Byzantine-Robust Learning on Heterogeneous Data via Gradient Splitting

Yuchen Liu [1 * †] **Chen Chen** [2 *] **Lingjuan Lyu** [2] **Fangzhao Wu** [3] **Sai Wu** [1] **Gang Chen** [1]

Abstract

Federated learning has exhibited vulnerabilities to Byzantine attacks, where the Byzantine attackers can send arbitrary gradients to a central server to destroy the convergence and performance of the global model. A wealth of robust AGgregation Rules (AGRs) have been proposed to defend against Byzantine attacks. However, Byzantine clients can still circumvent robust AGRs when data is non-Identically and Independently Distributed (non-IID). In this paper, we first reveal the root causes of performance degradation of current robust AGRs in non-IID settings: the curse of dimensionality and gradient heterogeneity. In order to address this issue, we propose GAS, a GrAdient Splitting approach that can successfully adapt existing robust AGRs to non-IID settings. We also provide a detailed convergence analysis when the existing robust AGRs are combined with GAS. Experiments on various real-world datasets verify the efficacy of our proposed GAS. The implementation code is provided in https://github.com/YuchenLiu-a/byzantine-gas.

1. Introduction

Federated Learning (FL) (McMahan et al., 2017; Lyu et al., 2020; Zhao et al., 2020) provides a privacy-aware and distributed machine learning paradigm. It has recently attracted widespread attention as a result of emerging data silos and growing privacy awareness. In this paradigm, data owners (clients) repeatedly use their private data to compute local gradients and send them to a central server for aggregation. In this way, clients can collaborate to train a model without exposing their private data. However, the distributed property of FL also makes it vulnerable to Byzantine attacks (Blanchard et al., 2017; Guerraoui et al., 2018; Chen et al., 2020), in which Byzantine clients can send arbitrary messages to the central server to bias the global model. Moreover, it is challenging for the server to identify the Byzantine clients, since the server can neither access clients' training data nor monitor their local training process.

In order to defend against Byzantine attacks, the community has proposed a wealth of defenses (Blanchard et al., 2017; Guerraoui et al., 2018; Yin et al., 2018). Most defenses abandon the averaging step adopted by conventional FL frameworks, e.g., FedAvg (McMahan et al., 2017). Instead, they use robust AGgregation Rules (AGRs) to aggregate local gradients and compute the global gradient. Most existing robust AGRs assume that the data distribution on different clients is Identically and Independently Distributed (IID) (Bernstein et al., 2018; Ghosh et al., 2019). In fact, the data is usually heterogeneous, i.e., non-IID, in real-world FL applications (McMahan et al., 2017; Kairouz et al., 2021; Lyu et al., 2022; Zhang et al., 2023; Chen et al., 2022a). In this paper, we focus on defending against Byzantine attacks in the more realistic non-IID setting.

In the non-IID setting, defending against Byzantine attacks becomes more challenging (Karimireddy et al., 2022; Acharya et al., 2022). Robust AGRs that try to include *all* the honest gradients in aggregation (Blanchard et al., 2017; Shejwalkar & Houmansadr, 2021) fail to handle the curse of dimensionality (Guerraoui et al., 2018). Byzantine clients can take advantage of the high dimension of gradients and participate in aggregation. As a result, the global gradient is manipulated away from the optimal gradient, i.e., the average of honest gradients. Other robust AGRs (Guerraoui et al., 2018; Yin et al., 2018) aggregate *fewer* gradients to ensure that only honest gradients participate in aggregation. However, the global gradient is still of limited utility due to gradient heterogeneity (Li et al., 2020; Karimireddy et al., 2020) in the non-IID setting. In summary, most existing AGRs fail to address both the curse of dimensionality (Guerraoui et al., 2018) and gradient heterogeneity (Karimireddy et al., 2022) at the same time. Consequently, they fail to achieve satisfactory performance in the non-IID setting.

Motivated by the above observations, we propose a GrAdient Splitting based approach called GAS for Byzantine robustness in non-IID settings. In particular, to ad-

[*]Equal contribution [†]Work partly done during internship at Sony AI [1]Key Lab of Intelligent Computing Based Big Data of Zhejiang Province, Zhejiang University, Hangzhou, China [2]Sony AI [3]Microsoft. Correspondence to: Lingjuan Lyu <lingjuan.lv@sony.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

dress the curse of dimensionality, GAS splits each high-dimensional gradient into low-dimensional sub-vectors and detects Byzantine gradients with the sub-vectors. To handle the gradient heterogeneity, GAS aggregates all the identified honest gradients.

Our contributions in this work are summarized below.

- We reveal the root causes of defending against Byzantine attacks in the non-IID setting: the gradient heterogeneity and the curse of dimensionality. Gradient heterogeneity makes it hard for Byzantine defenses to obtain a global gradient close to the optimal. The curse of dimensionality enables the Byzantine gradients to circumvent defenses that aggregate more gradients. To the best of our knowledge, no existing defense can address both issues at the same time.
- We propose a novel and compatible approach called GAS which consists of three steps: 1. splitting the high-dimensional gradients into low-dimensional sub-vectors; 2. penalizing each gradient by a score with a robust AGR based on their split low-dimensional sub-vectors to circumvent the curse of dimensionality; 3. identifying the gradients with low scores as honest ones and aggregating all the identified honest gradients to tackle the gradient heterogeneity issue. In step 2, GAS can apply any robust AGR to low-dimensional sub-vectors for identification, offering great compatibility.
- We provide convergence analysis for our proposed GAS. Extensive experiments on four real-world datasets across various non-IID settings empirically validate the effectiveness and superiority of our GAS.

2. Related Works

IID defenses. Blanchard et al. (2017) first introduce Byzantine robust learning and propose a distance-based AGR called Multi-Krum. Yin et al. (2018) theoretically analyze the statistical optimality of Median and Trimmed Mean. Guerraoui et al. (2018) propsoe Bulyan that applies a variant of Trimmed Mean as a post-processing method to handle the curse of dimensionality. Pillutla et al. (2019) discuss the Byzantine robustness of Geometric Median and propose a computationally efficient approximation of Geometric Median. Shejwalkar & Houmansadr (2021) propose to perform dimensionality reduction using random sampling, followed by spectral-based outlier removal. These defenses assume the data is IID. Their efficacy is therefore limited in more realistic FL applications where the data is non-IID.

Non-IID defenses. Recent works have also explored defenses applicable to the non-IID setting. Park et al. (2021) can only achieve Byzantine robustness when the server has a validation set, which compromises the privacy principle of the FL (McMahan et al., 2017). Data & Diggavi (2021) adapt a robust mean estimation approach to FL in order to combat the Byzantine attack in the non-IID setting. However, it requires $\Omega(d^2)$ time (d is the number of model parameters), which is unacceptable due to the high dimensionality of model parameters. El-Mhamdi et al. (2021) consider Byzantine robustness in the asynchronous communication and unconstrained topologies settings. Acharya et al. (2022) propose to apply geometric median only to the sparsified gradients to save computation cost. Karimireddy et al. (2022) perform a bucketing process before aggregation to reduce the gradient heterogeneity. However, most of these methods ignore the curse of dimensionality (Guerraoui et al., 2018), which becomes intractable in the non-IID setting (refer to Section 4 for more discussion). As a result, they fail to achieve satisfactory performance in the non-IID setting.

3. Notations and Preliminaries

Notations. For any positive integer $n \in \mathbb{N}^+$, we denote the set $\{1, \ldots, n\}$ by $[n]$. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. We denote the ℓ_2 norm of vector $\boldsymbol{x}$ by $\|\boldsymbol{x}\|$. We use $[\boldsymbol{x}]_j$ to represent the j-th component of vector $\boldsymbol{x}$. The sub-vector of vector $\boldsymbol{x}$ indexed by index set $\mathcal{J}$ is denoted by $[\boldsymbol{x}]_{\mathcal{J}} = ([\boldsymbol{x}]_{j_1}, \ldots, [\boldsymbol{x}]_{j_k})$, where $\mathcal{J} = \{j_1, \ldots, j_k\}$, and $k = |\mathcal{J}|$ is the number of indices. For a random variable X, we use $\mathbb{E}[X]$ and $\mathrm{Var}[X]$ to denote the expectation and variance of X, respectively.

Federated learning. We consider the Federated Learning (FL) system with a central server and n clients following (Blanchard et al., 2017; Yin et al., 2018; Chen et al., 2022b). Then the objective is to minimize loss $\mathcal{L}(\boldsymbol{w})$ defined as follows.

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

$$\text{where } \mathcal{L}_i(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\xi}_i}[\mathcal{L}(\boldsymbol{w}; \boldsymbol{\xi}_i)], i \in [n], \tag{2}$$

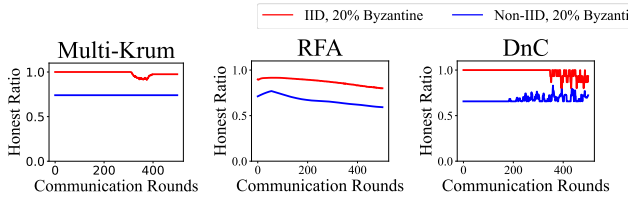
where $\boldsymbol{w}$ is the model parameter, $\mathcal{L}_i$ is the loss function on the i-th client, $\boldsymbol{\xi}_i$ is the data distribution on the i-th client, and $\mathcal{L}(\boldsymbol{w}; \boldsymbol{\xi})$ is the loss function.

In the t-th communication round, the server distributes the parameter $\boldsymbol{w}^t$ to the clients. Each client i conducts several epochs of local training on local data to obtain the updated local parameter $\boldsymbol{w}_i^t$. Then, client i computes the local gradient $\boldsymbol{g}_i^t$ as follows and sends it to the server.

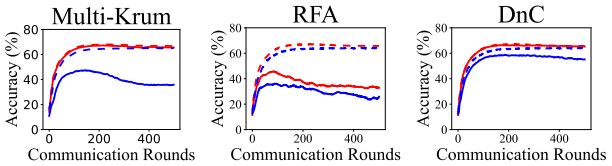
$$\boldsymbol{g}_i^t = \boldsymbol{w}^t - \boldsymbol{w}_i^t. \tag{3}$$

Finally, the server collects the local gradients and uses the average gradient to update the global model.

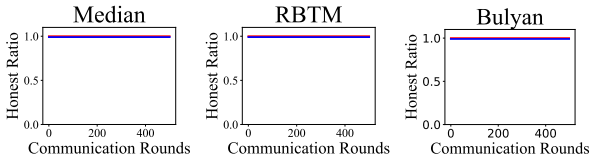
$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \boldsymbol{g}^t, \quad \boldsymbol{g}^t = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^t. \tag{4}$$



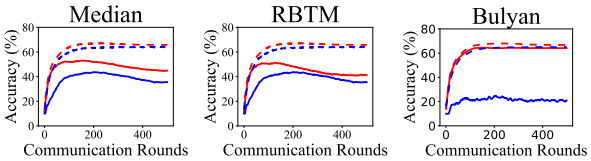
(a) The ratio of honest gradients that participate in the aggregation for robust AGRs that try to include all honest gradients (Multi-Krum (Blanchard et al., 2017), RFA (Pillutla et al., 2019), DnC (Shejwalkar & Houmansadr, 2021)).



(b) The test accuracy of robust AGRs that try to include all honest gradients (Multi-Krum (Blanchard et al., 2017), RFA (Pillutla et al., 2019), DnC (Shejwalkar & Houmansadr, 2021)).



(c) The ratio of honest gradients that participate in the aggregation of robust AGRs that aggregate fewer gradients (Median (Yin et al., 2018), RBTM (El-Mhamdi et al., 2021), Bulyan (Guerraoui et al., 2018)).



(d) The test accuracy of robust AGRs that aggregate few gradients (Median (Yin et al., 2018), RBTM (El-Mhamdi et al., 2021), Bulyan (Guerraoui et al., 2018)).

Figure 1: The experiments are conducted under the attack of 20% Byzantine clients on CIFAR-10 (Krizhevsky et al., 2009) dataset in both IID and non-IID settings. More detailed setups are covered in Appendix A.

The process is repeated until the number of communication rounds reaches the set value T.

