizhevsky et al., 2009) dataset in a non-IID manner among 100 clients. For more setup details, please refer to Appendix A.1. We run FedAvg (McMahan et al., 2017) for 200 communication rounds. We randomly sample a communication round and use Locally Linear Embedding (LLE) (Roweis & Saul, 2000) to visualize the gradients in this communication round in Figure 1. From Figure 1, we observe that the majority of honest gradients (blue circles) skew away from the optimal gradient (green stars). More visualization results can be found in Appendix A.2. We name this phenomenon ”gradient skew”.

We formulate the definition of gradient skew for further analysis. The idea behind this definition is to measure the skewness of honest gradients by the distance between the majority of honest gradients and the optimal gradient, i.e., the average of honest gradients.

Definition 2 ((f, γ)-skewed)**.** The set of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ is called (f, γ)-skewed if there exists a set $\mathcal{S} \subseteq \mathcal{H}$ of size $n - 2f$ such that

$$\mathbb{E}\left[\|\bar{\boldsymbol{g}}_{\mathcal{S}} - \bar{\boldsymbol{g}}\|^2\right] \geq \gamma \rho_{\mathcal{S}}^2, \tag{7}$$

where $\bar{\boldsymbol{g}} = \sum_{i \in \mathcal{H}} \boldsymbol{g}_i/(n-f)$, $\bar{\boldsymbol{g}}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \boldsymbol{g}_i/(n-2f)$, and $\rho_{\mathcal{S}}^2 = \mathbb{E}\left[\max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2\right]$ is a measure of gradient heterogeneity introduced by El-Mhamdi et al. (2021). Here, gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$ are called the *skewed majority*, and γ is called the skewness of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$.

In Definition 2, γ measures the skew degree of the honest gradients. A larger γ indicates a higher skew degree. The size of the skewed majority $|\mathcal{S}| = n - 2f$ is carefully chosen as follows: We aim to deceive robust AGR into believing that Byzantine gradients and majority gradients are honest, and the non-majority gradients are Byzantine. There are $n - f$ honest gradients and f Byzantine gradients. Thus the number of majority gradients is determined by $|\mathcal{S}| = (n - f) - f = n - 2f$.

4.2 Robust AGRs are Brittle under Gradient Skew

When the honest gradients are skewed, robust AGRs are extremely vulnerable. In fact, we can hide Byzantine gradients within the majority of honest gradients. This attack strategy makes Byzantine gradients stealthy and difficult to detect. The skewed nature of the majority further allows Byzantine gradients to deviate the global gradient away from the optimal gradient. The above argument can be formulated as the following lower bound.

Proposition 1 (Vulnerability under skew)**.** *Given any* (f, λ)*-resilient AGR* $\mathcal{A}$*, if the set of honest gradients* $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ *is* (f, γ)*-skewed, then there exist Byzantine gradients* $\{\boldsymbol{g}_i \mid i \in \mathcal{B}\}$ *such that*

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n) - \bar{\boldsymbol{g}}\|^2] \geq \Omega(\frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho_{\mathcal{S}}^2). \tag{8}$$

where $\bar{\boldsymbol{g}} = \sum_{i \in \mathcal{H}} \boldsymbol{g}_i/(n-f)$ *is the optimal gradient,* $\rho_{\mathcal{S}}^2 = \mathbb{E}[\max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2]$*,* $\mathcal{S}$ *is the index set of the skewed majority.*

The detailed proof is provided in Appendix B.1.2. Proposition 1 suggests that when the honest gradients are skewed, we can always launch Byzantine attacks to deviate the global gradient from the optimal gradient. Moreover, the more skewed the honest gradients are, the farther the global gradient is from the optimal gradient. An interesting result in Proposition 1 is that smaller λ leads to a larger lower bound in Equation (8), which implies that our attack is even more effective on robust AGRs with stronger resilience. This is because the global gradient obtained by robust AGRs with stronger resilience is closer to the majority of uploaded gradients (including Byzantine and honest). And the majority of uploaded gradients are away from the optimal gradients under our attack. Therefore, a robust AGR with stronger resilience is even more sensitive to our attack.

We further show that the above vulnerability enables us to prevent the global model from converging to the optimum for any L-smooth global loss function and unbiased honest gradients. These assumptions are standard in Byzantine robust learning (Karimireddy et al., 2021; Farhadkhani et al., 2022).

Assumption 1 (L-smooth)**.** The loss function is L-smooth, i.e.,

$$\|\nabla \mathcal{L}(\boldsymbol{w}) - \nabla \mathcal{L}(\boldsymbol{w}')\| \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|, \quad \forall \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d. \tag{9}$$

Assumption 2 (Unbias)**.** The stochastic gradients sampled from any local data distribution are unbiased estimators of local gradients for all clients, i.e.,

$$\mathbb{E}[\boldsymbol{g}_i^t] = \nabla \mathcal{L}(\boldsymbol{w}^t), \quad \forall i = 1, \ldots n, t = 0, \ldots, T-1. \tag{10}$$

Now we present our main result.

Proposition 2. *Given any* (f, λ)*-resilient AGR* $\mathcal{A}$*, L-smooth global loss function* $\mathcal{L}$*, and learning rate* $\eta \leq 1/L$*, if honest gradients* $\{\boldsymbol{g}_i^t \mid i \in \mathcal{H}\}$ *are* (f, γ)*-skewed for all* $t = 0, \ldots, T-1$*, then there exist Byzantine gradients* $\{\boldsymbol{g}_b^t \mid b \in \mathcal{B}, t = 0, \ldots, T-1\}$ *such that the global model parameter is bounded away from the global optimum* $\boldsymbol{w}^*$*:*

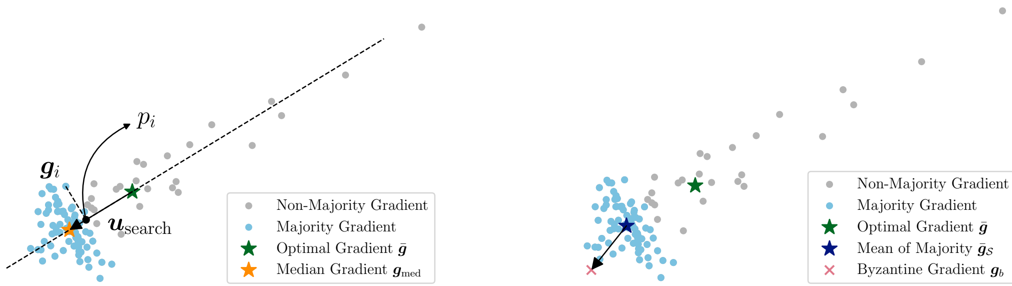
$$\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2] \geq \Omega(\eta^2(1 - L\eta)^2 \cdot \frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2), \quad t = 1, \ldots, T, \tag{11}$$

where $\boldsymbol{w}^t$ *is the parameter of global model in the* t*-th communication round,* $\boldsymbol{w}^*$ *is the global optimum of global loss function* $\mathcal{L}$*,* $\rho^2 = \min_{t=0,\ldots,T-1} \mathbb{E}[\max_{i,j \in \mathcal{S}^t} \|\boldsymbol{g}_i^t - \boldsymbol{g}_j^t\|^2]$*, and* $\mathcal{S}^t$ *is the index set of the skewed majority of honest gradients in* t*-th communication round.*

The proof of Proposition 2 can be found in Appendix B.1.3. Proposition 2 indicates that under gradient skew, we can establish Byzantine attacks to keep the global model away from the optimum. The lower bound in Proposition 2 is also aligned with the one in Proposition 1: a larger skewness γ would lead to a larger lower bound, and so does a smaller λ. Note that we do not require the loss function to be non-convex, which implies that Proposition 2 also applies to more challenging convex loss functions.

5 Proposed Attack

In this section, we introduce the proposed two-Stage aTtack based on gRadIent sKEw called STRIKE. As discussed in Section 4, the attack principle of STRIKE is to hide Byzantine gradients within the skewed majority of honest gradients. To achieve this goal, we carry out STRIKE attack in two stages: in the first stage, we search for the skewed majority of honest gradients; in the second stage, we construct Byzantine gradients within the skewed majority found in the first stage. The procedure of STRIKE attack is shown in Algorithm 1.



(a) We search along the direction $\boldsymbol{u}_{\text{search}} = \boldsymbol{g}_{\text{med}} - \bar{\boldsymbol{g}}$. The honest gradients with the largest scalar projection p_i are selected as the skewed majority of honest gradients (blue circles).

(b) We start from the average of skewed majority $\bar{\boldsymbol{g}}_{\mathcal{S}}$ (dark blue star) and select α such that Byzantine gradient $\boldsymbol{g}_b$ (pink cross) lies within the skewed majority.

Figure 2: Illustration of the proposed two-stage attack STRIKE: in the first stage, STRIKE searches for the skewed majority of honest gradients; in the second stage, STRIKE hides Byzantine gradients within the skewed majority.

Search for the skewed majority. To hide the Byzantine gradient in the skewed majority of the honest gradients, we first need to find the skewed majority. Naively searching the majority according to Definition 2 is computationally expensive. Therefore, we perform a heuristic search motivated by Karl Pearson's formula (Knoke et al., 2002; Moore et al., 2009). Figure 2a illustrates the search procedure in this stage.

As visualized in Figure 1, gradients are densely distributed within the skewed majority, which implies that the population mode coincides with the skewed majority with high probability. Karl Pearson's formula (Knoke et al., 2002; Moore et al., 2009) implies that the mode and median lie on the same side of the mean. Therefore, we search for the skewed majority along the direction $\boldsymbol{u}_{\text{search}}$ defined as:

$$\boldsymbol{u}_{\text{search}} = \boldsymbol{g}_{\text{med}} - \bar{\boldsymbol{g}}, \tag{12}$$

where $\boldsymbol{g}_{\text{med}}$ is the coordinate-wise median of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$, i.e., the k-th coordinate of $\boldsymbol{g}_{\text{med}}$ is $(\boldsymbol{g}_{\text{med}})_k = \text{median}\{(\boldsymbol{g}_i)_k \mid i \in \mathcal{H}\}$, and $\bar{\boldsymbol{g}} = \sum_{i \in \mathcal{H}} \boldsymbol{g}_i/(n-f)$ is the average of honest gradients.

For each honest gradient $\boldsymbol{g}_i$, we compute its scalar projection p_i on the searching direction $\boldsymbol{u}_{\text{search}}$:

$$p_i = \langle \boldsymbol{g}_i, \frac{\boldsymbol{u}_{\text{search}}}{\|\boldsymbol{u}_{\text{search}}\|} \rangle, \quad \forall i \in \mathcal{H}, \tag{13}$$

where $\langle \cdot, \cdot \rangle$ represents the inner product. The $n - 2f$ gradients with the highest scalar projection values are identified as the skewed majority. The goal is to have AGR consider the selected $n - 2f$ gradients as honest and the unselected f gradients as Byzantine. Let $\mathcal{S}$ denote index set, that is

$$\mathcal{S} = \text{Set of } (n-2f) \text{ indices of the gradients with the highest scalar projection } p_i, \tag{14}$$

then the skewed majority of honest gradients are $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$.

Hide Byzantine gradients within the skewed majority. In this stage, we aim to hide Byzantine gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{B}\}$ within the skewed majority $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$ identified in stage 1. The primary goal of our attack is to disguise Byzantine gradients and the skewed majority $\{\boldsymbol{g}_i \mid i \in \mathcal{B} \cup \mathcal{S}\}$ as honest gradients. Meanwhile, the secondary goal is to maximize the attack effect, i.e., maximize the distance between these "fake" honest gradients and the optimal gradient. The hiding procedure in this stage is illustrated in Figure 2b.

According to Definition 1, robust AGRs are sensitive to the diameter of gradients. Therefore, we ensure that the Byzantine gradients lie within the diameter of the skewed majority in order not to be detected.

$$\|\boldsymbol{g}_b - \boldsymbol{g}_s\| \leq \max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|, \quad \forall b \in \mathcal{B}, s \in \mathcal{S}. \tag{15}$$

Algorithm 1 STRIKE Attack

Require: Honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$, hyperparameter $\nu > 0$ that controls attack strength (default $\nu = 1$)

$\boldsymbol{g}_{\text{med}} \leftarrow$ Coordinate-wise median of $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ ▷ Stage 1: search for the skewed majority
$\boldsymbol{u}_{\text{search}} \leftarrow \boldsymbol{g}_{\text{med}} - \bar{\boldsymbol{g}}$
for $i \in \mathcal{H}$ **do**
  $p_i \leftarrow \langle \boldsymbol{g}_i, \boldsymbol{u}_{\text{search}} / \|\boldsymbol{u}_{\text{search}}\| \rangle$
end for
$\mathcal{S} \leftarrow$ Set of $n - f$ indices of honest gradients with the highest p_i
$\bar{\boldsymbol{g}}_{\mathcal{S}} \leftarrow \sum_{i \in \mathcal{S}} \boldsymbol{g}_i / (n - 2f)$ ▷ Stage 2: hide Byzantine gradients within the skewed majority
$\boldsymbol{\sigma}_{\mathcal{S}} \leftarrow$ Coordinate-wise standard deviation of $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$
solve Equation (19) for α
for $b \in \mathcal{B}$ **do**
  $\boldsymbol{g}_b \leftarrow \bar{\boldsymbol{g}}_{\mathcal{S}} + \nu\alpha \cdot \text{sign}(\bar{\boldsymbol{g}}_{\mathcal{S}} - \bar{\boldsymbol{g}}) \odot \boldsymbol{\sigma}_{\mathcal{S}}$
end for
return Byzantine gradients $\{\boldsymbol{g}_b \mid g \in \mathcal{B}\}$

Meanwhile, we want to maximize the attack effect. Therefore, we need to maximize the distance between $\bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}} = \sum_{i \in \mathcal{S}\cup\mathcal{B}} \boldsymbol{g}_i/(n-f)$ and the optimal gradient.

$$\max_{\{\boldsymbol{g}_b \mid b \in \mathcal{B}\}} \|\bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}} - \bar{\boldsymbol{g}}\|. \tag{16}$$

In summary, our objective can be formulated as the following constrained optimization problem.

$$\max_{\{\boldsymbol{g}_b \mid b \in \mathcal{B}\}} \|\bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}} - \bar{\boldsymbol{g}}\| \quad \text{s.t.} \quad \begin{cases} \bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}} = \sum_{i \in \mathcal{S}\cup\mathcal{B}} \boldsymbol{g}_i/(n-f) \\ \|\boldsymbol{g}_b - \boldsymbol{g}_s\| \le \max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|, \quad \forall b \in \mathcal{B}, s \in \mathcal{S} \end{cases} \tag{17}$$

Equation (17) is too complex to be solved due to the high complexity of its feasible region. Therefore, we restrict $\{\boldsymbol{g}_b \mid b \in \mathcal{B}\}$ to the following form:

$$\boldsymbol{g}_b = \bar{\boldsymbol{g}}_{\mathcal{S}} + \alpha \cdot \text{sign}(\bar{\boldsymbol{g}}_{\mathcal{S}} - \bar{\boldsymbol{g}}) \odot \boldsymbol{\sigma}_{\mathcal{S}}, \quad \forall b \in \mathcal{B}, \tag{18}$$

where $\bar{\boldsymbol{g}}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \boldsymbol{g}_i/(n-2f)$ is the average of the skewed majority of honest gradients, α is a non-negative real number that controls the attack strength, $\text{sign}(\cdot)$ returns the element-wise indication of the sign of a number, $\odot$ is the element-wise multiplication, and $\boldsymbol{\sigma}_{\mathcal{S}}$ is the element-wise standard deviation of skewed majority $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$. $\bar{\boldsymbol{g}}_{\mathcal{S}}$ lies within the feasible region of Equation (17), which ensures that $\{\boldsymbol{g}_b \mid b \in \mathcal{B}\}$ are feasible when $\alpha = 0$. $\text{sign}(\bar{\boldsymbol{g}}_{\mathcal{S}} - \bar{\boldsymbol{g}})$ controls the element-wise attack direction, and ensures that $\boldsymbol{g}_b$ is farther away from the optimal gradient $\bar{\boldsymbol{g}}$ under a larger α. $\boldsymbol{\sigma}_{\mathcal{S}}$ controls the element-wise attack strength and ensures that Byzantine gradients are covert in each dimension.

With the restriction in Equation (18), Equation (17) can be simplified to the following optimization problem,

$$\max \alpha \quad \text{s.t.} \quad \|\bar{\boldsymbol{g}}_{\mathcal{S}} + \alpha \cdot \text{sign}(\bar{\boldsymbol{g}}_{\mathcal{S}}) \odot \boldsymbol{\sigma}_{\mathcal{S}} - \boldsymbol{g}_s\| \le \max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|, \quad \forall s \in \mathcal{S}, \tag{19}$$

which can be easily solved by the bisection method described in Appendix C. While α that solves Equation (19) is theoretically provable (as shown in Appendix B.1.2, the proof of Proposition 1), we find in practice that an adjusted attack strength can further improve the effect of STRIKE. We use an additional hyperparameter $\nu(> 0)$ to control the attack strength of STRIKE. STRIKE sets $\boldsymbol{g}_b = \bar{\boldsymbol{g}}_{\mathcal{S}} + \nu\alpha \cdot \text{sign}(\bar{\boldsymbol{g}}_{\mathcal{S}}) \odot \boldsymbol{\sigma}_{\mathcal{S}} - \boldsymbol{g}_i$ for all $b \in \mathcal{B}$ and uploads Byzantine gradients to the server. Higher ν implies higher attack strength. We discuss the performance of STRIKE with different ν in Appendix D.2.1.

6 Experiments

6.1 Experimental Setups

Datasets. Our experiments are conducted on three real-world datasets: CIFAR-10 (Krizhevsky et al., 2009), a subset of ImageNet (Russakovsky et al., 2015) refered as ImageNet-12 (Li et al.,

Table 1: Accuracy (mean±std) under different attacks against different defenses on CIFAR-10, ImageNet-12, and FEMNIST. The best attack performance is in bold (the *lower*, the better).

CIFAR-10							
Attack	Multi-Krum	Median	RFA	Aksel	CClip	DnC	RBTM
BitFlip	54.76 ± 0.06	53.73 ± 2.05	56.04 ± 3.13	51.99 ± 2.04	54.44 ± 0.46	60.81 ± 0.56	55.21 ± 3.72
LIE	57.89 ± 0.22	49.20 ± 3.27	53.90 ± 5.43	46.73 ± 4.86	63.11 ± 0.43	61.58 ± 2.85	58.84 ± 0.64
IPM	47.55 ± 1.75	51.68 ± 1.85	55.36 ± 2.10	56.85 ± 2.07	58.75 ± 5.59	62.30 ± 3.60	48.43 ± 0.17
MinMax	59.44 ± 3.41	57.27 ± 0.63	60.20 ± 1.63	57.17 ± 5.50	59.38 ± 5.15	62.53 ± 2.67	57.72 ± 2.94
MinSum	55.47 ± 1.70	52.27 ± 0.53	54.59 ± 2.38	56.43 ± 1.74	54.70 ± 1.96	61.89 ± 1.62	46.78 ± 0.32
Mimic	56.00 ± 4.26	52.55 ± 0.89	53.61 ± 0.86	57.19 ± 2.50	51.00 ± 0.11	62.10 ± 5.22	46.77 ± 2.52
STRIKE (Ours)	**42.90** ± 1.97	**48.29** ± 0.40	**52.92** ± 1.75	**38.31** ± 0.47	**50.67** ± 0.27	**59.16** ± 1.84	**44.82** ± 0.97
ImageNet-12							
Attack	Multi-Krum	Median	RFA	Aksel	CClip	DnC	RBTM
BitFlip	59.62 ± 0.73	58.56 ± 4.80	59.71 ± 5.00	61.64 ± 1.98	14.87 ± 1.58	59.78 ± 1.50	58.49 ± 1.99
LIE	62.66 ± 0.30	51.41 ± 1.52	60.99 ± 1.22	54.14 ± 3.14	16.19 ± 3.95	67.85 ± 2.87	67.12 ± 0.39
IPM	52.66 ± 2.01	59.20 ± 2.44	61.25 ± 0.62	59.17 ± 1.27	14.33 ± 5.95	66.31 ± 3.60	55.93 ± 0.57
MinMax	68.17 ± 1.91	67.76 ± 0.07	63.05 ± 0.75	59.33 ± 3.85	20.99 ± 3.07	68.05 ± 1.59	65.99 ± 1.26
MinSum	57.50 ± 3.09	58.78 ± 2.10	64.04 ± 0.69	67.15 ± 0.32	16.38 ± 2.70	68.69 ± 1.18	61.70 ± 1.62
Mimic	66.86 ± 0.04	59.39 ± 6.07	60.45 ± 7.09	58.94 ± 1.27	11.35 ± 2.26	69.07 ± 4.69	55.26 ± 1.30
STRIKE (Ours)	**27.24** ± 1.63	**42.98** ± 1.62	**43.30** ± 3.13	**38.11** ± 1.02	**8.33** ± 1.85	**53.40** ± 4.94	**38.81** ± 0.65
FEMNIST							
Attack	Multi-Krum	Median	RFA	Aksel	CClip	DnC	RBTM
BitFlip	82.67 ± 5.13	71.57 ± 3.61	83.41 ± 4.33	81.42 ± 3.45	83.85 ± 8.50	83.58 ± 5.20	82.58 ± 6.08
LIE	68.11 ± 6.86	58.38 ± 7.06	66.19 ± 7.93	38.48 ± 3.32	73.03 ± 3.86	77.42 ± 5.60	53.35 ± 5.17
IPM	84.12 ± 3.06	72.60 ± 8.42	83.42 ± 4.13	78.28 ± 7.37	84.93 ± 4.41	83.03 ± 5.02	83.21 ± 6.42
MinMax	68.42 ± 5.91	66.44 ± 5.88	71.55 ± 5.98	34.22 ± 4.94	72.12 ± 4.39	75.40 ± 3.78	59.23 ± 3.41
MinSum	62.06 ± 3.13	65.46 ± 3.66	70.36 ± 7.24	44.91 ± 3.90	75.40 ± 4.88	77.11 ± 3.61	68.10 ± 8.86
Mimic	83.15 ± 3.46	74.00 ± 4.79	83.87 ± 3.00	79.06 ± 7.21	83.94 ± 5.25	82.22 ± 5.40	81.92 ± 3.40
STRIKE (Ours)	**22.13** ± 7.78	**55.19** ± 3.49	**39.43** ± 5.06	**16.58** ± 3.63	**18.88** ± 4.30	**17.56** ± 5.95	**39.33** ± 11.98

2021b) and FEMNIST (Caldas et al., 2018). Please refer to Appendix D.1.1 for more details about the data distribution.