Byzantine threat model. In real-world applications, not all clients in FL systems are honest. In other words, there may exist Byzantine clients in FL systems (Blanchard et al., 2017). Suppose that among total n clients, f clients are Byzantine. Let $\mathcal{B} \subseteq [n]$ denote set of Byzantine clients and $\mathcal{H} = [n] \setminus \mathcal{B}$ denote the set of honest clients. In the presence of Byzantine clients, the uploaded message of client i in the t-th communication round is

$$\boldsymbol{g}_i^t = \begin{cases} \boldsymbol{w}^t - \boldsymbol{w}_i^{t+1}, & i \in \mathcal{H}, \\ *, & i \in \mathcal{B}, \end{cases} \tag{5}$$

where $*$ represents an arbitrary value.

Robust AGRs. Most existing Byzantine defenses replace the averaging step with a robust AGR to defend against Byzantine attacks. More specifically, the server aggregates the gradients and updates the global model as follows.

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \hat{\boldsymbol{g}}^t, \quad \hat{\boldsymbol{g}}^t = \mathcal{A}(\boldsymbol{g}_1^t, \ldots, \boldsymbol{g}_n^t), \tag{6}$$

where $\hat{\boldsymbol{g}}^t$ is the aggregated gradient, and $\mathcal{A}$ is a robust AGR, e.g., Multi-Krum (Blanchard et al., 2017) and Bulyan (Guerraoui et al., 2018).

For notation simplicity, we omit the superscript t of the gradient symbols when there is no ambiguity in the rest of this paper.

4. The Challenges of Byzantine Robustness in Non-IID Setting

Most robust AGRs focus on Byzantine robustness in the IID setting (Blanchard et al., 2017; Guerraoui et al., 2018). When the data is non-IID (Kairouz et al., 2021; Zhang et al., 2022), the performance of these robust AGRs drops drastically (Shejwalkar & Houmansadr, 2021; Karimireddy et al., 2022). In order to understand the root cause of this performance drop, we perform an experimental study on various robust AGRs. Particularly, we examine their behaviors under the attack of 20% Byzantines in both IID and non-IID settings on CIFAR-10 (Krizhevsky et al., 2009) in Figure 1. More detailed setups are covered in Appendix A.

Some robust AGRs try to include *all* honest gradients in aggregation (the number of aggregated gradients is no less than $n - f$, i.e., the number of honest clients) (Blanchard et al., 2017; Shejwalkar & Houmansadr, 2021; Pillutla et al., 2019). However, they fail to address *the curse of dimensionality* (Guerraoui et al., 2018) on heterogeneous data. Byzantine clients can take advantage of the high dimension of gradients and easily circumvent these defenses. As shown in Figure 1a, these defenses include significantly more Byzantine gradients in aggregation in the non-IID setting. As a result, the global gradient is manipulated away from the optimal gradient, which leads to an ineffectual global model in the non-IID setting as shown in Figure 1b.

Other robust AGRs aggregate fewer gradients (less than $n - f$) to get rid of Byzantine clients (Guerraoui et al., 2018; Yin et al., 2018; El-Mhamdi et al., 2021). The results in Fig-

ure 1c imply that they can exclude Byzantine clients from the aggregation in both IID and non-IID settings. However, their performance still degrades in the non-IID setting as shown in Figure 1d. In fact, this degradation comes from *gradient heterogeneity* (Li et al., 2020; Karimireddy et al., 2020) in the non-IID setting. As a price for removing Byzantine gradients, these robust AGRs exclude a proportion of honest gradients from the aggregation. Since the honest gradients are heterogeneous, such exclusion causes the aggregated gradient to deviate far from the optimal gradient, i.e., the average of honest gradients. The deviation further leads to an ineffectual global model. Therefore, they fail to achieve satisfactory performance in the non-IID setting.

In summary, no existing robust AGR is capable of handling both the curse of dimensionality and gradient heterogeneity at the same time. A new strategy is needed to tackle both challenges in the non-IID setting.

5. Gradient Splitting Based Approach

Our observations in Section 4 clearly motivate the need for a more robust defense to tackle both the curse of dimensionality and gradient heterogeneity to defeat Byzantine attacks in the non-IID setting. Inspired by these observations, we propose a novel GrAdient Splitting based approach called GAS, which consists of three steps as follows.

Splitting. First, GAS splits the gradients to mitigate the curse of dimensionality for the next identification step. The splitting is specified by a partition of set $[d]$, where d is the dimension of gradients. In particular, we randomly partition $[d]$ into p subsets, with each subset having no more than $\lceil d/p \rceil$ dimensions. Let $\{\mathcal{J}_1, \dots \mathcal{J}_p\}$ denote the partition. Each gradient $\boldsymbol{g}_i$ is correspondingly split into p sub-vectors as follows.

$$\boldsymbol{g}_i^{(q)} = [\boldsymbol{g}_i]_{\mathcal{J}_q}, \quad i \in [n], q \in [p], \tag{7}$$

where $\boldsymbol{g}_i^{(q)}$ is the q-th sub-vector of gradient $\boldsymbol{g}_i$.

Identification. Then, GAS applies robust AGR $\mathcal{A}$ to each group of sub-vectors corresponding to $\mathcal{J}_q$:

$$\hat{\boldsymbol{g}}^{(q)} = \mathcal{A}(\boldsymbol{g}_1^{(q)}, \dots, \boldsymbol{g}_n^{(q)}), \quad q \in [p], \tag{8}$$

where $\hat{\boldsymbol{g}}^{(q)}$ is the aggregation result of group q. By performing aggregation on each group of low-dimensional sub-vectors separately, GAS can circumvent the curse of dimensionality and get rid of Byzantine gradients.

Note that $\hat{\boldsymbol{g}}^{(q)}$ may still deviate from the optimal gradient due to the gradient heterogeneity (Karimireddy et al., 2022) as illustrated in Section 4. Therefore, it is *inappropriate* to directly use the aggregation results $\{\hat{\boldsymbol{g}}^{(q)}, q \in [p]\}$ as the final output. Instead, we use $\hat{\boldsymbol{g}}^{(q)}$ as an honest reference to compute identification scores for each client as follows.

$$s_i^{(q)} = \|\boldsymbol{g}_i^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|, \quad i \in [n], q \in [p]. \tag{9}$$

Since the group-wise aggregation result $\hat{\boldsymbol{g}}^{(q)}$ can get rid of Byzantine gradients, the identification score $s_i^{(q)}$ can provably characterize the potential for the $\boldsymbol{g}_i^{(q)}$ being a sub-vector of a Byzantine gradient.

Then, GAS collects the identification scores from all groups and computes the final aggregation result. In particular, the final identification score s_i of each client is composed of its identification scores received from all groups as follows.

$$s_i = \sum_{q=1}^{p} s_i^{(q)}, \quad i \in [n]. \tag{10}$$

Aggregation. To handle the gradient heterogeneity issue, GAS selects total $n - f$ gradients with the lowest identification scores for aggregation. Let $\mathcal{I}$ denote the index set of selected gradients, where $|\mathcal{I}| = n - f$. Then the average of selected gradients is output as the final aggregation result as follows:

$$\hat{\boldsymbol{g}} = \frac{1}{n-f} \sum_{i \in \mathcal{I}} \boldsymbol{g}_i. \tag{11}$$

Note that in the second step (Identification) of GAS, $\mathcal{A}$ could be any (f, λ)-resilient AGR (Definition 1). The key difference lies in that all the existing robust AGRs (Multi-Krum, Bulyan, etc.) directly operate on the original gradients; instead, we propose to apply robust AGRs on the split gradients, followed by identification before aggregation. In this way, we can help enhance the ability to handle both the curse of dimensionality and gradient heterogeneity of the current robust AGRs that satisfy the (f, λ)-resilient property (Definition 1) in the non-IID setting. We also analyze the computation cost of our proposed GAS in Appendix B.

Moreover, our GAS is a compatible approach that can be combined with most existing robust AGRs, e.g., Multi-Krum (Blanchard et al., 2017), Bulyan (Guerraoui et al., 2018).

6. Theoretical Analysis

In this section, we provide a convergence analysis for our GAS approach.

We analyze a popular FL model widely considered by Karimireddy et al. (2021; 2022); Acharya et al. (2022). In particular, each local gradient is computed by SGD as follows.

$$\boldsymbol{g}_i^t = \eta \nabla \mathcal{L}(\boldsymbol{w}^t; \boldsymbol{\xi}_i^t), \quad i \in \mathcal{H}, \tag{12}$$

where η is learning rate, $\boldsymbol{\xi}_i^t$ represents a minibatch uniformly sampled from the local data distribution $\boldsymbol{\xi}_i$ in the t-th communication round, and $\nabla \mathcal{L}(\boldsymbol{w}^t, \boldsymbol{\xi}_i^t)$ represents the gradient of loss over the minibatch $\boldsymbol{\xi}_i^t$.

We make the following assumptions, which are standard in FL (Karimireddy et al., 2021; 2022; Acharya et al., 2022).

Assumption 1 (Unbiased Estimator)**.** The stochastic gradients sampled from any local data distribution are unbiased estimators of local gradients over $\mathbb{R}^d$ for all honest clients, i.e.,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}_i^t}[\nabla \mathcal{L}_i(\boldsymbol{w}; \boldsymbol{\xi}_i^t)] = \nabla \mathcal{L}_i(\boldsymbol{w}), \\ &\forall \boldsymbol{w} \in \mathbb{R}^d, i \in \mathcal{H}, t \in \mathbb{N}^+. \end{aligned} \tag{13}$$

Assumption 2 (Bounded Variance)**.** The variance of stochastic gradients sampled from any local data distribution is uniformly bounded over $\mathbb{R}^d$ for all honest clients, i.e., there exists $\sigma \geq 0$ such that

$$\begin{aligned} &\mathbb{E}\|\nabla \mathcal{L}_i(\boldsymbol{w}; \boldsymbol{\xi}_i^t) - \nabla \mathcal{L}_i(\boldsymbol{w})\|^2 \leq \sigma^2, \\ &\forall \boldsymbol{w} \in \mathbb{R}^d, i \in \mathcal{H}, t \in \mathbb{N}^+. \end{aligned} \tag{14}$$

Assumption 3 (Gradient Dissimilarity)**.** The difference between the local gradients and the global gradient is uniformly bounded over $\mathbb{R}^d$ for all honest clients, i.e., there exists $\kappa \geq 0$ such that

$$\|\nabla \mathcal{L}_i(\boldsymbol{w}) - \nabla \mathcal{L}(\boldsymbol{w})\|^2 \leq \kappa^2, \quad \forall \boldsymbol{w} \in \mathbb{R}^d, i \in \mathcal{H}. \tag{15}$$

We consider arbitrary non-convex loss function $\mathcal{L}(\cdot)$ that satisfies the following Lipschitz condition. This condition is widely applied in the convergence analysis of Byzantine-robust federated learning (Karimireddy et al., 2022; Allen-Zhu et al., 2020; El-Mhamdi et al., 2021).

Assumption 4 (Lipschitz Smoothness)**.** The loss function is L-Lipschitz smooth over $\mathbb{R}^d$, i.e.,

$$\begin{aligned} &\|\nabla \mathcal{L}(\boldsymbol{w}) - \nabla \mathcal{L}(\boldsymbol{w}')\| \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|, \\ &\forall \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d. \end{aligned} \tag{16}$$

We consider robust AGRs that satisfy the following robustness criterion (Definition 1) introduced by Farhadkhani et al. (2022). A wide class of state-of-the-art robust AGRs satisfy this criterion (Farhadkhani et al., 2022).

Definition 1 ((f, λ)-resilient)**.** For integer $f < n/2$ and real value $\lambda > 0$, an AGR $\mathcal{A}$ is called (f, λ)-resilient if for any input $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and any set $\mathcal{S} \subseteq [n]$ of size $n - f$, the output of $\mathcal{A}$ satisfies:

$$\|\mathcal{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \bar{\boldsymbol{x}}_{\mathcal{S}}\| \leq \lambda \max_{i, i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|, \tag{17}$$

where $\bar{\boldsymbol{x}}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \boldsymbol{x}_i / |\mathcal{S}|$.

We show that given any (f, λ)-resilient base AGR $\mathcal{A}$, our GAS can help the global model to reach a better parameter.