Baseline attacks. We consider six state-of-the-art attacks: BitFlip (Allen-Zhu et al., 2020), LIE (Baruch et al., 2019), IPM (Xie et al., 2020), Min-Max (Shejwalkar & Houmansadr, 2021), Min-Sum (Shejwalkar & Houmansadr, 2021), and Mimic (Karimireddy et al., 2022). The detailed hyperparameter settings of these attacks are shown in Appendix D.1.2.

Evaluated defenses. We evaluate the performance of our attack on the following robust AGRs: Multi-Krum Blanchard et al. (2017), Median Yin et al. (2018), RFA Pillutla et al. (2019), Aksel Boussetta et al. (2021), CClip Karimireddy et al. (2021) DnC Shejwalkar & Houmansadr (2021), and RBTM El-Mhamdi et al. (2021). Besides, we also consider a simple yet effective bucketing scheme Karimireddy et al. (2022) that adapts existing robust AGRs to the non-IID setting. The detailed hyperparameter settings of the above robust AGRs are listed in Appendix D.1.3.

More detailed setups are deferred to Appendix D.1.

6.2 Experiment Results

Attacking against various robust AGRs. Table 1 demonstrates the performance of seven different attacks against seven robust AGRs on CIFAR-10, ImageNet-12, and FEMNIST datasets. From Table 1, we can observe that our STRIKE attack generally outperforms all the baseline attacks against various defenses on all datasets, verifying the efficacy of our STRIKE attack. On ImageNet-12 and FEMNIST, the improvement of STRIKE over the best baselines is more significant. We hypothesize that this is because the skew degree is higher on ImageNet-12 and FEMNIST compared to CIFAR-10. Since STRIKE exploits gradient skew to launch Byzantine attacks, it is more effective on ImageNet-12 and FEMNIST. DnC demonstrates almost the strongest resilience to previous baseline attacks. This is because these attacks fail to be aware of the skew nature of honest gradients in FL. By contrast, our STRIKE attack can take advantage of gradient skew and circumvent DnC defense. The above observations clearly validate the superiority of STRIKE.

Attacking against robust AGRs with bucketing. Figure 3 demonstrates the performance of seven different attacks against the bucketing scheme (Karimireddy et al., 2022) with different robust AGRs.

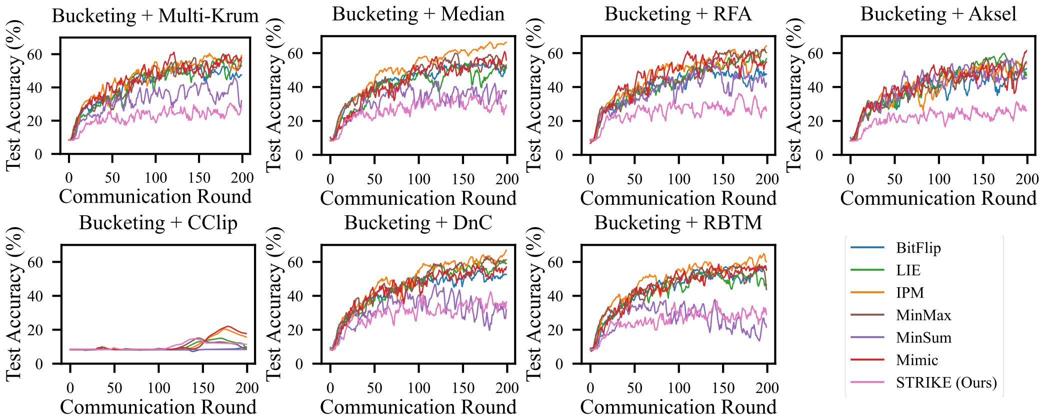


Figure 3: Accuracy under different attacks against seven robust AGRs with bucketing on ImageNet-12. The *lower*, the better.

The results demonstrate that our STRIKE attack works best against Multi-Krum, RFA, and Aksel. When attacking against DnC, Median, and RBTM, only MinSum attack is comparable to our STRIKE attack.

Imparct of ν on STRIKE attack. We study the influence of ν on ImageNet-12 dataset. We report the test accuracy under STRIKE attack with ν in $\{0.25 * i \mid i = 1, \ldots, 8\}$ against seven different defenses on ImageNet-12. As shown in Figure 5, the performance of STRIKE is generally competitive with varying ν. In most cases, simply setting $\nu = 1$ can beat almost all the attacks (except for CClip, yet we observe that the performance is low enough to make the model useless).

The effectiveness of STRIKE attack under different non-IID levels. We vary Dirichlet concentration parameter β in $\{0.1, 0.2, 0.5, 0.7, 0.9\}$ to study how our attack behaves under different non-IID levels. We additionally test the performance in the IID setting. As shown in Figure 6, the accuracy generally increases as β decreases for all attacks. The accuracy under our STRIKE attack is consistently lower than that of all the baseline attacks. Besides, we also note that the accuracy gap between our STRIKE attack and other baseline attacks gets smaller when the non-IID level decreases. We hypothesize the reason is that gradient skew becomes milder as the non-IID level decreases, which aligns with our theoretical results in Propositions 1 and 2. Even in the IID setting, our STRIKE attack is competitive compared to other baselines.

The performance of STRIKE attack with different Byzantine client ratio. We vary the number of Byzantine clients f in $\{5, 10, 15, 20\}$ and fix the total number of clients n to be 50. In this way, Byzantine client ratio f/n varies in $\{0.1, 0.2, 0.3, 0.4\}$ to study how our attack behaves under different Byzantine client ratio. As shown in Figure 7, the accuracy generally decreases as f/n increases for all attacks. The accuracy under our STRIKE attack is consistently lower than that under all the baseline attacks.

The performance of STRIKE attack with different client number. We vary the number of total clients n in $\{10, 30, 50, 70, 90\}$ and set the number of Byzantine clients $f = 0.2n$ accordingly. The results are plotted in Figure 8 in Appendix D.2.4. As shown in Figure 8, the accuracy generally decreases as client number n increases for all attacks. The accuracy under our STRIKE attack is consistently lower than that under all the baseline attacks with different number of clients.

7 Conclusion

In this paper, we theoretically analyze the vulnerability of existing defenses in the non-IID setting due to the skewed nature of honest gradients. Based on the analysis, we propose a novel attack called STRIKE that can exploit the vulnerability. Generally, STRIKE hides Byzantine gradients within the skewed majority of honest gradients. To this end, STRIKE first searches for the skewed majority of honest gradients, and then constructs Byzantine gradients within the skewed majority by solving a constrained optimization problem. Empirical studies on three real-world datasets confirm the efficacy of our STRIKE attack.

Ethics and Broader Impact

The proposed skew-aware Byzantine attack STRIKE can present a threat to federated learning. Our goal with this work is thus to preempt these harms and encourage Byzantine defenses that are robust to skew-aware attacks in the future.

Reproducibility Statement

The implementation code is provided in Supplementary Materials. All datasets and the code platform (PyTorch) we use are public. Detail experiment setups are provided in the Appendix D.

References

Zeyuan Allen-Zhu, Faeze Ebrahimianghazani, Jerry Li, and Dan Alistarh. Byzantine-resilient nonconvex stochastic gradient descent. In *International Conference on Learning Representations*, 2020.

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. *arXiv preprint arXiv:2302.01772*, 2023.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Amine Boussetta, El Mahdi El Mhamdi, Rachid Guerraoui, Alexandre David Olivier Maurer, and Sébastien Louis Alexandre Rouault. Aksel: Fast byzantine sgd. In *Proceedings of the 24th International Conference on Principles of Distributed Systems (OPODIS 2020)*, number CONF. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2021.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

El Mahdi El-Mhamdi, Sadegh Farhadkhani, Rachid Guerraoui, Arsany Guirguis, Lê-Nguyên Hoang, and Sébastien Rouault. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in Neural Information Processing Systems*, 34:25044–25057, 2021.

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1605–1622, 2020.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6246–6283. PMLR, 17–23 Jul 2022.

Rachid Guerraoui, Sébastien Rouault, et al. The hidden vulnerability of distributed learning in byzantium. In *International Conference on Machine Learning*, pp. 3521–3530. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319. PMLR, 2021.

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`.

D. Knoke, G.W. Bohrnstedt, and A.P. Mee. *Statistics for Social Data Analysis*. F.E. Peacock Publishers, 2002. ISBN 9780875814483. URL `https://books.google.co.jp/books?id=1wNHAAAAMAAJ`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

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

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34, 2021b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

David S Moore, George P Mccabe, and Bruce A Craig. Introduction to the practice of statistics, 2009.

Krishna Pillutla, Sham M Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Virat Shejwalkar and Amir Houmansadr. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *Uncertainty in Artificial Intelligence*, pp. 261–270. PMLR, 2020.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261, 2019.

A VISUALIZATION OF GRADIENT SKEW

In order to gain insight into the gradient distribution, we use Locally Linear Embedding (LLE) [1] (Roweis & Saul, 2000) to visualize the gradients. From the visualization results, we observe that the distribution of gradient is skewed throughout FL training process when the data across different clients is non-IID. In this section, we first provide the detailed experimental setups of the observation experiments and then present the visualization results.

A.1 EXPERIMENTAL SETUPS

For CIFAR-10, we set the number of clients $n = 100$ and the Dirichlet concentration parameter $\beta = 0.1$. For ImageNet-12, we set the number of clients $n = 50$ and the Dirichlet concentration parameter $\beta = 0.1$. For FEMNIST, we adopt its natural data partition as introduced in Section 6.1. For all three datasets, we set the number of Byzantine clients $f = 0$. For CIFAR-10 and FEMNIST, we sample 100 clients to participate in training in each communication round. More visualized gradients would help us capture the characteristic of gradient distribution. For ImageNet-12, we sample 50 clients in each communication round. This is because we train ResNet-18 on ImageNet-12 and LLE on 100 gradients of ResNet-18 would be intractable due to the high dimensionality. Other setups align with Table 4.

For LLE, we set the number of neighbors to be $k = 0.1m$, where m is the number of sampled clients, to capture both local and global geometry of gradient distribution.

A.2 GRADIENT VISUALIZATION RESULTS

On each dataset, we run FedAvg for T communication round. Among the total T communication rounds, we randomly sample 6 rounds for visualization. For each round, we use LLE to visualize all the gradients and the optimal gradient (the average of all gradients) in this round. Please note that LLE is not linear. Therefore, the optimal gradient after the LLE may not be the average of all uploaded gradients after LLE. The visualization results are posted in Figure 4 below. In Figure 4, the majority of gradients skew away from the optimal gradient. These results imply that the gradient distribution is skewed during the entire training process.

B THEORETICAL ANALYSIS: EXPLOIT GRADIENT SKEW TO CIRCUMVENT BYZANTINE DEFENSES

We first recall all the definitions and assumptions for the integrity of this section.