Proposition 1. *Suppose Assumptions 1 to 4 hold, and let learning rate $\eta = 1/2L$. Given any (f, λ)-resilient robust AGR $\mathcal{A}$, we start from $\boldsymbol{w}^0$ and run GAS for T communication rounds, it satisfies*

$$\mathcal{L}(\boldsymbol{w}^0) \geq \frac{3}{16L} \sum_{t=1}^{T} (\|\nabla \mathcal{L}(\boldsymbol{w}^t)\|^2 - e^2), \tag{18}$$

where

$$e^2 = \mathcal{O}((\kappa^2 + \sigma^2) \tag{19}$$

$$\cdot (1 + \frac{n - f + 1}{p})(1 + \lambda^2 + \frac{1}{n - f}) \frac{f^2}{(n - f)^2}). \tag{20}$$

Please refer to Appendix C.1 for the proof. Proposition 1 provides an upper bound for the sum of gradient norms in the presence of Byzantine gradients. Equation (18) indicates that as the number of communication rounds increases, we can find an approximate optimal parameter $\boldsymbol{w}$ such that $\|\nabla \mathcal{L}(\boldsymbol{w})\|$ can be arbitrary close to e. κ^2 and σ^2 in Equation (19) are positively related to the gradient dimension d (Guerraoui et al., 2018). Therefore, the convergence error e grows larger when d increases. As the number of sub-vectors p increases, the approximation becomes better, i.e., e^2 decreases, which validates the efficacy of our approach. From another aspect, Proposition 1 also characterizes the fundamental difficulties of Byzantine-robust federated learning in the non-IID setting. The negative term $-e^2$ on the RHS of Equation (18) implies that FL may never converge to an optimal parameter. By contrast, the global model may wander among sub-optimal points. What's more, even after reaching the convergence point, the global model may step into another sub-optimal in the next communication round. It aligns with the previous lower bound in (Karimireddy et al., 2022). A detailed comparison of the convergence results between our approach and recent works is presented in Appendix C.2.

7. Experiments

7.1. Experimental Setups

Datasets. Our experiments are conducted on four real-world datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), a subset of ImageNet (Russakovsky et al., 2015) refered as ImageNet-12 (Li et al., 2021b) and FEMNIST (Caldas et al., 2018).

Data distribution. For CIFAR-10, CIFAR-100, and ImageNet-12, we use Dirichlet distribution to generate non-IID data by following Yurochkin et al. (2019); Li et al.

Table 1: Accuracy (mean±std) of different defenses under 6 attacks on CIFAR-10, CIFAR-100, FEMNIST, and ImageNet-12.

Dataset	Attack	BitFlip	LabelFlip	LIE	Min-Max	Min-Sum	IPM
CIFAR-10	Multi-Krum	43.19 ± 0.38	43.90 ± 0.03	37.03 ± 1.62	39.06 ± 0.07	23.68 ± 0.18	36.47 ± 0.22
	GAS (Multi-Krum)	**59.23** ± 0.55	**61.47** ± 0.26	**55.66** ± 0.93	**49.19** ± 0.72	**53.59** ± 0.96	**56.94** ± 3.60
	Bulyan	54.10 ± 0.19	55.12 ± 0.14	30.58 ± 0.75	29.03 ± 1.10	46.19 ± 0.92	33.88 ± 0.61
	GAS (Bulyan)	**59.14** ± 0.01	**61.21** ± 0.60	**48.90** ± 0.83	**48.35** ± 1.58	**53.74** ± 0.71	**56.53** ± 1.51
	Median	45.41 ± 0.44	51.88 ± 0.62	28.75 ± 0.35	32.72 ± 0.81	37.39 ± 0.90	43.21 ± 0.47
	GAS (Median)	**59.28** ± 0.24	**61.24** ± 1.34	**46.60** ± 0.13	**49.37** ± 1.13	**53.32** ± 1.90	**56.33** ± 0.82
	RFA	49.61 ± 0.31	44.35 ± 0.31	15.39 ± 0.37	16.62 ± 0.83	18.22 ± 0.43	45.92 ± 0.13
	GAS (RFA)	**53.35** ± 0.30	**62.25** ± 0.56	**52.69** ± 0.89	**52.64** ± 1.48	**56.16** ± 0.91	**62.26** ± 1.27
	DnC	58.63 ± 1.29	60.82 ± 1.56	61.07 ± 0.72	60.42 ± 0.59	53.71 ± 0.96	**59.99** ± 0.82
	GAS (DnC)	**58.96** ± 0.60	**61.02** ± 0.27	**61.87** ± 0.51	**61.04** ± 1.18	**54.36** ± 1.12	57.92 ± 1.71
	RBTM	54.27 ± 1.63	59.60 ± 1.76	47.67 ± 2.51	49.02 ± 0.31	50.74 ± 0.06	55.27 ± 1.60
	GAS (RBTM)	**59.41** ± 0.20	**60.75** ± 0.19	**52.10** ± 1.28	**49.60** ± 0.17	**53.63** ± 0.58	**56.65** ± 1.52
CIFAR-100	Multi-Krum	34.27 ± 0.28	35.57 ± 0.94	17.17 ± 0.08	16.77 ± 0.78	22.89 ± 0.61	15.93 ± 2.00
	GAS (Multi-Krum)	**42.41** ± 0.58	**42.55** ± 0.12	**27.81** ± 0.32	**31.18** ± 1.48	**41.33** ± 0.50	**42.62** ± 1.53
	Bulyan	35.77 ± 0.18	42.60 ± 0.07	35.41 ± 0.40	35.53 ± 1.38	39.13 ± 0.12	40.27 ± 1.64
	GAS (Bulyan)	**42.28** ± 1.61	**43.77** ± 0.46	**38.39** ± 0.19	**36.33** ± 1.51	**40.73** ± 0.39	**42.88** ± 0.14
	Median	36.62 ± 0.12	41.64 ± 0.76	22.75 ± 0.04	23.21 ± 0.71	30.68 ± 0.26	40.98 ± 0.38
	GAS (Median)	**42.41** ± 0.66	**42.62** ± 0.09	**35.16** ± 1.08	**36.46** ± 0.10	**41.08** ± 0.04	**43.63** ± 2.85
	RFA	21.32 ± 0.84	28.76 ± 1.33	25.63 ± 0.20	26.46 ± 1.83	28.33 ± 0.93	21.36 ± 0.54
	GAS (RFA)	**42.64** ± 0.44	**42.42** ± 0.25	**26.30** ± 1.08	**30.30** ± 0.12	**41.09** ± 0.66	**43.45** ± 0.52
	DnC	41.77 ± 0.62	42.93 ± 0.07	42.95 ± 1.03	40.15 ± 0.70	40.02 ± 1.07	41.23 ± 2.29
	GAS (DnC)	**43.35** ± 0.41	**43.57** ± 1.11	**43.64** ± 0.11	**41.66** ± 0.78	**41.02** ± 1.39	**43.25** ± 0.43
	RBTM	36.35 ± 0.17	42.67 ± 1.55	24.06 ± 0.09	26.24 ± 1.04	36.51 ± 0.40	43.12 ± 1.12
	GAS (RBTM)	**43.44** ± 0.81	**43.19** ± 2.65	**33.14** ± 0.58	**34.35** ± 0.76	**41.51** ± 0.93	**43.20** ± 0.76
FEMNIST	Multi-Krum	67.65 ± 0.23	57.43 ± 1.25	44.58 ± 0.07	28.32 ± 0.31	29.98 ± 0.45	12.26 ± 1.34
	GAS (Multi-Krum)	**84.29** ± 1.76	**85.45** ± 0.40	**74.76** ± 1.74	**57.46** ± 0.33	**70.65** ± 1.35	**81.46** ± 0.18
	Bulyan	77.58 ± 1.30	79.39 ± 2.14	56.43 ± 0.45	35.10 ± 0.69	44.83 ± 1.40	5.91 ± 0.17
	GAS (Bulyan)	**84.90** ± 0.69	**83.68** ± 0.76	**71.43** ± 1.07	**66.22** ± 0.47	**71.76** ± 0.99	**82.97** ± 1.04
	Median	80.25 ± 0.06	76.86 ± 1.96	64.88 ± 0.23	50.67 ± 0.37	61.33 ± 0.13	71.98 ± 0.77
	GAS (Median)	**84.59** ± 0.14	**85.67** ± 0.48	**76.19** ± 0.43	**65.84** ± 0.41	**70.84** ± 0.86	**82.18** ± 0.40
	RFA	5.46 ± 0.06	5.46 ± 0.01	5.46 ± 0.05	5.46 ± 0.03	5.46 ± 0.02	5.59 ± 0.09
	GAS (RFA)	**84.86** ± 0.78	**84.59** ± 0.20	**69.82** ± 0.33	**69.18** ± 0.09	**77.67** ± 1.31	**86.08** ± 2.51
	DnC	8.90 ± 0.31	77.71 ± 0.03	78.52 ± 0.28	8.29 ± 0.37	74.18 ± 0.03	74.70 ± 1.57
	GAS (DnC)	**84.71** ± 0.39	**85.39** ± 0.64	**82.54** ± 0.26	**74.37** ± 0.50	**75.41** ± 0.22	**82.73** ± 1.22
	RBTM	82.57 ± 0.34	81.57 ± 1.12	59.93 ± 0.20	65.20 ± 0.60	71.82 ± 0.73	76.88 ± 1.75
	GAS (RBTM)	**84.89** ± 1.94	**85.44** ± 0.20	**73.38** ± 0.31	**66.24** ± 0.94	**75.50** ± 1.13	**82.58** ± 1.85
ImageNet-12	Multi-Krum	44.36 ± 1.52	34.04 ± 1.69	45.38 ± 1.04	48.72 ± 0.16	57.69 ± 0.30	33.14 ± 0.86
	GAS (Multi-Krum)	**66.79** ± 1.08	**63.04** ± 0.14	**57.15** ± 0.19	**59.94** ± 0.32	**64.07** ± 1.38	**61.92** ± 0.04
	Bulyan	62.28 ± 0.84	59.84 ± 1.09	48.04 ± 2.22	48.97 ± 1.87	59.94 ± 0.51	60.67 ± 0.07
	GAS (Bulyan)	**66.76** ± 0.72	**62.28** ± 0.32	**57.44** ± 0.39	**58.81** ± 0.05	**65.00** ± 0.08	**62.76** ± 0.14
	Median	55.93 ± 0.55	58.14 ± 0.18	46.67 ± 1.01	49.07 ± 1.19	58.40 ± 0.03	43.62 ± 1.72
	GAS (Median)	**66.28** ± 0.41	**62.34** ± 1.10	**60.74** ± 1.24	**59.26** ± 0.31	**64.78** ± 2.10	**62.24** ± 0.51
	RFA	61.12 ± 1.26	61.31 ± 1.68	49.49 ± 1.33	53.04 ± 0.13	61.92 ± 0.67	63.97 ± 0.93
	GAS (RFA)	**66.92** ± 1.58	**63.88** ± 0.94	**61.41** ± 0.02	**59.42** ± 0.64	**67.02** ± 0.54	**66.67** ± 0.38
	DnC	54.94 ± 0.04	5.59 ± 0.06	58.01 ± 1.52	58.11 ± 0.41	60.42 ± 1.60	59.99 ± 0.50
	GAS (DnC)	**65.19** ± 1.63	**63.01** ± 0.27	**64.42** ± 0.19	**65.03** ± 1.23	**65.38** ± 1.68	**65.03** ± 0.04
	RBTM	60.06 ± 1.76	60.44 ± 0.37	55.77 ± 0.82	57.50 ± 0.10	63.91 ± 0.78	56.19 ± 1.05
	GAS (RBTM)	**66.99** ± 0.38	**61.92** ± 1.22	**59.87** ± 0.72	**59.81** ± 1.34	**64.94** ± 0.72	**63.40** ± 0.97

Table 2: Accuracy of different robust AGRs combined with Bucketing or GAS under six attacks on CIFAR-10.