Definition 1 ((f, λ)-resilient)**.** Given $f < n$ and $\lambda \geq 0$, an AGR $\mathcal{A}$ is (f, λ)-resilient if for any collection of n vectors $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$ and any set $\mathcal{G} \subseteq \{1, \ldots, n\}$ of size $n - f$,

$$\|\mathcal{A}(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n) - \bar{\boldsymbol{g}}_{\mathcal{G}}\| \leq \lambda \max_{i,j \in \mathcal{G}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|, \tag{20}$$

where $\bar{\boldsymbol{g}}_{\mathcal{G}} = \sum_{i \in \mathcal{G}} \boldsymbol{g}_i/(n - f)$ is the average of gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{G}\}$.

[1]Compared to LLE, t-SNE (Van der Maaten & Hinton, 2008) is a more popular visualization technique. Since t-SNE adjusts Gaussian bandwidth to locally normalize the density of data points, t-SNE can not capture the distance information of data. However, gradient skew relies heavily on distance information. Therefore, t-SNE is not appropriate for the visualization of gradient skew. In contrast, LLE can preserve the distance information of data distribution.

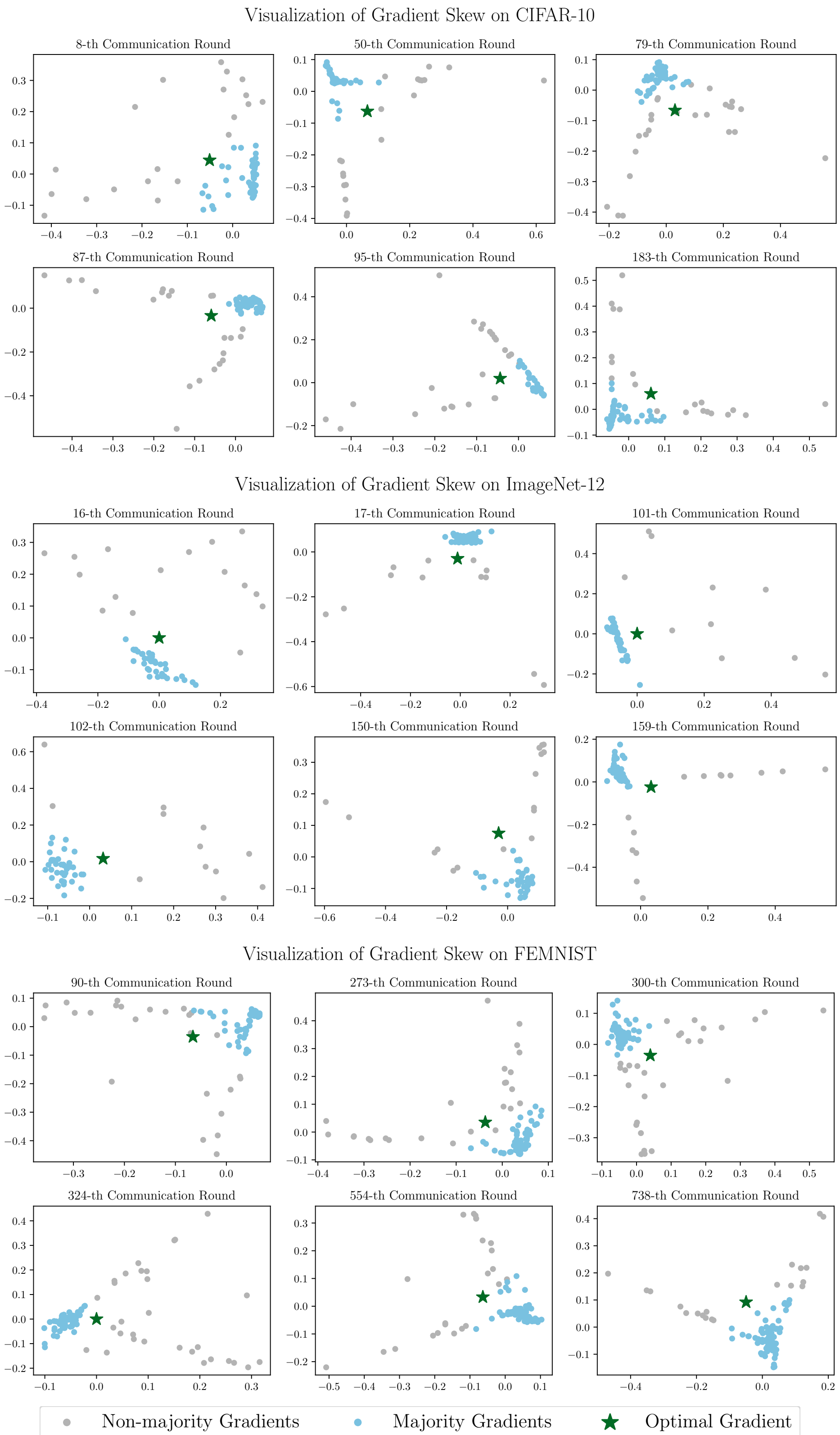


Figure 4: Visualization of gradient skew on three benchmark datasets.

Definition 2 ((f,γ)-skewed)**.** The set of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ is called (f,γ)-skewed if there exists a set $\mathcal{S} \subseteq \mathcal{H}$ of size $n-2f$ such that

$$\mathbb{E}[\|\bar{\boldsymbol{g}}_{\mathcal{S}} - \bar{\boldsymbol{g}}\|^2] \geq \gamma \rho_{\mathcal{S}}^2, \tag{21}$$

where $\bar{\boldsymbol{g}} = \sum_{i\in\mathcal{H}} \boldsymbol{g}_i/(n-f)$, $\bar{\boldsymbol{g}}_{\mathcal{S}} = \sum_{i\in\mathcal{S}} \boldsymbol{g}_i/(n-2f)$, and $\rho_{\mathcal{S}}^2 = \mathbb{E}[\max_{i,j\in\mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2]$ is a measure of gradient heterogeneity introduced by El-Mhamdi et al. (2021). Here, gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$ are called the *skewed majority* (of honest gradients), and γ is called the skewness of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$.

Assumption 1 (L-smooth)**.** The loss function is L-smooth, i.e.,

$$\|\nabla\mathcal{L}(\boldsymbol{w}) - \nabla\mathcal{L}(\boldsymbol{w}')\| \leq \|\boldsymbol{w} - \boldsymbol{w}'\|, \quad \forall \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d. \tag{22}$$

Assumption 2 (Unbias)**.** The stochastic gradients sampled from any local data distribution are unbiased estimators of local gradients for all clients, i.e.,

$$\mathbb{E}[\boldsymbol{g}_i^t] = \nabla\mathcal{L}(\boldsymbol{w}^t), \quad \forall i = 1, \ldots n, t = 0, \ldots, T-1. \tag{23}$$

B.1 Proofs

B.1.1 Supporting Lemma

We start with proving the lemma stated below.

Lemma 1. *Given any d-dimensional random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, the following inequalities hold:*

$$(\sqrt{\mathbb{E}[\|\boldsymbol{X}\|^2]} - \sqrt{\mathbb{E}[\|\boldsymbol{Y}\|^2]})^2 \leq \mathbb{E}[\|\boldsymbol{X}+\boldsymbol{Y}\|^2] \leq (\sqrt{\mathbb{E}[\|\boldsymbol{X}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{Y}\|^2]})^2 \tag{24}$$

Proof. $\mathbb{E}[\|\boldsymbol{X}+\boldsymbol{Y}\|^2]$ can be written as follows,

$$\mathbb{E}[\|\boldsymbol{X}+\boldsymbol{Y}\|^2] = \mathbb{E}[\|\boldsymbol{X}\|^2 + \|\boldsymbol{Y}\|^2 + 2\langle\boldsymbol{X},\boldsymbol{Y}\rangle] = \mathbb{E}[\|\boldsymbol{X}\|^2] + \mathbb{E}[\|\boldsymbol{Y}\|^2] + 2\mathbb{E}[\langle\boldsymbol{X},\boldsymbol{Y}\rangle]. \tag{25}$$

According to the Cauchy–Schwarz inequality, we have

$$|\mathbb{E}[\langle\boldsymbol{X},\boldsymbol{Y}\rangle]| \leq \mathbb{E}[|\langle\boldsymbol{X},\boldsymbol{Y}\rangle|] \leq \mathbb{E}[\|\boldsymbol{X}\|\|\boldsymbol{Y}\|] \leq \mathbb{E}[\|\boldsymbol{X}\|^2]\mathbb{E}[\|\boldsymbol{Y}\|^2]. \tag{26}$$

That is

$$-\mathbb{E}[\|\boldsymbol{X}\|^2]\mathbb{E}[\|\boldsymbol{Y}\|^2] \leq \mathbb{E}[\langle\boldsymbol{X},\boldsymbol{Y}\rangle] \leq \mathbb{E}[\|\boldsymbol{X}\|^2]\mathbb{E}[\|\boldsymbol{Y}\|^2]. \tag{27}$$

Combine Equation (25) and Inequality (27), then we have

$$\mathbb{E}[\|\boldsymbol{X}+\boldsymbol{Y}\|^2] \geq \mathbb{E}[\|\boldsymbol{X}\|^2] + \mathbb{E}[\|\boldsymbol{Y}\|^2] - 2\mathbb{E}[\|\boldsymbol{X}\|^2]\mathbb{E}[\|\boldsymbol{Y}\|^2] = (\sqrt{\mathbb{E}[\|\boldsymbol{X}\|^2]} - \sqrt{\mathbb{E}[\|\boldsymbol{Y}\|^2]})^2, \tag{28}$$

and

$$\mathbb{E}[\|\boldsymbol{X}+\boldsymbol{Y}\|^2] \leq \mathbb{E}[\|\boldsymbol{X}\|^2] + \mathbb{E}[\|\boldsymbol{Y}\|^2] + 2\mathbb{E}[\|\boldsymbol{X}\|^2]\mathbb{E}[\|\boldsymbol{Y}\|^2] = (\sqrt{\mathbb{E}[\|\boldsymbol{X}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{Y}\|^2]})^2. \tag{29}$$

□

B.1.2 Proof of Proposition 1

We recall the proposition statement below.

Proposition 1 (Vulnerability under skew)**.** *Given any (f,λ)-resilient AGR $\mathcal{A}$, if the set of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ is (f,γ)-skewed, then there exist Byzantine gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{B}\}$ such that*

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n) - \bar{\boldsymbol{g}}\|^2] \geq \Omega(\frac{\gamma}{\lambda^2}\cdot\frac{f^2}{(n-f)^2}\cdot\rho_{\mathcal{S}}^2). \tag{30}$$

where $\bar{\boldsymbol{g}} = \sum_{i\in\mathcal{H}} \boldsymbol{g}_i/(n-f)$ is the optimal gradient, $\rho_{\mathcal{S}}^2 = \mathbb{E}[\max_{i,j\in\mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2]$, $\mathcal{S}$ is the index set of the skewed majority.

Proof. According to Definition 2, there exists $\mathcal{S} \subseteq \mathcal{H}$ of size $n-2f$ and $\gamma > 1$ such that

$$\mathbb{E}[\|\bar{\boldsymbol{g}}_{\mathcal{S}} - \bar{\boldsymbol{g}}\|^2] = \gamma \rho_{\mathcal{S}}^2. \tag{31}$$

For all $i \in \mathcal{B}$, we set Byzantine gradient $\boldsymbol{g}_i = \bar{\boldsymbol{g}}_{\mathcal{S}}$. We then show that, under this attack, the aggregation error is lower-bounded as shown in Equation (8).

We consider the average and heterogeneity of the forged honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{S} \cup \mathcal{B}\}$.

The average is computed as follows.

$$\begin{aligned}
\bar{\boldsymbol{g}}_{\mathcal{B}\cup\mathcal{S}} &= \frac{1}{n-f} \sum_{i \in \mathcal{B}\cup\mathcal{S}} \boldsymbol{g}_i && (32)\\
&= \frac{1}{n-f} (\sum_{i \in \mathcal{B}} \boldsymbol{g}_i + \sum_{i \in \mathcal{S}} \boldsymbol{g}_i) && (33)\\
&= \frac{1}{n-f} (f \bar{\boldsymbol{g}}_{\mathcal{S}} + (n-2f) \bar{\boldsymbol{g}}_{\mathcal{S}}) && (34)\\
&= \bar{\boldsymbol{g}}_{\mathcal{S}}. && (35)
\end{aligned}$$

Then we consider the heterogeneity of gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{S} \cup \mathcal{B}\}$ $\rho_{\mathcal{S}\cup\mathcal{B}}$.

For all $b \in \mathcal{B}$ and $i \in \mathcal{S}$,

$$\begin{aligned}
\|\boldsymbol{g}_b - \boldsymbol{g}_i\|^2 &= \|\bar{\boldsymbol{g}}_{\mathcal{S}} - \boldsymbol{g}_i\|^2 && (36)\\
&= \|\frac{1}{n-2f} \sum_{j \in \mathcal{S}} \boldsymbol{g}_j - \boldsymbol{g}_i\|^2 && (37)\\
&= \|\frac{1}{n-2f} \sum_{j \in \mathcal{S}} (\boldsymbol{g}_j - \boldsymbol{g}_i)\|^2 && (38)\\
&\leq \frac{1}{n-2f} \sum_{j \in \mathcal{S}} \|\boldsymbol{g}_j - \boldsymbol{g}_i\|^2 && (39)\\
&\leq \max_{j \in \mathcal{S}} \|\boldsymbol{g}_j - \boldsymbol{g}_i\|^2 && (40)
\end{aligned}$$

where Inequality (39) comes from the Cauchy inequality.

Then for the heterogeneity of $\{\boldsymbol{g}_i \mid i \in \mathcal{S} \cup \mathcal{B}\}$, we have:

$$\begin{aligned}
\rho_{\mathcal{B}\cup\mathcal{S}}^2 =& \mathbb{E}[\max_{i,j \in \mathcal{B}\cup\mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2] && (41)\\
=& \mathbb{E}[\max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2] && (42)\\
=& \rho_{\mathcal{S}}^2. && (43)
\end{aligned}$$

For notation simplicity, we denote $\mathcal{A}(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$ by $\hat{\boldsymbol{g}}$. Then we can lower bound $\mathbb{E}[\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}\|^2]$ as follows

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}\|^2] &= \mathbb{E}[\|(\bar{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}}) - (\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}}\|))^2] && (44)\\
&= \mathbb{E}[\|(\bar{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{S}}) - (\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}}\|))^2] && (45)\\
&\geq (\sqrt{\mathbb{E}[\|\bar{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{S}}\|]^2} - \sqrt{\mathbb{E}[\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}}\|^2]})^2. && (46)
\end{aligned}$$

Here, Equation (45) is due to Equation (32), Inequality (46) relies on Lemma 1

We can lower bound term $\sqrt{\mathbb{E}[\|\bar{\boldsymbol{g}}-\bar{\boldsymbol{g}}_{\mathcal{S}}\|^2]}-\sqrt{\mathbb{E}[\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}}\|^2]}$ as follows.

$$\begin{aligned}
\sqrt{\mathbb{E}[\|\bar{\boldsymbol{g}}-\bar{\boldsymbol{g}}_{\mathcal{S}}\|^2]}-\sqrt{\mathbb{E}[\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}_{\mathcal{S}\cup\mathcal{B}}\|^2]} &\geq \sqrt{\gamma\cdot\rho_{\mathcal{S}}^2}-\sqrt{\lambda^2\rho_{\mathcal{S}\cup\mathcal{B}}^2} && (47)\\
&= \sqrt{\gamma\cdot\rho_{\mathcal{S}}^2}-\sqrt{\lambda^2\rho_{\mathcal{S}}^2} && (48)\\
&= (\frac{\sqrt{\gamma}}{\lambda}-1)\lambda\rho_{\mathcal{S}} && (49)\\
&\geq (\frac{\sqrt{\gamma}}{\lambda}-1)\frac{f}{n-f}\cdot\rho_{\mathcal{S}} && (50)\\
&= \Omega(\frac{\sqrt{\gamma}}{\lambda}\cdot\frac{f}{n-f}\cdot\rho_{\mathcal{S}}) && (51)
\end{aligned}$$

where Equation (47) results from Equation (31) and Equation (6), Equation (48) relies on Equation (43). In Inequality (50), we use the fact $\lambda \geq f/(n-f)$ from Farhadkhani et al. (2022).

We combine Inequality (46) and Equation (51) for the final conclusion in Equation (8):

$$\mathbb{E}[\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}\|^2]=\Omega(\frac{\gamma}{\lambda^2}\cdot\frac{f^2}{(n-f)^2}\cdot\rho_{\mathcal{S}}^2). \tag{52}$$

□

B.1.3 Proof for Proposition 2

We recall the proposition statement below.

Proposition 2. *Given any (f,λ)-resilient AGR $\mathcal{A}$, L-smooth global loss function $\mathcal{L}$, and learning rate $\eta \leq 1/L$, if honest gradients $\{\boldsymbol{g}_i^t \mid i\in\mathcal{H}\}$ are (f,γ)-skewed for all $t=0,\ldots,T-1$, then there exists Byzantine gradients $\{\boldsymbol{g}_b^t \mid b\in\mathcal{B}, t=0,\ldots,T-1\}$ such that the global model parameter is bounded away from the global optimum $\mathbf{w}^*$:*

$$\mathbb{E}[\|\boldsymbol{w}^t-\boldsymbol{w}^*\|^2]\geq\Omega(\eta^2(1-L\eta)^2\cdot\frac{\gamma}{\lambda^2}\cdot\frac{f^2}{(n-f)^2}\cdot\rho^2),\quad t=1,\ldots,T, \tag{53}$$

where $\mathbf{w}^t$ is the parameter of global model in the t-th communication round, $\mathbf{w}^*$ is the global optimum of global loss function $\mathcal{L}$, $\rho^2=\min_{t=0,\ldots,T-1}\mathbb{E}[\max_{i,j\in\mathcal{S}^t}\|\boldsymbol{g}_i^t-\boldsymbol{g}_j^t\|^2]$, and $\mathcal{S}^t$ is the index set of the skewed majority of honest gradients in t-th communication round.

Proof. According to Proposition 1, for all $t=0,\ldots,T-1$, there exist Byzantine gradients $\{\boldsymbol{g}_i^t \mid i\in\mathcal{B}\}$ such that

$$\mathbb{E}[\|\hat{\boldsymbol{g}}^t-\bar{\boldsymbol{g}}^t\|^2]\geq C\cdot\frac{\gamma}{\lambda^2}\cdot\frac{f^2}{(n-f)^2}\cdot(\rho^t)^2, \tag{54}$$

where C is a constant, and $(\rho^t)^2=\max_{i,j\in\mathcal{S}^t}\|\boldsymbol{g}_i^t-\boldsymbol{g}_j^t\|^2$, and $\mathcal{S}^t$ is the skewed majority of the honest gradients in t-th communication round. Let $\rho^2=\min_{t=1,\ldots,T-1}(\rho^t)^2$, then we have

$$\mathbb{E}[\|\hat{\boldsymbol{g}}^t-\bar{\boldsymbol{g}}^t\|^2]\geq C\cdot\frac{\gamma}{\lambda^2}\cdot\frac{f^2}{(n-f)^2}\cdot\rho^2, \tag{55}$$

We prove Equation (8) in the following two different cases.

Case 1. $\mathbb{E}[\|\boldsymbol{w}^t-\boldsymbol{w}^*\|^2]<C\eta^2\gamma f^2\rho^2/4\lambda^2(n-f)^2$.

Since $\boldsymbol{w}^{t+1}=\boldsymbol{w}^t-\eta\hat{\boldsymbol{g}}^t$, we can rewrite $\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^*\|^2$ as follows.

$$\begin{aligned}
\|\boldsymbol{w}^{t+1}-\boldsymbol{w}^*\|^2 &= \|(\boldsymbol{w}^t-\eta\hat{\boldsymbol{g}}^t)-\boldsymbol{w}^*\|^2 && (56)\\
&= \|(\nabla\mathcal{L}(\boldsymbol{w}^t)-\eta\hat{\boldsymbol{g}}^t)+(\boldsymbol{w}^t-\boldsymbol{w}^*-\eta\nabla\mathcal{L}(\boldsymbol{w}^t))\|^2 && (57)\\
&= \|(\nabla\mathcal{L}(\boldsymbol{w}^t)-\eta\hat{\boldsymbol{g}}^t)+(\boldsymbol{w}^t-\boldsymbol{w}^*-\eta(\nabla\mathcal{L}(\boldsymbol{w}^t)-\nabla\mathcal{L}(\boldsymbol{w}^*)))\|^2. && (58)
\end{aligned}$$

In Equation (58) we use the fact that $\nabla\mathcal{L}(\boldsymbol{w}^*)=\mathbf{0}$.

Combine Equation (58) and Lemma 1, we can lower bound $\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2]$ as follows,

$$\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2] = \|(\nabla \mathcal{L}(\boldsymbol{w}^t) - \eta \hat{\boldsymbol{g}}^t) + (\boldsymbol{w}^t - \boldsymbol{w}^* - \eta(\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*)))\|^2 \tag{59}$$

$$\geq (\eta \underbrace{\sqrt{\mathbb{E}[\|\nabla \mathcal{L}(\boldsymbol{w}^t) - \hat{\boldsymbol{g}}^t\|^2]}}_{A} - \underbrace{\sqrt{\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^* - \eta(\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*))\|^2]}}_{B})^2. \tag{60}$$

To obtain a further lower bound for Equation (60) amounts to give lower and upper bound for terms A and B, respectively.