Attack	BitFlip	LabelFlip	LIE	Min-Max	Min-Sum	IPM
Bucketing (Multi-Krum)	47.87	49.86	45.90	43.53	44.92	50.28
GAS (Multi-Krum)	**59.23**	**61.47**	**55.66**	**49.19**	**53.59**	**56.94**
Bucketing (Bulyan)	51.79	61.16	46.02	45.90	52.30	56.44
GAS (Bulyan)	**59.14**	**61.21**	**48.90**	**48.35**	**53.74**	**56.53**
Bucketing (Median)	53.17	59.50	**47.13**	47.93	51.52	52.69
GAS (Median)	**59.28**	**61.24**	46.60	**49.37**	**53.32**	**56.33**
Bucketing (RFA)	52.55	58.44	48.71	47.51	52.29	55.19
GAS (RFA)	**53.35**	**62.25**	**52.69**	**52.64**	**56.16**	**62.26**
Bucketing (DnC)	57.79	59.39	57.53	55.09	53.83	54.01
GAS (DnC)	**58.96**	**61.02**	**61.87**	**61.04**	**54.36**	**57.92**
Bucketing (RBTM)	53.25	60.10	51.87	49.32	53.56	53.77
GAS (RBTM)	**59.41**	**60.75**	**52.10**	**49.60**	**53.63**	**56.65**

Table 3: Accuracy of GAS with different number of sub-vectors p under LIE attack on CIFAR-10. d represents the number of model parameters.

p	100	1000	10000	100000	1000000	2472266 (d)
GAS (Multi-Krum)	55.07	63.23	**63.86**	60.16	58.31	57.70
GAS (Bulyan)	50.29	57.11	59.82	60.42	**60.47**	59.90

(2021a). We follow Li et al. (2021a) and set the number of clients $n = 50$ and the concentration parameter of Dirichlet distribution $\beta = 0.5$ as default. FEMNIST is a dataset with a natural non-IID partition. In particular, the data is partitioned into 3,597 clients based on the writer of the digit/character. For each client, we randomly sample 0.9 portion of data as training data and let the rest 0.1 portion of data be test data by following Caldas et al. (2018).

Evaluated attacks. We consider six representative attacks BitFlip (Allen-Zhu et al., 2020), LabelFlip (Allen-Zhu et al., 2020), LIE (Baruch et al., 2019), Min-Max (Shejwalkar & Houmansadr, 2021), Min-Sum (Shejwalkar & Houmansadr, 2021) and IPM (Xie et al., 2020). The detailed hyperparameter setting of the attacks are shown in Table 9 in Appendix D.

Baselines. We consider six representative robust AGRs: Multi-Krum (Blanchard et al., 2017), Bulyan (Guerraoui et al., 2018), Median (Yin et al., 2018), RFA (Pillutla et al., 2019), DnC (Shejwalkar & Houmansadr, 2021), RBTM (El-Mhamdi et al., 2021). We compare each AGR with its GAS variant and name them GAS (Multi-Krum), GAS (Bulyan), GAS (Median), GAS (RFA), GAS (DnC), and GAS (RBTM), respectively. The detailed hyperparameter settings of the robust AGRs are listed in Table 10 in Appendix D. We also compare our GAS against Bucketing (Karimireddy et al., 2022).

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. We run each experiment for five times and report the mean and standard deviation of the highest accuracy during the training process.

Other settings. We utilize AlexNet (Krizhevsky et al., 2017), SqueezeNet (Iandola et al., 2016), ResNet-18 (He et al., 2016) and a four-layer CNN (Caldas et al., 2018) for CIFAR-10, CIFAR-100, ImageNet-12 and FEMNIST, respectively. The number of Byzantine clients of all datasets is set to $f = 0.2 \cdot n$. We also consider up to $f = 0.3 \cdot n$ Byzantine clients in the ablation study. Please refer to Table 8 in Appendix D for more details.

7.2. Experiment Results

Main results. Table 1 illustrates the results of different defenses against popular attacks on CIFAR-10, CIFAR-100, ImageNet-12 and FEMNIST. From these tables, we observe that:

(1) Integrating current robust AGRs into our GAS generally outperform all their original versions on all datasets, which verifies the efficacy of our proposed GAS. For example, GAS improves the accuracy of Median by 15.93% under Min-Sum attack on CIFAR-10.

(2) The improvement of GAS (DnC) over DnC is relatively mild on CIFAR-10. Our interpretation is that when the dataset is relatively small and simple, DnC is capable of obtaining a rational gradient estimation. Nevertheless,

Table 4: The accuracy of GAS with $\delta = 0.1, 0.3$ under 20% LIE attack on CIFAR-10. N/A represents the case where the number of Byzantine clients f is known to the server and the server can exclude exactly f clients, i.e., δ is N/A.

δ	Multi-Krum	Bulyan	median	RFA	DnC	RBTM
N/A	**55.66**	**48.90**	**46.60**	**52.69**	**61.87**	**52.10**
0.1	54.30	46.94	45.74	52.21	56.93	51.47
0.3	50.48	44.96	43.09	52.04	60.50	50.21

Table 5: Accuracy (mean±std) of different defenses against LIE attack under different non-IID levels on CIFAR-10. A smaller β implies a higher non-IID level.

β	Multi-Krum	GAS (Multi-Krum)	Bulyan	GAS (Bulyan)	Median	GAS (Median)
0.3	12.19 ± 1.04	**52.80** ± 0.74	28.16 ± 0.44	**42.81** ± 0.63	25.62 ± 0.83	**40.97** ± 0.89
0.7	31.01 ± 0.54	**55.64** ± 0.60	44.72 ± 1.43	**51.29** ± 0.35	34.04 ± 0.29	**53.34** ± 0.08
β	RFA	GAS (RFA)	DnC	GAS (DnC)	RBTM	GAS (RBTM)
0.3	20.08 ± 0.13	**48.77** ± 0.84	59.99 ± 1.81	**60.21** ± 0.62	37.67 ± 0.18	**49.27** ± 0.05
0.7	18.11 ± 0.24	**53.25** ± 1.41	62.15 ± 0.73	**62.48** ± 0.52	48.43 ± 0.22	**52.25** ± 1.16

on larger and more complex datasets, i.e., FEMNIST and ImageNet-12, DnC fails to achieve satisfactory performance under Byzantine attacks.

(3) Although RFA collapses on FEMNIST, combining with our GAS can still improve it to satisfactory performance. Our illustration is that although the aggregated gradient of RFA deviates from the optimal gradient, it can still assist in identifying honest gradients when combined with GAS. As a result, GAS (RFA) is still effective on FEMNIST.

GAS v.s. Bucketing. We also compare our GAS method against Bucketing (Karimireddy et al., 2022) on CIFAR-10. For each robust AGR, we combine it with GAS or Bucketing separately and compare their performance. The results are posted in Table 2. As shown in Table 2, our GAS outperforms Bucketing in most cases. Except for LIE attack, the test accuracy of GAS (Median) is slightly lower than Bucketing (Median).

Number of sub-vectors. We vary sub-vector number p across $\{100, 1000, 10000, 100000, 1000000, 2472266(d)\}$ under LIE attack on the heterogeneous CIFAR-10 dataset. Other setups align with the main experiments. The results are provided in Table 3. As shown in Table 3, when p increases, the accuracy of GAS first increases, then slightly drops. Compared to GAS (Multi-Krum), GAS (Bulyan) demonstrates the best performance at a larger p and declines more slowly as p continues to increase. These results imply that: (1) GAS with a moderate p is more likely to achieve better performance; (2) the best p for different base AGRs differs.

Performance of GAS when number of Byzantine clients f is unknown. We run additional experiments to evaluate the performance of GAS when the number of Byzantine clients f is unknown to the server. In this case, GAS removes a fixed fraction of δ sampled clients in each communication round, where $\delta \in [0, 0.5)$ is the estimated ratio of Byzantine clients. We test for $\delta = 0.1, 0.3$ when there are 20% Byzantine clients under LIE attack on CIFAR-10. From results in Table 4, we can summarize that: (1) When the server knows the number of Byzantine clients (i.e., δ is N/A), GAS achieves the best performance; (2) When the number of Byzantine clients is unknown, the performance degradation of GAS is relatively mild; (3) Compared to excluding fewer clients ($\delta = 0.1$) from aggregation, the performance of GAS is generally better when excluding more clients ($\delta = 0.3$). We hypothesize this is because gradient heterogeneity is more impactful than LIE attack. Therefore, excluding honest gradients ($\delta = 0.3$) is more harmful to the performance of GAS compared to including Byzantine gradients ($\delta = 0.1$).

Results on different levels of non-IID. We discuss the impact of non-IID levels of data distributions. We modify the concentration parameter β to change the non-IID level. A smaller β implies a higher non-IID level. Table 5 demonstrates the accuracy of different defenses under LIE attack on CIFAR-10 dataset across $\beta = \{0.3, 0.7\}$. Other setups follow the default setup of the main experiments as illustrated in Section 7.1 and Appendix D. As shown in Table 5, all the existing AGRs achieve better performances than their original versions when combined with GAS, which validates the efficacy of our GAS under different non-IID levels. Moreover, when the level of non-IID is higher, the improve-

Table 6: Accuracy (mean±std) of different defenses against LIE attack with different Byzantine client numbers $f = \{5, 15\}$ on CIFAR-10. The number of total clients is fixed to $n = 50$.

f	Multi-Krum	GAS (Multi-Krum)	Bulyan	GAS (Bulyan)	Median	GAS (Median)
5	41.65 ± 1.78	**61.24** ± 0.01	56.28 ± 1.44	**58.27** ± 0.17	46.91 ± 1.36	**57.69** ± 1.81
15	10.00 ± 0.00	**34.70** ± 0.28	10.00 ± 0.00	**31.67** ± 0.19	18.85 ± 1.54	**30.95** ± 0.42
f	RFA	GAS (RFA)	DnC	GAS (DnC)	RBTM	GAS (RBTM)
5	22.37 ± 1.00	**58.06** ± 1.29	62.27 ± 0.04	**63.14** ± 0.20	55.92 ± 0.10	**59.72** ± 0.16
15	16.16 ± 0.14	**40.37** ± 0.26	57.28 ± 1.37	**60.14** ± 1.64	34.93 ± 1.36	**35.78** ± 1.51

Table 7: Accuracy (mean±std) of different defenses against LIE attack under different client numbers on CIFAR-10.

n	Multi-Krum	GAS (Multi-Krum)	Bulyan	GAS (Bulyan)	Median	GAS (Median)
75	28.72 ± 0.71	**54.89** ± 0.16	23.37 ± 1.22	**51.11** ± 0.00	44.89 ± 2.98	**52.22** ± 1.64
100	32.49 ± 1.22	**56.51** ± 0.01	21.93 ± 0.55	**46.49** ± 1.33	33.82 ± 0.21	**46.12** ± 0.17
n	RFA	GAS (RFA)	DnC	GAS (DnC)	RBTM	GAS (RBTM)
75	16.89 ± 1.38	**49.85** ± 0.06	59.31 ± 1.33	**59.75** ± 0.42	45.06 ± 0.96	**50.24** ± 0.31
100	14.01 ± 1.34	**49.85** ± 1.97	58.88 ± 1.45	**59.61** ± 1.19	40.38 ± 0.48	**47.02** ± 0.03

ment on robust AGRs is more significant. The results further confirm that our GAS can overcome the failures aggravated under a higher non-IID level.

Results on different number of Byzantine clients. We also conduct experiments across different number of Byzantine clients (the total number of clients n is fixed). Other setups follow the default setup of the main experiments in Section 7.1 and Appendix D. Table 6 demonstrates the results of different defenses under LIE attack across $f = \{5, 15\}$ Byzantine clients on CIFAR-10 dataset. As shown in Table 6, our GAS outperforms the corresponding baselines across all Byzantine client numbers.

Results on different number of clients. We further analyze the efficacy of our GAS under different number of clients. We test the performance of different defenses under LIE attack across $n = \{75, 100\}$ clients on CIFAR-10 dataset. The number of Byzantine clients is set to $f = 0.2 \cdot n$ correspondingly. Other setups follow the default setup of the main experiments in Section 7.1 and Appendix D. These results demonstrate that all the robust AGRs consistently outperform all their original versions when combined with our GAS, which validates that our GAS can effectively defend against Byzantine across different numbers of clients.