To lower bound term A, again we use Lemma 1,

$$\mathbb{E}[\|\nabla \mathcal{L}(\boldsymbol{w}^t) - \hat{\boldsymbol{g}}^t\|^2] = \mathbb{E}[\|(\bar{\boldsymbol{g}}^t - \hat{\boldsymbol{g}}^t) + (\nabla \mathcal{L}(\boldsymbol{w}^t) - \bar{\boldsymbol{g}}^t)\|^2] \tag{61}$$

$$\geq (\sqrt{\mathbb{E}[\|\bar{\boldsymbol{g}}^t - \hat{\boldsymbol{g}}^t\|^2]} - \sqrt{\mathbb{E}[\|\nabla \mathcal{L}(\boldsymbol{w}^t) - \bar{\boldsymbol{g}}^t\|^2]})^2 \tag{62}$$

$$\geq (\sqrt{C} \cdot \frac{\sqrt{\gamma}}{\lambda} \cdot \frac{f}{n-f} \cdot \rho - \frac{\sigma}{\sqrt{n-f}})^2. \tag{63}$$

Here, $\sigma^2 = \sum_{i \in \mathcal{H}} \mathrm{Var}[\boldsymbol{g}_i^t]/(n-f)$ is the average variance of stochastic gradients. Inequality (63) is a combined result of Equation (55) and the law of large numbers.

We apply Lemma 1 to upper-bound term B as follows,

$$\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^* - (\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*))\|^2] \leq \mathbb{E}[(\|\boldsymbol{w}^t - \boldsymbol{w}^*\| + \eta \cdot L\|\boldsymbol{w}^t - \boldsymbol{w}^*\|)^2] \tag{64}$$

$$= (1 + L\eta)^2 \mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2] \tag{65}$$

$$\leq (1 + L\eta)^2 \cdot \frac{C}{4} \cdot \eta^2 \cdot \frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2 \tag{66}$$

Combine Inequality (63) and Inequality (66), we have

$$\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2] \geq (\eta(\sqrt{C} \cdot \frac{\sqrt{\gamma}}{\lambda} \cdot \frac{f}{n-f} \cdot \rho - \frac{\sigma}{\sqrt{n-f}}) - \frac{\eta(1+L\eta)}{2} \cdot \frac{\sqrt{C\gamma}}{\lambda} \cdot \frac{f}{n-f} \cdot \rho)^2 \tag{67}$$

$$= (\frac{\eta(1-L\eta)}{2} \cdot \frac{\sqrt{C\gamma}}{\lambda} \cdot \frac{f}{n-f} \cdot \rho - \frac{\sigma}{\sqrt{n-f}})^2 \tag{68}$$

$$= \Omega(\eta^2(1-L\eta)^2 \cdot \frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2) \tag{69}$$

Here Equation (69) uses the fact that SGD variance σ^2 is negligible with respect to the gradient heterogeneity ρ^2.

Case 2. $\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2] \geq C\eta^2\gamma f^2\rho^2/4\lambda^2(n-f)^2$. In this case, we let Byzantine gradients behave honestly such that $\hat{\boldsymbol{g}}^t = \bar{\boldsymbol{g}}^t$. Then $\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2]$ can be lower-bounded as follows.

$$\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2] = \mathbb{E}[\|(\boldsymbol{w}^t - \eta\bar{\boldsymbol{g}}^t) - \boldsymbol{w}^*\|^2] \tag{70}$$

$$= \mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^* - \eta(\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*)) - \eta(\bar{\boldsymbol{g}}^t - \nabla \mathcal{L}(\boldsymbol{w}^t))\|^2] \tag{71}$$

$$\geq (\sqrt{E[\|\boldsymbol{w}^t - \boldsymbol{w}^* - \eta(\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*))\|^2]} - \eta\sqrt{\mathbb{E}[\|\bar{\boldsymbol{g}}^t - \nabla \mathcal{L}(\boldsymbol{w}^t)\|^2]})^2. \tag{72}$$

In Equation (71) we use the fact that $\nabla \mathcal{L}(\boldsymbol{w}^*) = \mathbf{0}$, and Equation (72) comes from Lemma 1

We first lower-bound $\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^* - \eta(\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*))\|^2]$,

$$\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^* - \eta(\nabla \mathcal{L}(\boldsymbol{w}^t) - \nabla \mathcal{L}(\boldsymbol{w}^*))\|^2] \geq \mathbb{E}[(\|\boldsymbol{w}^t - \boldsymbol{w}^*\| - \eta \cdot L\|\boldsymbol{w}^t - \boldsymbol{w}^*\|)^2] \tag{73}$$

$$= (1 - L\eta)^2 \mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2] \tag{74}$$

$$\geq (1 - L\eta)^2 \cdot \frac{C}{4} \cdot \eta^2 \cdot \frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2 \tag{75}$$

Then we upper-bound $\mathbb{E}[\|\hat{\boldsymbol{g}}^t - \nabla\mathcal{L}(\boldsymbol{w}^t)\|^2]$

$$\mathbb{E}[\|\hat{\boldsymbol{g}}^t - \nabla\mathcal{L}(\boldsymbol{w}^t)\|^2] = \mathbb{E}[\|\bar{\boldsymbol{g}}^t - \nabla\mathcal{L}(\boldsymbol{w}^t)\|^2] \tag{76}$$

$$\leq \frac{\sigma^2}{n-f} \tag{77}$$

Here, $\sigma^2 = \sum_{i\in\mathcal{H}} \mathrm{Var}[\boldsymbol{g}_i^t]/(n-f)$ is the average variance of stochastic gradients.

Combining Equation (75) and Equation (77), we have

$$\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2] \geq (\frac{\eta(1-L\eta)}{2} \cdot \frac{\sqrt{C\gamma}}{\lambda} \cdot \frac{f}{n-f} \cdot \rho - \eta\frac{\sigma}{\sqrt{n-f}})^2 \tag{78}$$

$$= \Omega(\eta^2(1-L\eta)^2 \cdot \frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2) \tag{79}$$

Here Equation (79) uses the fact that SGD variance σ^2 is negligible with respect to the gradient heterogeneity ρ^2.

In both cases, we have

$$\mathbb{E}[\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^*\|^2] = \Omega(\eta^2(1-L\eta)^2 \cdot \frac{\gamma}{\lambda^2} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2), \quad t = 0, \ldots, T-1, \tag{80}$$

which completes the proof

□

B.2 Application to Other Definitions of Byzantine Resilience

In this section, we discuss how our analysis applies to other definitions of Byzantine resilience. In particular, we consider the definitions of Byzantine resilience in recent works of Karimireddy et al. (2022); Allouah et al. (2023).

B.2.1 Circumvent $(\delta_{\max}, c)$-AGRs

The following formulation of Byzantine resilience in Karimireddy et al. (2022) improves the upper bound by the fraction of Byzantine clients, and thus can recover the standard convergence rate when there are no Byzantine clients.

Definition 3 ($(\delta_{\max}, c)$-AGR)**.** A robust AGR gA is called a $(\delta_{\max}, c)$-AGR if, given any input $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$ of which a subset of at least size $|\mathcal{G}| > (1-\delta)n$ for $\delta \leq \delta_{\max} < 0.5$ and satisfies $\mathbb{E}[\|\boldsymbol{g}_i - \boldsymbol{g}_j\|] \leq \rho^2$, the output $\hat{\boldsymbol{g}}$ of AGR $\mathcal{A}$ satisfies:

$$\mathbb{E}[\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{G}}\|^2] \leq c\delta\rho^2 \quad \text{where} \quad \hat{\boldsymbol{g}} = \mathcal{A}_\delta(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n), \bar{\boldsymbol{g}}_{\mathcal{G}} = \sum_{i\in\mathcal{G}} \boldsymbol{g}_i/(n-f). \tag{81}$$

We show that any $(\delta_{\max}, c)$-AGR $\mathcal{A}$ also satisfies the resilience defined in Definition 1

Proposition 3. *Any $(\delta_{\max}, c)$-AGR $\mathcal{A}$ is (f, λ)-resilient for any $f \leq \delta_{\max}n$ and $\lambda = \sqrt{c\delta}$.*

Proof. Consider any deterministic vectors $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$, $f \leq \delta_{\max}n$, and subset $\mathcal{G} \subseteq \{1, \ldots, n\}$ of size $n - f$. According to Definition 3, we have

$$\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{G}}\|^2 \leq c\delta\rho^2 \tag{82}$$

where $\hat{\boldsymbol{g}} = \mathcal{A}_\delta(\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n)$, $\bar{\boldsymbol{g}}_{\mathcal{G}} = \sum_{i\in\mathcal{G}} \boldsymbol{g}_i/(n-f)$, $\delta = f/n$, and $\rho^2 \geq \max_{i,j\in\mathcal{G}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2$ The expectation is dropped since input vectors $\{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n\}$, $f \leq \delta_{\max}n$ are deterministic. We take $\rho^2 = \max_{i,j\in\mathcal{G}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|$ take the square root of both sides of Inequality (82), then we have

$$\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{G}}\| \leq \sqrt{c\delta} \max_{i,j\in\mathcal{G}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|. \tag{83}$$

Therefore, $\mathcal{A}$ is (f, λ)-resilient for any $f \leq \delta_{\max}n$ and $\lambda = \sqrt{cf/n}$. □

Combining Proposition 3 with Proposition 1 and Proposition 2, the following corollaries are obvious.

Corollary 1. *Given any $(\delta_{\max}, c)$-AGR $\mathcal{A}$ with $\delta_{\max} \geq f/n$, if the set of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ is (f,γ)-skewed, then there exist Byzantine gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{B}\}$ such that*

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n) - \bar{\boldsymbol{g}}\|^2] \geq \Omega(\frac{\gamma}{c} \cdot \frac{f}{n-f} \cdot \rho_{\mathcal{S}}^2). \tag{84}$$

where $\bar{\boldsymbol{g}} = \sum_{i\in\mathcal{H}} \boldsymbol{g}_i/(n-f)$ is the optimal gradient, $\rho_{\mathcal{S}}^2 = \mathbb{E}[\max_{i,j\in\mathcal{S}}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2]$, $\mathcal{S}$ is the index set of the skewed majority.

Corollary 2. *Given any $(\delta_{\max}, c)$-resilient AGR $\mathcal{A}$ with $\delta_{\max} \geq f/n$, L-smooth global loss function $\mathcal{L}$, and learning rate $\eta \leq 1/L$, if honest gradients $\{\boldsymbol{g}_i^t \mid i \in \mathcal{H}\}$ are (f,γ)-skewed for all $t = 0,\ldots,T-1$, then there exists Byzantine gradients $\{\boldsymbol{g}_b^t \mid b \in \mathcal{B}, t = 0,\ldots,T-1\}$ such that the global model parameter is bounded away from the global optimum $\boldsymbol{w}^*$:*

$$\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2] \geq \Omega(\eta^2(1-L\eta)^2 \cdot \frac{\gamma}{c} \cdot \frac{f}{n-f} \cdot \rho^2), \quad t = 1,\ldots,T, \tag{85}$$

where $\boldsymbol{w}^t$ is the parameter of global model in the t-th communication round, and $\boldsymbol{w}^*$ is the global optimum of global loss function $\mathcal{L}$.

B.2.2 Circumvent (f,κ)-robust AGRs

The following notion of Byzantine resilience in Allouah et al. (2023) is also a unified robustness criterion that is fine-grained to obtain tight convergence guarantees.

Definition 4 **((f,κ)-robust).** Let $f < n/2$ and $\kappa \geq 0$, a robust AGR gA is called (f,κ)-robust] if for any input $\{\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n\}$ and any set $\mathcal{G} \subseteq \mathcal{G}$ of size $n-f$, the output $\hat{\boldsymbol{g}}$ of AGR $\mathcal{A}$ satisfies:

$$\|\mathcal{A}(\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n) - \bar{\boldsymbol{g}}_{\mathcal{G}}\| \leq \frac{\kappa}{n-f}\sum_{i\in\mathcal{S}}\|\boldsymbol{g}_i - \bar{\boldsymbol{g}}_{\mathcal{G}}\|^2 \quad \text{where} \quad \bar{\boldsymbol{g}}_{\mathcal{G}} = \sum_{i\in\mathcal{G}} \boldsymbol{g}_i/(n-f). \tag{86}$$

We show that any (f,κ)-robust $\mathcal{A}$ also satisfies the resilience defined in Definition 1.

Proposition 4. *Any (f,κ)-robust AGR $\mathcal{A}$ is (f,λ)-resilient for $\lambda = \sqrt{\kappa}$.*

Proof. Given any deterministic vectors $\{\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n\}$ and subset $\mathcal{G} \subseteq \{1,\ldots,n\}$ of size $n-f$. According to Definition 4, we have

$$\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{G}}\|^2 \leq \frac{\kappa}{n-f}\sum_{i\in\mathcal{S}}\|\boldsymbol{g}_i - \bar{\boldsymbol{g}}_{\mathcal{G}}\|^2 \leq \frac{\kappa}{n-f}\sum_{i\in\mathcal{S}}\max_{j\in\mathcal{G}}\|\boldsymbol{g}_i - \boldsymbol{g}_j\| \leq \kappa \max_{i,j\in\mathcal{G}}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2 \tag{87}$$

We take take the square root of both sides of Inequality (87), then we have

$$\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}_{\mathcal{G}}\| \leq \sqrt{\kappa}\max_{i,j\in\mathcal{G}}\|\boldsymbol{g}_i - \boldsymbol{g}_j\| \tag{88}$$

Therefore, $\mathcal{A}$ is (f,λ)-resilient for $\lambda = \sqrt{\kappa}$. □

Combining Proposition 4 with Proposition 1 and Proposition 2, the following corollaries are obvious.

Corollary 3. *Given any (f,κ)-robust AGR $\mathcal{A}$, if the set of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{H}\}$ is (f,γ)-skewed, then there exist Byzantine gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{B}\}$ such that*

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{g}_1,\ldots,\boldsymbol{g}_n) - \bar{\boldsymbol{g}}\|^2] \geq \Omega(\frac{\gamma}{\kappa} \cdot \frac{f^2}{(n-f)^2} \cdot \rho_{\mathcal{S}}^2). \tag{89}$$

where $\bar{\boldsymbol{g}} = \sum_{i\in\mathcal{H}} \boldsymbol{g}_i/(n-f)$ is the optimal gradient, $\rho_{\mathcal{S}}^2 = \mathbb{E}[\max_{i,j\in\mathcal{S}}\|\boldsymbol{g}_i - \boldsymbol{g}_j\|^2]$, $\mathcal{S}$ is the index set of the skewed majority.

Corollary 4. *Given any (f,κ)-robust AGR $\mathcal{A}$, L-smooth global loss function $\mathcal{L}$, and learning rate $\eta \leq 1/L$, if honest gradients $\{\boldsymbol{g}_i^t \mid i \in \mathcal{H}\}$ are (f,γ)-skewed for all $t = 0,\ldots,T-1$, then there exists Byzantine gradients $\{\boldsymbol{g}_b^t \mid b \in \mathcal{B}, t = 0,\ldots,T-1\}$ such that the global model parameter is bounded away from the global optimum $\boldsymbol{w}^*$:*

$$\mathbb{E}[\|\boldsymbol{w}^t - \boldsymbol{w}^*\|^2] \geq \Omega(\eta^2(1-L\eta)^2 \cdot \frac{\gamma}{\kappa} \cdot \frac{f^2}{(n-f)^2} \cdot \rho^2), \quad t = 1,\ldots,T, \tag{90}$$

where $\boldsymbol{w}^t$ is the parameter of global model in the t-th communication round, $\boldsymbol{w}^*$ is the global optimum of global loss function $\mathcal{L}$, $\rho^2 = \min_{t=0,\ldots,T-1}\mathbb{E}[\max_{i,j\in\mathcal{S}^t}\|\boldsymbol{g}_i^t - \boldsymbol{g}_j^t\|^2]$, and $\mathcal{S}^t$ is the index set of the skewed majority of honest gradients in t-th communication round.

C Bisection Method to Solve Equation (19)

In this section, we present the bisection method used to solve Equation (19). We define $f(\cdot)$ as follows.

$$f(\alpha) = \max_{i \in \mathcal{S}} \|\bar{\boldsymbol{g}}_{\mathcal{S}} + \alpha \cdot \text{sign}(\bar{\boldsymbol{g}}_{\mathcal{S}}) \odot \boldsymbol{\sigma}_{\mathcal{S}} - \boldsymbol{g}_i\| - \max_{i,j \in \mathcal{S}} \|\boldsymbol{g}_i - \boldsymbol{g}_j\|, \quad \alpha \in [0, +\infty). \tag{91}$$

We can easily verify the following facts: 1. $f(0) \leq 0$, $f(\alpha) \to +\infty$ when $\alpha \to +\infty$; 2. $f(\cdot)$ is continuous; 3. $f(\cdot)$ has unique zero point in $[0, +\infty)$. Therefore, optimizing Equation (19) is equivalent to finding the zero point of $f(\cdot)$, which can be easily solved by bisection method in Algorithm 2.

Algorithm 2 Bisection method

Require: The skewed majority of honest gradients $\{\boldsymbol{g}_i \mid i \in \mathcal{S}\}$, tolerance $\varepsilon > 0$, max iteration $M > 0$

$\alpha_{\min} \leftarrow 0$
$\alpha_{\max} \leftarrow 1$
while $f(\alpha) < 0$ **do**
    $\alpha_{\max} \leftarrow 2\alpha_{\max}$
end while
$iter \leftarrow 0$
while $\alpha_{\max} - \alpha_{\min} > \varepsilon$ and $iter < M$ **do**
    $\alpha_{\text{mid}} \leftarrow (\alpha_{\max} + \alpha_{\min})/2$
    if $f(\alpha_{\text{mid}}) < 0$ **then**
        $\alpha_{\min} \leftarrow \alpha_{\text{mid}}$
    else
        $\alpha_{\max} \leftarrow \alpha_{\text{mid}}$
    end if
    $iter \leftarrow iter + 1$
end while
$\alpha \leftarrow (\alpha_{\max} + \alpha_{\min})/2$
return α

D Experimental Setups and Additional Experiments

D.1 Experimental Setups

D.1.1 Data Distribution

For CIFAR-10 Krizhevsky et al. (2009) and ImageNet-12, we use Dirichlet distribution to generate non-IID data by following Yurochkin et al. (2019); Li et al. (2021a). For each class c, we sample $\boldsymbol{q}_c \sim \text{Dir}_n(\beta)$ and allocate a $(\boldsymbol{q}_c)_i$ portion of training samples of class c to client i. Here, $\text{Dir}_n(\cdot)$ denotes the n-dimensional Dirichlet distribution, and $\beta > 0$ is a concentration parameter. We follow Li et al. (2021a) and set the number of clients $n = 50$ and the concentration parameter $\beta = 0.5$ as default.

For FEMNIST, the data is naturally partitioned into 3,597 clients based on the writer of the digit/character. Thus, the data distribution across different clients is naturally non-IID. For each client, we randomly sample a 0.9 portion of data as the training data and let the remaining 0.1 portion of data be the test data following Caldas et al. (2018).

D.1.2 Hyperparameter Setting of Baselines Attacks

The compared baseline attacks are: BitFlip (Allen-Zhu et al., 2020), LIE (Baruch et al., 2019), IPM (Xie et al., 2020), Min-Max (Shejwalkar & Houmansadr, 2021), Min-Sum (Shejwalkar & Houmansadr, 2021), and Mimic (Karimireddy et al., 2022). The hyperparameter setting of the above attacks is listed in the following table.

Table 2: The hyperparameter setting of six baseline attacks. N/A represents there is no hyperparameter required for this attack.

Attacks	Hyperparameters
BitFlip	N/A
LIE	$z = 1.5$
IPM	$\varepsilon = 0.1$
Min-Max	$\gamma_{\text{init}} = 10, \tau = 1 \times 10^{-5}$, ∇^p: coordinate-wise standard deviation
Min-Sum	$\gamma_{\text{init}} = 10, \tau = 1 \times 10^{-5}$, ∇^p: coordinate-wise standard deviation
Mimic	N/A

D.1.3 The Hyperparameter Setting of Evaluated Defenses

The performance of our attack is evaluated on seven recent robust defenses: Multi-Krum Blanchard et al. (2017), Median Yin et al. (2018), RFA Pillutla et al. (2019), Aksel Boussetta et al. (2021), CClipKarimireddy et al. (2021) DnC Shejwalkar & Houmansadr (2021), and RBTM El-Mhamdi et al. (2021). The hyperparameter setting of the above defenses is listed in the following table. we

Table 3: The hyperparameter setting of seven evaluated defenses. N/A represents there is no hyperparameter required for this defense.

Defenses	Hyperparameters
Multi-Krum	N/A
Median	N/A
RFA	$T = 8$
Aksel	N/A
CClip	$L = 1, \tau = 10$
DnC	$c = 1, \text{niters} = 1, b = 1000$
RBTM	N/A

also consider a simple yet effective bucketing scheme (Karimireddy et al., 2022) that adapts existing defenses to the non-IID setting. We follow the original paper and set the bucket size to be $s = 2$.

D.1.4 Evaluation

We use top-1 accuracy, i.e., the proportion of correctly predicted testing samples to total testing samples, to evaluate the performance of global models. The *lower* the accuracy, the more effective the attack. We run each experiment five times and report the mean and standard deviation of the highest accuracy during the training process.

D.1.5 Compute

All experiments are run on the same machine with Intel E5-2665 CPU, 32GB RAM, and four GeForce GTX 1080Ti GPU.

D.1.6 Other Setups

The number of Byzantine clients of all datasets is set to $f = 0.2 \cdot n$. We test STRIKE with $\nu \in \{0.25 \cdot i \mid i = 1, \ldots, 8\}$ and report the lowest test accuracy (highest attack effectiveness).

The hyperparameter setting for datasets FEMNIST (Caldas et al., 2018), CIFAR-10 (Krizhevsky et al., 2009) and ImageNet-12 (Russakovsky et al., 2015) are listed in below Table 4.

Table 4: Hyperparameter setting for FEMNIST, CIFAR-10 and ImageNet-12. # is the number sign. For example, # Communication rounds represents the number of communication rounds.

Dataset	FEMNIST	CIFAR-10	ImageNet-12
Architecture	CNN (Caldas et al., 2018)	AlexNet (Krizhevsky et al., 2017)	ResNet-18 (He et al., 2016)
# Communication rounds	800	200	200
# Sampled Clients	10	50	50
# Local epochs	1	1	1
Optimizer	SGD	SGD	SGD
Batch size	128	128	128
Learning rate	0.5	0.1	0.1
Momentum	0.5	0.9	0.9
Weight decay	0.0001	0.0001	0.0001
Gradient clipping	Yes	Yes	Yes
Clipping norm	2	2	2

D.2 Additional Experiments

D.2.1 Performance under Varying Hyperparameter ν

We study the influence of ν on ImageNet-12 dataset. We report the test accuracy under STRIKE attack with ν in $\{0.25{*}i \mid i = 1, \dots, 8\}$ against seven different defenses on ImageNet-12 in Figure 5. We also report the lowest test accuracy (best performance) of six baseline attacks introduced in Section 6.1 as a reference. Please note that a *lower* accuracy implies higher attack effectiveness.