8. Conclusion and Discussion

In this work, we identify two main challenges of Byzantine robustness in the non-IID setting: the curse of dimensionality and gradient heterogeneity. Robust AGRs that try to include all honest gradients in aggregation suffer from the curse of dimensionality. Other robust AGRs that aggregate fewer gradients to get rid of Byzantines fail due to gradient heterogeneity. Motivated by the above discoveries, we propose a novel GrAdient Splitting (GAS) based approach that is compatible with most existing robust AGRs and overcomes the high dimensionality and gradient heterogeneity. GAS splits each high-dimensional gradient into low-dimensional sub-vectors and detects Byzantine gradients with the sub-vectors to address the curse of dimensionality. Then, GAS aggregates all the identified honest gradients to handle the gradient heterogeneity to alleviate the gradient heterogeneity issue. We also provide a detailed convergence analysis of our proposed GAS. Empirical studies on four real-world datasets justify the efficacy of GAS.

Discussion. In the first step of GAS, we use an equal splitting mechanism for splitting. In fact, there are many other mechanisms, e.g., split gradients by layer. A future research direction is to discover more effective splitting mechanisms for GAS. Note that our GAS can also be combined with adaptive client selection strategies (Wan et al., 2022) to achieve better Byzantine robustness. We would discuss it more in our future work.

Acknowledgements

This work is supported by the National Key R&D Program of China (No.2022YFB3304100) and by the Zhejiang University-China Zheshang Bank Co., Ltd. Joint Research Center. This work is also sponsored by Sony AI.

References

Acharya, A., Hashemi, A., Jain, P., Sanghavi, S., Dhillon, I. S., and Topcu, U. Robust training in high dimensions via block coordinate geometric median descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 11145–11168. PMLR, 2022.

Allen-Zhu, Z., Ebrahimianghazani, F., Li, J., and Alistarh, D. Byzantine-resilient non-convex stochastic gradient descent. In *International Conference on Learning Representations*, 2020.

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

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Chen, C., Zhang, J., Tung, A. K., Kankanhalli, M., and Chen, G. Robust federated recommendation system. *arXiv preprint arXiv:2006.08259*, 2020.

Chen, C., Liu, Y., Ma, X., and Lyu, L. Calfat: Calibrated federated adversarial training with label skewness. In *NeurIPS*, 2022a.

Chen, C., Lyu, L., Yu, H., and Chen, G. Practical attribute reconstruction attack against federated learning. *IEEE Transactions on Big Data*, 2022b.

Data, D. and Diggavi, S. Byzantine-resilient high-dimensional sgd with local iterations on heterogeneous data. In *International Conference on Machine Learning*, pp. 2478–2488. PMLR, 2021.

El-Mhamdi, E. M., Farhadkhani, S., Guerraoui, R., Guirguis, A., Hoang, L.-N., and Rouault, S. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). *Advances in Neural Information Processing Systems*, 34:25044–25057, 2021.

Farhadkhani, S., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. Byzantine machine learning made easy by resilient averaging of momentums. In *International Conference on Machine Learning*, pp. 6246–6283. PMLR, 2022.

Ghosh, A., Hong, J., Yin, D., and Ramchandran, K. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

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

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

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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

Karimireddy, S. P., He, L., and Jaggi, M. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pp. 5311–5319. PMLR, 2021.

Karimireddy, S. P., He, L., and Jaggi, M. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

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

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

Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34, 2021b.

Lyu, L., Yu, H., and Yang, Q. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., and Philip, S. Y. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*, 2022.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Park, J., Han, D.-J., Choi, M., and Moon, J. Sageflow: Robust federated learning against both stragglers and adversaries. *Advances in Neural Information Processing Systems*, 34:840–851, 2021.

Peng, J., Wu, Z., Ling, Q., and Chen, T. Byzantine-robust variance-reduced federated learning over distributed non-iid data. *Information Sciences*, 616:367–391, 2022.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

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

Wan, W., Hu, S., Lu, j., Zhang, L. Y., Jin, H., and He, Y. Shielding federated learning: Robust aggregation with adaptive client selection. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 753–760. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/106. URL https://doi.org/10.24963/ijcai.2022/106. Main Track.

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

Yang, G. and Schoenholz, S. Mean field residual networks: On the edge of chaos. *Advances in neural information processing systems*, 30, 2017.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, pp. 5650–5659. PMLR, 2018.

Yu, S. and Kar, S. Secure distributed optimization under gradient attacks. *arXiv preprint arXiv:2210.15821*, 2022.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261, 2019.

Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., and Wu, C. Dense: Data-free one-shot federated learning. *Advances in Neural Information Processing Systems*, 35:21414–21428, 2022.

Zhang, J., Li, B., Chen, C., Lyu, L., Wu, S., Ding, S., and Wu, C. Delving into the adversarial robustness of federated learning. *arXiv preprint arXiv:2302.09479*, 2023.

Zhao, Y., Zhao, J., Jiang, L., Tan, R., Niyato, D., Li, Z., Lyu, L., and Liu, Y. Privacy-preserving blockchain-based federated learning for iot devices. *IEEE Internet of Things Journal*, 8(3):1817–1829, 2020.

A. Setups for Experiments in Section 4

The experiments are conducted on CIFAR-10 (Krizhevsky et al., 2009).

For both IID and non-IID settings, the number of clients is set to $n = 50$. For IID data distribution, all 50,000 samples are randomly partitioned into 50 clients each containing 1,000 samples. For non-IID data distribution, the samples are partitioned in a Dirichlet manner with concentration parameter $\beta = 0.5$. Please refer to Section 7.1 for the details of Dirichlet partition.

The number of Byzantine clients is set to $f = 10$. LIE (Baruch et al., 2019) attack with $z = 1.5$ is considered.

We use AlexNet (Krizhevsky et al., 2017) as the model architecture. The number of communication rounds is set to 500. In each communication round, all clients participate in the training.

For local training, the number of local epochs is set to 1, batch size is set to 64, the optimizer is set to SGD. For SGD optimizer, learning rate is set to 0.1, momentum is set to 0.5, weight decay coefficient is set to 0.0001. We also adopt gradient clipping with clipping norm 2.

Six robust AGRs are considered: Bulyan (Guerraoui et al., 2018), Median (Yin et al., 2018), RBTM (El-Mhamdi et al., 2021), Multi-Krum (Blanchard et al., 2017), RFA (Pillutla et al., 2019), DnC (Shejwalkar & Houmansadr, 2021)

B. Computation Cost of GAS

We first give the computation cost of the proposed GAS method. The computation cost of GAS is closely related to the computation cost of the base robust AGR $\mathcal{A}$. We use $\texttt{cost}_{\mathcal{A}}(d, n)$ to denote the computation cost of the base AGR $\mathcal{A}$ given n gradients of dimensionality d.

Our GAS method has three steps: splitting, identification, and aggregation.

Splitting. The splitting step is of complexity $\mathcal{O}(d)$;

Identification. The identification step consists of two parts: apply AGR $\mathcal{A}$ to sub-vectors ($\mathcal{O}(p\texttt{cost}_{\mathcal{A}}(d/p, n))$) and compute identification scores ($\mathcal{O}(nd + np)$).

Aggregation. The complexity of aggregation step is $\mathcal{O}(n \log(n - f) + (n - f)d)$.

In summary, the overall complexity for GAS is $\mathcal{O}(d + p\texttt{cost}_{\mathcal{A}}(d/p, n) + nd + np + n \log(n - f) + (n - f)d) = \mathcal{O}(n(d + \log(n - f)) + p\texttt{cost}_{\mathcal{A}}(d/p, n))$.

Then we analyze the computation cost of GAS $\mathcal{O}(n(d + \log(n - f)) + p\texttt{cost}_{\mathcal{A}}(d/p, n))$. Since $n \ll d$ [7], the first term $\mathcal{O}(n(d + \log(n - f))) \approx \mathcal{O}(d) \ll \Omega(d^2)$. The second term $\mathcal{O}(p\texttt{cost}_{\mathcal{A}}(d/p, n))$ relies on $\texttt{cost}_{\mathcal{A}}(d/p, n)$, the cost of base AGR $\mathcal{A}$. The computation cost of popular AGRs are usually $\mathcal{O}(d/p)$ under assumption $n \ll d$ [7], e.g., Krum ($\mathcal{O}(n^2 d/p)$), Bulyan ($\mathcal{O}(n^2 d/p)$). Therefore, the second term usually satisfies $\mathcal{O}(p\texttt{cost}_{\mathcal{A}}(d/p, n)) \approx \mathcal{O}(d) \ll \Omega(d^2)$. In summary, the computation cost of GAS is generally $\mathcal{O}(d)$ (consider only d and omit n), which is much smaller than $\Omega(d^2)$.

C. Convergence Analysis

In this section, we provide the proof for our convergence results in Proposition 1 and the comparison of our convergence results with recent works.

We first restate the assumptions, the definition and the proposition for the integrity of this section.

Assumption 1 (Unbiased Estimator)**.** The stochastic gradients sampled from any local data distribution are unbiased estimators of local gradients over $\mathbb{R}^d$ for all honest clients, i.e.,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\xi}_i^t}[\nabla \mathcal{L}_i(\boldsymbol{w}; \boldsymbol{\xi}_i^t)] = \nabla \mathcal{L}_i(\boldsymbol{w}), \\ &\forall \boldsymbol{w} \in \mathbb{R}^d, i \in \mathcal{H}, t \in \mathbb{N}^+. \end{aligned} \tag{13}$$

Assumption 2 (Bounded Variance)**.** The variance of stochastic gradients sampled from any local data distribution is

uniformly bounded over $\mathbb{R}^d$ for all honest clients, i.e., there exists $\sigma \geq 0$ such that

$$\begin{aligned} \mathbb{E}\|\nabla \mathcal{L}_i(\boldsymbol{w}; \boldsymbol{\xi}_i^t) - \nabla \mathcal{L}_i(\boldsymbol{w})\|^2 \leq \sigma^2, \\ \forall \boldsymbol{w} \in \mathbb{R}^d, i \in \mathcal{H}, t \in \mathbb{N}^+. \end{aligned} \tag{14}$$

Assumption 3 (Gradient Dissimilarity)**.** The difference between the local gradients and the global gradient is uniformly bounded over $\mathbb{R}^d$ for all honest clients, i.e., there exists $\kappa \geq 0$ such that

$$\|\nabla \mathcal{L}_i(\boldsymbol{w}) - \nabla \mathcal{L}(\boldsymbol{w})\|^2 \leq \kappa^2, \quad \forall \boldsymbol{w} \in \mathbb{R}^d, i \in \mathcal{H}. \tag{15}$$

Assumption 4 (Lipschitz Smoothness)**.** The loss function is L-Lipschitz smooth over $\mathbb{R}^d$, i.e.,

$$\begin{aligned} \|\nabla \mathcal{L}(\boldsymbol{w}) - \nabla \mathcal{L}(\boldsymbol{w}')\| \leq L\|\boldsymbol{w} - \boldsymbol{w}'\|, \\ \forall \boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d. \end{aligned} \tag{16}$$

Definition 1 ((f, λ)-resilient)**.** For integer $f < n/2$ and real value $\lambda > 0$, an AGR $\mathcal{A}$ is called (f, λ)-resilient if for any input $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and any set $\mathcal{S} \subseteq [n]$ of size $n - f$, the output of $\mathcal{A}$ satisfies:

$$\|\mathcal{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \bar{\boldsymbol{x}}_{\mathcal{S}}\| \leq \lambda \max_{i,i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|, \tag{17}$$

where $\bar{\boldsymbol{x}}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \boldsymbol{x}_i / |\mathcal{S}|$.

Proposition 1. *Suppose Assumptions 1 to 4 hold, and let learning rate $\eta = 1/2L$. Given any (f, λ)-resilient robust AGR $\mathcal{A}$, we start from $\boldsymbol{w}^0$ and run GAS for T communication rounds, it satisfies*

$$\mathcal{L}(\boldsymbol{w}^0) \geq \frac{3}{16L} \sum_{t=1}^{T} (\|\nabla \mathcal{L}(\boldsymbol{w}^t)\|^2 - e^2), \tag{18}$$

where

$$e^2 = \mathcal{O}((\kappa^2 + \sigma^2) \tag{19}$$

$$\cdot (1 + \frac{n - f + 1}{p})(1 + \lambda^2 + \frac{1}{n - f}) \frac{f^2}{(n - f)^2}). \tag{20}$$

C.1. Proof for Proposition 1

Lemma 1. *For positive integer n, $f \leq n$ and real value λ, AGR $\mathcal{A}$ is (f, λ)-resilient. Then for any set of random variables $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $\mathcal{S} \subseteq [n]$ of size $n - f$ that satisfies,*

$$\mathbb{E}[\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2] \leq \rho^2, \quad \forall i, i' \in \mathcal{H} \tag{21}$$

we have

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \boldsymbol{x}_{\mathcal{S}}\|^2] \leq 4\lambda^2 \cdot \frac{(n - f - 1)^2}{n - f} \cdot \rho^2, \tag{22}$$

where $\boldsymbol{x}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \boldsymbol{x}_i / |\mathcal{S}|$.

Proof. Since $\mathcal{A}$ is (f, λ)-resilient, we have

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \bar{\boldsymbol{x}}_{\mathcal{S}}\|^2] \leq \mathbb{E}[\lambda^2 \max_{i,i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2] = \lambda^2 \mathbb{E}[\max_{i,i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2]. \tag{23}$$

Then we bound $\max_{i,i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2$ as follows.

$$\max_{i,i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 \leq \max_{i,i' \in \mathcal{S}} 2(\|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2 + \|\boldsymbol{x}_{\mathcal{S}} - \boldsymbol{x}_{i'}\|^2) \tag{24}$$

$$\leq \max_{i,i' \in \mathcal{S}} 2\|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2 + \max_{i,i' \in \mathcal{S}} 2\|\boldsymbol{x}_{\mathcal{S}} - \boldsymbol{x}_{i'}\|^2 \tag{25}$$

$$= 4 \max_{i \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2 \tag{26}$$

$$\leq 4 \sum_{i \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2 \tag{27}$$

Here Inequality (24) comes from the Cauchy inequality.

We further bound $\|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2$ for all $i \in \mathcal{S}$ as follows.

$$\|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2 = \frac{1}{(n-f)^2}\|\sum_{i' \in \mathcal{S}\setminus\{i\}} (\boldsymbol{x}_i - \boldsymbol{x}_{i'})\|^2 \tag{28}$$

$$\leq \frac{1}{(n-f)^2} \cdot (n-f-1) \sum_{i' \in \mathcal{S}\setminus\{i\}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2 \tag{29}$$

$$= \frac{n-f-1}{(n-f)^2} \sum_{i' \in \mathcal{S}\setminus\{i\}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2. \tag{30}$$

Here Equation (29) comes from the Cauchy inequality.

Combine Equations (23) and (30) and Inequality (27) and we have

$$\mathbb{E}[\|\mathcal{A}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) - \boldsymbol{x}_{\mathcal{S}}\|^2] \leq \lambda^2 \mathbb{E}[\max_{i,i' \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2] \tag{31}$$

$$\leq \lambda^2 \mathbb{E}[4 \sum_{i \in \mathcal{S}} \|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{S}}\|^2] \tag{32}$$

$$\leq \lambda^2 \mathbb{E}[4 \sum_{i \in \mathcal{S}} \frac{n-f-1}{(n-f)^2} \sum_{i' \in \mathcal{S}\setminus\{i\}} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2] \tag{33}$$

$$\leq 4\lambda^2 \cdot \frac{n-f-1}{(n-f)^2} \sum_{i,i' \in \mathcal{S}, i \neq i'} \mathbb{E}[\|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2] \tag{34}$$

$$\leq 4\lambda^2 \cdot \frac{n-f-1}{(n-f)^2} \cdot (n-f)(n-f-1)\rho^2 \tag{35}$$

$$= 4\lambda^2 \cdot \frac{(n-f-1)^2}{n-f} \cdot \rho^2 \tag{36}$$

□

We state and prove the following lemma for the proof of Lemma 3.

Lemma 2. *For any random vector $\boldsymbol{X}$, we have*

$$\mathrm{Var}[\|\boldsymbol{X}\|] \leq \mathbb{E}\|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}\|^2. \tag{37}$$

Proof. From the definition of variance, we have

$$\mathrm{Var}[\|\boldsymbol{X}\|] = \mathbb{E}(\|\boldsymbol{X}\| - \mathbb{E}\|\boldsymbol{X}\|)^2 \tag{38}$$

$$= \mathbb{E}(\|\boldsymbol{X}\| - \|\mathbb{E}\boldsymbol{X}\|)^2 - (\|\mathbb{E}\boldsymbol{X}\| - \mathbb{E}\|\boldsymbol{X}\|)^2 \tag{39}$$

$$\leq \mathbb{E}(\|\boldsymbol{X}\| - \|\mathbb{E}\boldsymbol{X}\|)^2 \tag{40}$$

$$\leq \mathbb{E}\|\boldsymbol{X} - \mathbb{E}\boldsymbol{X}\|^2. \tag{41}$$

The second inequality comes from triangular inequality. □

Lemma 3 (Aggregation error)**.** *Suppose Assumptions 1 to 3 hold. Given an (f, λ)-resilient robust AGR $\mathcal{A}$, for any $t > 0$, it satisfies*

$$\mathbb{E}[\|\hat{\boldsymbol{g}} - \bar{\boldsymbol{g}}\|^2] \leq \mathcal{O}((\kappa^2 + \sigma^2)(1 + \frac{n-f+1}{p})(1 + \lambda^2 + \frac{1}{n-f})\frac{f^2}{(n-f)^2}) \tag{42}$$

Proof. We rewrite $\hat{\boldsymbol{g}}$ as follows.

$$\hat{\boldsymbol{g}} = \frac{1}{n-f}\sum_{i\in\mathcal{I}}\boldsymbol{g}_i = \frac{1}{n-f}(\sum_{h\in\tilde{\mathcal{H}}}\boldsymbol{g}_h + \sum_{b\in\tilde{\mathcal{B}}}\boldsymbol{g}_b) = \frac{|\tilde{\mathcal{H}}|}{n-f}\boldsymbol{g}_{\tilde{\mathcal{H}}} + \frac{|\tilde{\mathcal{B}}|}{n-f}\boldsymbol{g}_{\tilde{\mathcal{B}}}. \tag{43}$$

Here $\tilde{\mathcal{H}} = \mathcal{H}\cap\mathcal{I}$, $\tilde{\mathcal{B}} = \mathcal{B}\cap\mathcal{I}$, and $\boldsymbol{g}_{\mathcal{S}} = \sum_{i\in\mathcal{S}}\boldsymbol{g}_i/|\mathcal{S}|$ for all $\mathcal{S}\in[n]$.

Then, we can bound $\mathbb{E}\left\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}\right\|^2$ as follows

$$\begin{aligned}
\mathbb{E}[\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}\|^2] &= \mathbb{E}[\|\frac{|\tilde{\mathcal{H}}|}{n-f}\boldsymbol{g}_{\tilde{\mathcal{H}}} + \frac{|\tilde{\mathcal{B}}|}{n-f}\boldsymbol{g}_{\tilde{\mathcal{B}}} - \bar{\boldsymbol{g}}\|^2] && (44)\\
&= \mathbb{E}[\|\frac{|\tilde{\mathcal{H}}|}{n-f}(\boldsymbol{g}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}) + \frac{|\tilde{\mathcal{B}}|}{n-f}(\boldsymbol{g}_{\tilde{\mathcal{B}}} - \bar{\boldsymbol{g}})\|^2] && (45)\\
&\le \frac{2|\tilde{\mathcal{H}}|^2}{(n-f)^2}\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}\|^2] + \frac{2|\tilde{\mathcal{B}}|^2}{(n-f)^2}\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{B}}} - \bar{\boldsymbol{g}}\|^2]. && (46)
\end{aligned}$$

We bound $\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}\|^2]$ as follows.

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}\|^2] &= \mathbb{E}[\|(\boldsymbol{g}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}_{\tilde{\mathcal{H}}}) + (\bar{\boldsymbol{g}}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}})\|^2] && (47)\\
&= \mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}_{\tilde{\mathcal{H}}}\|^2] + \|\bar{\boldsymbol{g}}_{\tilde{\mathcal{H}}} - \bar{\boldsymbol{g}}\|^2 && (48)\\
&\le \frac{\sigma^2}{|\tilde{\mathcal{H}}|} + \kappa^2, && (49)
\end{aligned}$$

where $\bar{\boldsymbol{g}}_{\tilde{\mathcal{H}}} = \mathbb{E}[\boldsymbol{g}_{\tilde{\mathcal{H}}}]$.

Then we consder $\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{B}}} - \bar{\boldsymbol{g}}\|^2]$. According to the law of total expectation, we have

$$\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{B}}} - \bar{\boldsymbol{g}}\|^2] = \sum_{\tilde{f}=0}^{f}\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{B}}} - \bar{\boldsymbol{g}}\|^2 \mid |\tilde{\mathcal{B}}| = \tilde{f}]\Pr(|\tilde{\mathcal{B}}| = \tilde{f}). \tag{50}$$

For all parameter group $q\in[p]$ and $i,j\in\mathcal{H}$, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{g}_i^{(q)} - \boldsymbol{g}_j^{(q)}\|^2] &= \mathbb{E}[\|(\boldsymbol{g}_i^{(q)} - \bar{\boldsymbol{g}}_i^{(q)}) + (\bar{\boldsymbol{g}}_i^{(q)} - \bar{\boldsymbol{g}}^{(q)}) + (\bar{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}_j^{(q)}) + (\bar{\boldsymbol{g}}_j^{(q)} - \boldsymbol{g}_j^{(q)})\|^2] && (51)\\
&= \mathbb{E}[\|\boldsymbol{g}_i^{(q)} - \bar{\boldsymbol{g}}_i^{(q)}\|^2] + \|\bar{\boldsymbol{g}}_i^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|^2 + \|\bar{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}_j^{(q)}\|^2 + \mathbb{E}[\|\bar{\boldsymbol{g}}_j^{(q)} - \boldsymbol{g}_j^{(q)}\|^2]\\
&\quad + 2\langle\bar{\boldsymbol{g}}_i^{(q)} - \bar{\boldsymbol{g}}^{(q)}, \bar{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}_j^{(q)}\rangle && (52)\\
&\le \mathbb{E}[\|\boldsymbol{g}_i^{(q)} - \bar{\boldsymbol{g}}_i^{(q)}\|^2] + \|\bar{\boldsymbol{g}}_i^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|^2 + \|\bar{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}_j^{(q)}\|^2 + \mathbb{E}[\|\bar{\boldsymbol{g}}_j^{(q)} - \boldsymbol{g}_j^{(q)})\|^2]\\
&\quad + 2\|\bar{\boldsymbol{g}}_i^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|\cdot\|\bar{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}_j^{(q)}\| && (53)\\
&\le 2\sigma^2 + 4\kappa^2 && (54)
\end{aligned}$$

Here Equation (52) is due to the independence of $\boldsymbol{g}_i^{(q)}$ and $\boldsymbol{g}_j^{(q)}$, Inequality (53) comes from the Cauchy inequality, and Inequality (54) follows Assumptions 2 and 3.

Then according to the Lemma 1, we have

$$\mathbb{E}[\|\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2] \le c^2\max_{i,j\in\mathcal{H}}\mathbb{E}[\|\boldsymbol{g}_i^{(q)} - \boldsymbol{g}_j^{(q)}\|^2] \le c^2(2\sigma^2 + 4\kappa^2), \tag{55}$$

where $c^2 = 4\lambda^2(n-f-1)^2/(n-f)$.