As shown in the Figure 5, the performance of STRIKE is generally competitive with varying ν. In most cases, simply setting $\nu = 1$ can beat other attacks (except for CClip, yet we observe that the performance is low enough to make the model useless). The impact of ν value is different for different robust AGRs: for Median and RFA, the accuracy is relatively stable under different νs; for CClip and Multi-Krum, the accuracy is lower with larger νs; for Aksel and DnC, the accuracy first decreases and then increases as ν increases.

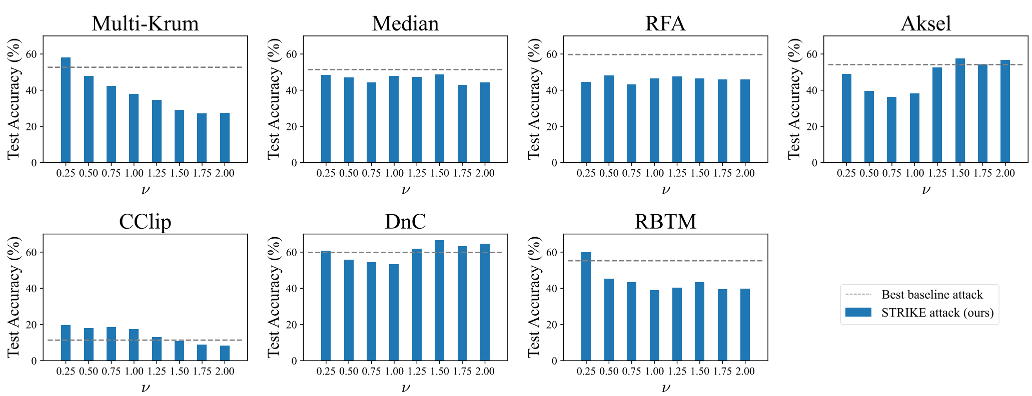


Figure 5: Accuracy under STRIKE attack with ν in $\{0.25 * i \mid i = 1, \dots, 8\}$ against seven different defenses on ImageNet-12. The gray dashed line in each figure represents the lowest test accuracy (best performance) of six baseline attacks introduced in Section 6.1. We include it as a reference. The *lower* the accuracy, the more effective the attack. Other experimental setups align with the main experiment as introduced in Section 6.1.

D.2.2 PERFORMANCE UNDER DIFFERENT NON-IID LEVELS

As shown in Table 1, DnC demonstrates the strongest robustness against various attacks on all datasets. Therefore, we fix the defense to be DnC in this experiment. As discussed in Appendix D.2.1, simply setting $\nu = 1$ yields satisfactory performance of our STRIKE attack. Thus, we fix $\nu = 1$ in this experiment. We vary Dirichlet concentration parameter β in $\{0.1, 0.2, 0.5, 0.7, 0.9\}$ to study how our attack behaves under different non-IID levels. Lower β implies a higher non-IID level. We additionally test the performance in the IID setting. Other setups align with the main experiment as introduced in Section 6.1. The results are posted in Figure 6 below.

As shown in Figure 6, the accuracy generally increases as β decreases for all attacks. The accuracy under our STRIKE attack is consistently lower than all the baseline attacks. Besides, we also note that the accuracy gap between our STRIKE attack and other baseline attacks gets smaller when the non-IID level decreases. We hypothesize the reason is that gradient skew is milder as the non-IID level decreases, which aligns with our theoretical results in Propositions 1 and 2. Even in the IID setting, our STRIKE attack is competitive compared to other baselines.

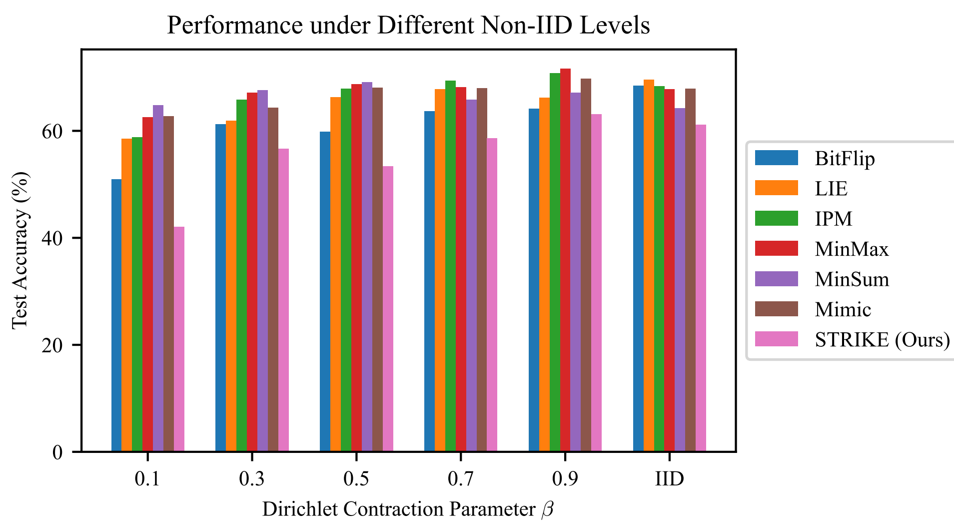


Figure 6: Accuracy under different attacks against DnC under different non-IID levels on ImageNet12. Lower β implies a higher non-IID level. "IID" implies that the data is IID distributed. The *lower*, the better. Other setups align with the main experiment as introduced in Section 6.1.

D.2.3 PERFORMANCE UNDER DIFFERENT BYZANTINE CLIENT RATIO

As shown in Table 1, DnC demonstrates the strongest robustness against various attacks on all datasets. Therefore, we fix the defense to be DnC in this experiment. As discussed in Appendix D.2.1, simply setting $\nu = 1$ yields satisfactory performance of our STRIKE attack. Thus, we fix $\nu = 1$ in this experiment. We vary the number of Byzantine clients f in $\{5, 10, 15, 20\}$ and fix the total number of clients n to be 50. In this way, Byzantine client ratio f/n varies in $\{0.1, 0.2, 0.3, 0.4\}$ to study how our attack behaves under different Byzantine client ratio. Other setups align with the main experiment as introduced in Section 6.1. The results are posted in Figure 7 below.

As shown in Figure 7, the accuracy generally decreases as f/n increases for all attacks. The accuracy under our STRIKE attack is consistently lower than all the baseline attacks. The results suggest that all attacks are more effective when there are more Byzantine clients. Meanwhile, our attack is the most effective under different Byzantine client number.

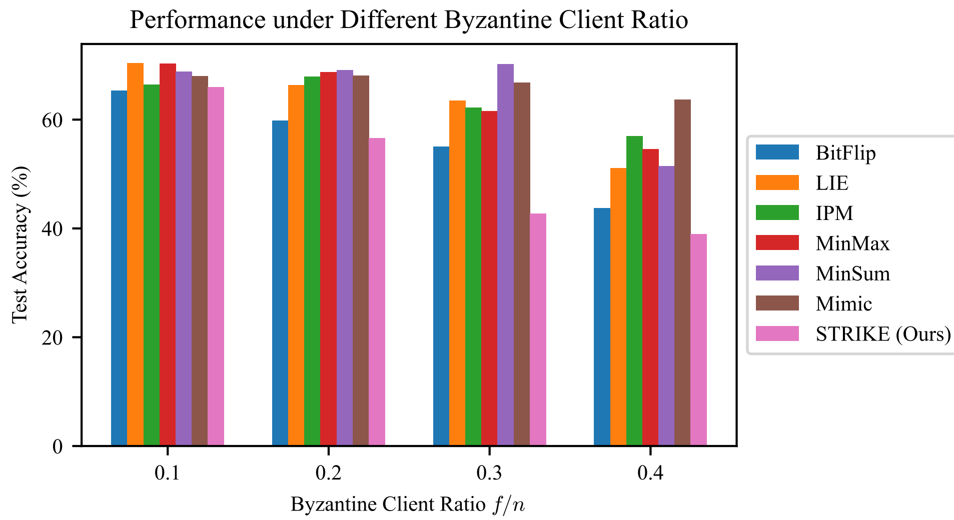


Figure 7: Accuracy under different attacks against DnC under different Byzantine client ratio on ImageNet12. The *lower*, the better. Other setups align with the main experiment as introduced in Section 6.1.

D.2.4 Performance under Different Client Number

As shown in Table 1, DnC demonstrates the strongest robustness against various attacks on all datasets. Therefore, we fix the defense to be DnC in this experiment. As discussed in Appendix D.2.1, simply setting $\nu = 1$ yields satisfactory performance of our STRIKE attack. Thus, we fix $\nu = 1$ in this experiment. We vary the number of total clients n in $\{10, 30, 50, 70, 90\}$ and set the number of Byzantine clients $f = 0.2n$ accordingly. In this way, We can study how our attack behaves under different client number. Other setups align with the main experiment as introduced in Section 6.1. The results are posted in Figure 8 below.

As shown in Figure 8, the accuracy generally decreases as client number n increases for all attacks. The accuracy under our STRIKE attack is consistently lower than all the baseline attacks under different client number.

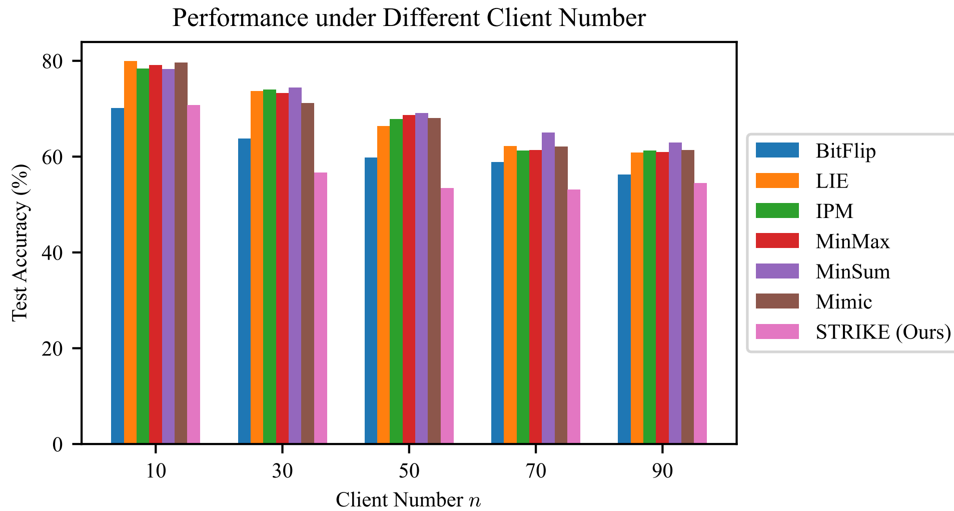


Figure 8: Accuracy under different attacks against DnC under different client number on ImageNet12. The *lower*, the better. Other setups align with the main experiment as introduced in Section 6.1.