For honest client h, the expectation of abnormal score $s_h^{(q)}$ from group q can be bounded as follows.

$$\mathbb{E}[s_h^{(q)}] = \mathbb{E}[\|\boldsymbol{g}_h^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|] \tag{56}$$
$$\leq \mathbb{E}[\|\boldsymbol{g}_h^{(q)} - \bar{\boldsymbol{g}}_h^{(q)}\| + \|\bar{\boldsymbol{g}}_h^{(q)} - \bar{\boldsymbol{g}}^{(q)}\| + \|\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\| + \|\boldsymbol{g}^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|] \tag{57}$$
$$= \mathbb{E}[\|\boldsymbol{g}_h^{(q)} - \bar{\boldsymbol{g}}_h^{(q)}\|] + \|\bar{\boldsymbol{g}}_h^{(q)} - \bar{\boldsymbol{g}}^{(q)}\| + \mathbb{E}[\|\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|] + \mathbb{E}[\|\boldsymbol{g}^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|] \tag{58}$$
$$\leq \sqrt{\mathbb{E}[\|\boldsymbol{g}_h^{(q)} - \bar{\boldsymbol{g}}_h^{(q)}\|^2]} + \|\bar{\boldsymbol{g}}_h^{(q)} - \bar{\boldsymbol{g}}^{(q)}\| + \sqrt{\mathbb{E}[\|\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{g}^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|^2]} \tag{59}$$
$$\leq (1 + \frac{1}{\sqrt{n-f}})\sigma + \kappa + c\sqrt{2\sigma^2 + 4\kappa^2}. \tag{60}$$

Here Inequality (57) is a result of triangular inequality, Inequality (59) comes from Cauchy inequality, and Inequality (60) is a combined result of Equation (55) and Assumptions 2 and 3.

The variance of $s_h^{(q)}$ can be bounded as follows.

$$\mathrm{Var}[s_h^{(q)}] = \mathbb{E}[(s_h^{(q)})^2] - (\mathbb{E}[s_h^{(q)}])^2 \tag{61}$$
$$\leq \mathbb{E}[(s_h^{(q)})^2] \tag{62}$$
$$= \mathbb{E}[\|\boldsymbol{g}_h^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|^2] \tag{63}$$
$$= \mathbb{E}[\|(\boldsymbol{g}_h^{(q)} - \bar{\boldsymbol{g}}_h^{(q)}) + (\bar{\boldsymbol{g}}_h^{(q)} - \bar{\boldsymbol{g}}^{(q)}) + (\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}) + (\boldsymbol{g}^{(q)} - \hat{\boldsymbol{g}}^{(q)})\|^2] \tag{64}$$
$$\leq 4\mathbb{E}[\|\boldsymbol{g}_h^{(q)} - \bar{\boldsymbol{g}}_h^{(q)}\|^2 + \|\bar{\boldsymbol{g}}_h^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|^2 + \|\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2 + \|\boldsymbol{g}^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|^2]. \tag{65}$$

Here Inequality (65) is a result of Cauchy inequality.

We bound $\mathbb{E}[\|\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2]$ as follows.

$$\mathbb{E}[\|\bar{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2] = \mathbb{E}[\|\frac{1}{n-f}\sum_{i\in\mathcal{H}}(\bar{\boldsymbol{g}}_i^{(q)} - \boldsymbol{g}_i^{(q)})\|^2] \tag{66}$$
$$= \frac{1}{(n-f)^2}\sum_{i\in\mathcal{H}}\mathbb{E}[\|\bar{\boldsymbol{g}}_i^{(q)} - \boldsymbol{g}_i^{(q)}\|^2] \tag{67}$$
$$\leq \frac{1}{(n-f)^2}\sum_{i\in\mathcal{H}}\sigma^2 \tag{68}$$
$$= \frac{\sigma^2}{n-f} \tag{69}$$

Here Equation (67) comes from the independence of minibatches sampling across different clients, and Inequality (68) is a result of Assumption 2.

Applying Assumptions 2 and 3 and Equations (55) and (69) to Inequality (65), we have

$$\mathrm{Var}[s_h^{(q)}] \leq 4(\sigma^2 + \kappa^2 + \frac{\sigma^2}{n-f} + c(2\sigma^2 + 4\kappa^2)) \tag{70}$$
$$= (4 + 8c^2 + \frac{4}{n-f})\sigma^2 + (4 + 16c^2)\kappa^2. \tag{71}$$

According to Inequality (60) and Equation (71), we can bound the expectation and variance of total abnormal score s_h of an honest client h.

$$\mathbb{E}[s_h] = \mathbb{E}[\sum_{q=1}^{p} s_h^{(q)}] \leq p(\sigma + \kappa + c\sqrt{2\sigma^2 + 4\kappa^2}) := A, \tag{72}$$

$$\mathrm{Var}[s_h] = \sum_{q=1}^{p}\mathrm{Var}[s_h^{(q)}] \leq p((4 + 8c^2 + \frac{4}{n-f})\sigma^2 + (4 + 16c^2)\kappa^2) := B. \tag{73}$$

Here the addictive property of variance is a result of the independence of group abnormal scores $\{s_h^{(q)} \mid q \in [p]\}$, which comes from the independence of components in a gradient (Yang & Schoenholz, 2017).

From Chebyshev's inequality, for any $\Delta_h > 0$ and honest client $h \in [n] \setminus \mathcal{B}$, we have

$$P(s_h < \mathbb{E}[s_h] + \Delta_h) \geq 1 - \frac{\mathrm{Var}[s_h]}{\Delta_h^2}. \tag{74}$$

Consider the expectation of abnormal score $s_b^{(q)}$ from group q for Byzantine client $b \in \mathcal{B}$

$$\begin{aligned}
\mathbb{E}[s_b^{(q)}] &= \mathbb{E}[\|\boldsymbol{g}_b^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|] && (75)\\
&= \mathbb{E}[\|(\boldsymbol{g}_b^{(q)} - \bar{\boldsymbol{g}}^{(q)}) - (\hat{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}^{(q)})\|] && (76)\\
&\geq \mathbb{E}[\|\boldsymbol{g}_b^{(q)} - \bar{\boldsymbol{g}}^{(q)}\| - \|\hat{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|] && (77)\\
&\geq \mathbb{E}[\|\boldsymbol{g}_b^{(q)} - \bar{\boldsymbol{g}}^{(q)}\| - (\|\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\| + \|\boldsymbol{g}^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|)] && (78)\\
&\geq \|\boldsymbol{g}_b^{(q)} - \bar{\boldsymbol{g}}^{(q)}\| - (\sqrt{\mathbb{E}[\|\hat{\boldsymbol{g}}^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|^2]} + \sqrt{\mathbb{E}[\|\boldsymbol{g}^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|^2]}) && (79)\\
&\geq \delta_b - c\sqrt{2\sigma^2 + 4\kappa^2} - \frac{\sigma}{\sqrt{n-f}} && (80)
\end{aligned}$$

where $\delta_b = \|\boldsymbol{g}_b^{(q)} - \bar{\boldsymbol{g}}^{(q)}\|$ is the expected deviation of Byzantine client b from the average of honest gradients. Here the first and second inequalities come from triangular inequality, the third inequality is based on Cauchy inequality, and the 4-th inequality is a combined result of Equations (55) and (69).

The variance of abnormal score $s_b^{(q)}$ can be bounded as follows.

$$\begin{aligned}
\mathrm{Var}[s_b^{(q)}] &= \mathrm{Var}[\|\boldsymbol{g}_b^{(q)} - \hat{\boldsymbol{g}}^{(q)}\|] && (81)\\
&\leq \mathbb{E}[\|\boldsymbol{g}_b^{(q)} - \hat{\boldsymbol{g}}^{(q)} - \mathbb{E}[\boldsymbol{g}_b^{(q)} - \hat{\boldsymbol{g}}^{(q)}]\|^2] && (82)\\
&= \mathbb{E}[\|(\boldsymbol{g}_b^{(q)} - \mathbb{E}[\boldsymbol{g}_b^{(q)}]) - (\hat{\boldsymbol{g}}^{(q)} - \mathbb{E}[\hat{\boldsymbol{g}}^{(q)}])\|^2] && (83)\\
&\leq 2\mathbb{E}[\|\boldsymbol{g}_b^{(q)} - \mathbb{E}\boldsymbol{g}_b^{(q)}\|^2 + \|\hat{\boldsymbol{g}}^{(q)} - \mathbb{E}\hat{\boldsymbol{g}}^{(q)}\|^2] && (84)\\
&= 2\mathbb{E}\|\boldsymbol{g}_b^{(q)} - \mathbb{E}\boldsymbol{g}_b^{(q)}\|^2 + 2\mathbb{E}\|\hat{\boldsymbol{g}}^{(q)} - \mathbb{E}\hat{\boldsymbol{g}}^{(q)}\|^2 && (85)
\end{aligned}$$

The first inequality results from Lemma 2, and the second inequality comes from Cauchy inequality.

We bound $\|\hat{\boldsymbol{g}}^{(q)} - \mathbb{E}\hat{\boldsymbol{g}}^{(q)})\|$ as follows.

$$\begin{aligned}
\mathbb{E}\|\hat{\boldsymbol{g}}^{(q)} - \mathbb{E}\hat{\boldsymbol{g}}^{(q)}\| &= \mathbb{E}\|(\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}) + (\boldsymbol{g}^{(q)} - \mathbb{E}\boldsymbol{g}^{(q)}) - \mathbb{E}[\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}]\|^2] && (86)\\
&\leq 3\mathbb{E}[\|\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2 + \|\boldsymbol{g}^{(q)} - \mathbb{E}\boldsymbol{g}^{(q)}\|^2 + \|\mathbb{E}[\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}]\|^2] && (87)\\
&\leq 6\mathbb{E}\|\hat{\boldsymbol{g}}^{(q)} - \boldsymbol{g}^{(q)}\|^2 + 3\mathbb{E}\|\boldsymbol{g}^{(q)} - \mathbb{E}\boldsymbol{g}^{(q)}\|^2 && (88)\\
&\leq 48(\sigma^2 + \kappa^2) + \frac{3\sigma^2}{n-f} && (89)\\
&= (48 + \frac{3\sigma^2}{n-f})\sigma^2 + 48\kappa^2 && (90)
\end{aligned}$$

Apply Equation (90) to Equation (85), we have

$$\mathrm{Var}[s_b^{(q)}] \leq 2\sigma_b^2 + (96 + \frac{6}{n-f})\sigma^2 + 96\kappa^2, \tag{91}$$

where $\sigma_b^2 = \mathbb{E}\|\boldsymbol{g}_b^{(q)} - \mathbb{E}\boldsymbol{g}_b^{(q)}\|^2$ is the variance.

Similar to Equations (72) and (73), we utilize Equations (80) and (91) to bound the expectation and variance of total abnormal score s_b of a byzantine client b.

$$\mathbb{E}[s_b] = \mathbb{E}[\sum_{q=1}^{p} s_b^{(q)}] \geq p(\delta_b - 2\sqrt{2}c\sqrt{2\sigma^2 + 4\kappa^2} - \frac{\sigma}{\sqrt{n-f}}) := C, \tag{92}$$

$$\mathrm{Var}[s_b] = \sum_{q=1}^{p} \mathrm{Var}[s_b^{(q)}] \leq p(2\mathrm{const} + (96 + \frac{6}{n-f})\sigma^2 + 96\kappa^2) := D \tag{93}$$

where $\delta_b = \mathbb{E}\|\boldsymbol{g}_b - \bar{\boldsymbol{g}}\|$. According to Shejwalkar & Houmansadr (2021), σ_b^2 is bounded, i.e., $\sigma_b^2 \leq \mathrm{const}$.

Similarly, we apply Chebyshev's inequality to the abnormal score of a Byzantine client $b \in \mathcal{B}$.

$$\Pr(s_b \geq \mathbb{E}[s_b] - \Delta_b) \geq 1 - \frac{\mathrm{Var}[s_b]}{\Delta_b^2}, \quad b \in \mathcal{B}. \tag{94}$$

Combine Equations (72) to (74), and take $\Delta_h = (C - A)/(1 + \sqrt{D/B})$, we have

$$\begin{aligned}
\Pr(s_h < \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}) &= \Pr(s_h < A + \Delta_h) && (95)\\
&\geq \Pr(s_h < \mathbb{E}[s_h] + \Delta_h) && (96)\\
&\geq 1 - \frac{\mathrm{Var}[s_h]}{\Delta_h^2} && (97)\\
&\geq 1 - \frac{B}{\Delta_h^2} && (98)\\
&= 1 - \frac{(\sqrt{B} + \sqrt{D})^2}{(C - A)^2}, && (99)
\end{aligned}$$

Combine Equations (92) to (94), and take $\Delta_b = (C - A)/(1 + \sqrt{B/D})$, we have

$$\begin{aligned}
\Pr(s_b \geq \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}) &\geq \Pr(s_b > C - \Delta_b) && (100)\\
&\geq \Pr(s_b > \mathbb{E}[s_b] - \Delta) && (101)\\
&\geq 1 - \frac{\mathrm{Var}[s_b]}{\Delta^2} && (102)\\
&\geq 1 - \frac{D}{\Delta_b^2}, && (103)\\
&= 1 - \frac{(\sqrt{B} + \sqrt{D})^2}{(C - A)^2}, && (104)
\end{aligned}$$

Then consider the probability a Byzantine b is selected,

$$\begin{aligned}
\Pr(b \in \tilde{\mathcal{B}}) &= 1 - \Pr(b \notin \tilde{\mathcal{B}}) && (105)\\
&\leq 1 - \Pr(s_h < \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}, \forall h \in \mathcal{H}, s_b > \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}) && (106)\\
&= \Pr(s_h \geq \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}, \forall h \in \mathcal{H} \text{ or } s_b < \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}) && (107)\\
&\leq \sum_{h \in \mathcal{H}} \Pr(s_h \geq \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}) + \Pr(s_b < \frac{\sqrt{D}A + \sqrt{B}C}{\sqrt{B} + \sqrt{D}}) && (108)\\
&\leq (n - f + 1) \cdot \frac{(\sqrt{B} + \sqrt{D})^2}{(C - A)^2} && (109)
\end{aligned}$$

Solve $(n-f+1)\cdot(\sqrt{B}+\sqrt{D})^2/(C-A)^2 \leq \varepsilon$, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{g}_b-\bar{\boldsymbol{g}}\|] \geq (1+\frac{1}{\sqrt{n-f}})\sigma+\kappa+2c\sqrt{2\sigma^2+4\kappa^2} \\ +\sqrt{\frac{n-f+1}{p\varepsilon}}(\sqrt{(4+16c^2+\frac{4}{n-f})\sigma^2+(4+8c^2)\kappa^2} \\ +\sqrt{2\text{const}+(96+\frac{6}{n-f})\sigma^2+96\kappa^2})\end{aligned} \tag{110}$$

which implies that the Byzantine gradients that deviate from the optimal gradient will be filtered by GAS.

Therefore, for all $b\in\tilde{\mathcal{B}}$,

$$\mathbb{E}[\|\boldsymbol{g}_{\tilde{\mathcal{B}}}-\bar{\boldsymbol{g}}\|^2] \leq \mathcal{O}((\kappa^2+\sigma^2)(1+\lambda^2+\frac{1}{n-f})(1+\frac{n-f+1}{p})) := C_1^2 \tag{111}$$

The elimination of ε is due to the sub-Gaussian property of $\boldsymbol{g}_{\tilde{\mathcal{B}}}-\bar{\boldsymbol{g}}$, which comes from the Gaussian property of benign gradients.

Combine Equations (49) and (111), $\mathbb{E}[\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}\|^2]$ is finally bounded as follows.

$$\mathbb{E}[\|\hat{\boldsymbol{g}}-\bar{\boldsymbol{g}}\|^2] \tag{112}$$

$$\leq \frac{|\tilde{\mathcal{H}}|^2}{(n-f)^2}(\sigma^2/\tilde{h}+\kappa^2)+\frac{|\tilde{\mathcal{B}}|^2}{(n-f)^2}C_1^2 \tag{113}$$

$$\leq \frac{(n-2f)^2}{(n-f)^2}(\sigma^2/(n-2f)+\kappa^2)+\frac{f^2}{(n-f)^2}C_1^2, \tag{114}$$

$$= \mathcal{O}((\kappa^2+\sigma^2)(1+\frac{n-f+1}{p})(1+\lambda^2+\frac{1}{n-f})\frac{f^2}{(n-f)^2}) \tag{115}$$

which completes the proof. □

C.1.1. Proof For The Main Proposition

Proof. According to the Lipschitz property of loss function $\mathcal{L}$, we have

$$\mathcal{L}(\boldsymbol{w}^t)-\mathcal{L}(\boldsymbol{w}^{t+1}) \geq \left\langle\nabla\mathcal{L}(\boldsymbol{w}^t),\boldsymbol{w}^t-\boldsymbol{w}^{t+1}\right\rangle-\frac{L}{2}\|\boldsymbol{w}^t-\boldsymbol{w}^{t+1}\|^2. \tag{116}$$

Since $\boldsymbol{w}^t-\boldsymbol{w}^{t+1}=\nabla\mathcal{L}(\boldsymbol{w}^t)+(\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t))$, we can write Equation (116) as follows

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}^t)-\mathcal{L}(\boldsymbol{w}^{t+1}) \geq (\eta-\frac{L}{2}\eta^2)\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2 \\ +(\eta-\frac{L}{2}\eta^2)\langle\nabla\mathcal{L}(\boldsymbol{w}^t),\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\rangle \\ -\frac{L}{2}\eta^2\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2.\end{aligned} \tag{117}$$

Then, we bound inner product term $\langle\nabla\mathcal{L}(\boldsymbol{w}^t),\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\rangle$.

$$|\nabla\mathcal{L}(\boldsymbol{w}^t),\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\rangle| \leq \|\langle\nabla\mathcal{L}(\boldsymbol{w}^t)\|\cdot\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\| \tag{118}$$

$$\leq \frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2+2\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2 \tag{119}$$

Combine Equations (117) and (119) and we have

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}^t)-\mathcal{L}(\boldsymbol{w}^{t+1}) \geq & (\eta-\frac{L}{2}\eta^2)\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2 \\ & +(\eta-\frac{L}{2}\eta^2)\cdot-(\frac{1}{2}\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2+2\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2) \qquad (120)\\ & -\frac{L}{2}\eta^2\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2 \\ = & (\frac{1}{2}\eta-\frac{L}{4}\eta^2)\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2-(2\eta-\frac{L}{2}\eta^2)\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2 \qquad (121)\end{aligned}$$

Take the expectation on both sides of Equation (121), we have

$$\mathbb{E}[\mathcal{L}(\boldsymbol{w}^t)-\mathcal{L}(\boldsymbol{w}^{t+1})] \geq (\frac{1}{2}\eta-\frac{L}{4}\eta^2)\mathbb{E}[\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2]-(2\eta-\frac{L}{2}\eta^2)\mathbb{E}[\|\hat{\boldsymbol{g}}^t-\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2]. \qquad (122)$$

Apply Lemma 3 to Inequality (122) and sum over $t=0,1,\ldots,T-1$, then we have

$$\mathbb{E}[\mathcal{L}(\boldsymbol{w}^0)-\mathcal{L}(\boldsymbol{w}^T)] \geq (\frac{1}{2}\eta-\frac{L}{4}\eta^2)\sum_{t-1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2]-T(\frac{1}{2}\eta-\frac{L}{2}\eta^2)C^2. \qquad (123)$$

where $C^2=\mathcal{O}((\kappa^2+\sigma^2)(1+(n-f+1)/p)(1+\lambda^2+1/(n-f))\frac{f^2}{(n-f)^2})$

Take $\eta=1/2L$, and consider that the loss function is generally non-negative, e.g., cross-entropy loss, ℓ_2 loss,

$$\mathbb{E}[\mathcal{L}(\boldsymbol{w}^0)] \geq \frac{3}{16L}\sum_{t-1}^{T}(\mathbb{E}[\|\nabla\mathcal{L}(\boldsymbol{w}^t)\|^2]-\frac{2}{3}C^2), \qquad (124)$$

which completes the proof. □

C.2. Comparasion of Our Convergence Results with Recent Works

Recent works (Karimireddy et al., 2022; Yu & Kar, 2022; El-Mhamdi et al., 2021; Allen-Zhu et al., 2020) also analyze the convergence of Byzantine-robust FL in the non-IID setting. We compare our convergence results with them.

Similarities. We all guarantee that we can reach an approximate optimal point after a certain number of communication rounds. Moreover, we all admit that convergence in the presence of Byzantine clients may be impossible due to non-IID data, i.e., $\|\nabla\mathcal{L}(\boldsymbol{w})\|$ may never decrease to zero.

The difference from Karimireddy et al. (2022). Our result is orthogonal to one in Karimireddy et al. (2022) since our GAS method is orthogonal to the Bucketing scheme proposed by Karimireddy et al. (2022): we focus on how gradient splitting can alleviate the curse of dimensionality and gradient heterogeneity at the same time while (Karimireddy et al., 2022) considers how partitioning gradients into buckets can help with non-IID data. In fact, we can obtain a better convergence result by combining our method with Bucketing scheme (Karimireddy et al., 2022). The result would enjoy the strengths of both our GAS method and Bucketing scheme: (1) free from the curse of dimensionality; (2) handle gradient heterogeneity that comes from non-IID data; (3) the variance term diminishes when there is no Byzantine client.

The difference from El-Mhamdi et al. (2021). Technically, our result is orthogonal from one in El-Mhamdi et al. (2021). El-Mhamdi et al. (2021) consider how to improve robust AGRs to achieve optimal Byzantine resilience. We focus on how to handle the high-dimension nature of gradients. Moreover, El-Mhamdi et al. (2021) focus on decentralized FL with a server and provide an order optimal upper bound. However, this strong result requires a Byzantine ratio lower than $1/3$. By contrast, we consider a centralized FL setting and only assume the Byzantine ratio to be lower than $1/2$.

The difference from Peng et al. (2022). Peng et al. (2022) consider how client variance reduction and robust AGRs can jointly improve Byzantine resilience. And we concentrate more on gradient dimensions Peng et al. (2022) consider an ideal case where the objective function is strongly convex, while we consider a more general non-convex case.

The difference from Yu & Kar (2022). We considered different settings. We consider standard federated learning with a central server and (Yu & Kar, 2022) considers distributed optimization without a central server. Besides, the convergence analysis is based on different assumptions.

- Yu & Kar (2022) assume the strong convexity of the loss function (Assumption 3) while we do not. This assumption is restrictive since global models are neural networks in practical settings.
- Yu & Kar (2022) do not assume uniformly bounded gradient differences but assume a common global minimizer. Instead, they assume a common minimizer among different agents (clients).

Due to different settings and assumptions, our convergence results are different. Yu & Kar (2022) guarantee almost sure convergence while we ensure that we can approach an approximate optimal parameter. Note that our upper bound matches the lower bound in (Karimireddy et al., 2022).

D. Experiment Setup

D.1. Setup for Main Experiments in Section 7

Data distribution. For CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-12 (Li et al., 2021b), we use Dirichlet distribution to generate non-IID data by following Yurochkin et al. (2019); Li et al. (2021a). In particular, for each client i, we sample $p_i^y \sim \text{Dir}(\beta)$ and allocate a p_i^y proportion of the data of label y to client i, where $\text{Dir}(\beta)$ represents the Dirichlet distribution with a concentration parameter β. We follow Li et al. (2021a) and set the number of clients $n = 50$ and the concentration parameter $\beta = 0.5$ as default.

Other setups. The setups for datasets FEMNIST (Caldas et al., 2018), CIFAR-10 (Krizhevsky et al., 2009), CIFR-100 (Krizhevsky et al., 2009) and ImageNet-12 (Russakovsky et al., 2015) are listed in below Table 8.

Table 8: Default experimental settings for FEMNIST, CIFAR-10, CIFAR-100 and ImageNet-12.

Dataset	FEMNIST	CIFAR-10	CIFAR-100	ImageNet-12
Architecture	CNN (Caldas et al., 2018)	AlexNet (Krizhevsky et al., 2017)	SqueezeNet (Iandola et al., 2016)	ResNet-18 (He et al., 2016)
# Communication rounds	1000	200	400	200
Client sample ratio	0.005	0.1	0.1	0.1
# Local epochs	1	5	1	1
Optimizer	SGD	SGD	SGD	SGD
Batch size	64	64	64	128
Learning rate	0.5	0.1	0.1	0.1
Momentum	0.5	0.5	0.5	0.9
Weight decay	0.0001	0.0001	0.0001	0.0001
Learning rate decay	No	No	No	Reduce to 0.01 after 100-th communication round
Gradient clipping	Yes	Yes	Yes	Yes
Clipping norm	2	2	2	2

The hyperparameters of six attacks: BitFlip (Allen-Zhu et al., 2020), LabelFlip (Allen-Zhu et al., 2020), LIE (Baruch et al., 2019), Min-Max (Shejwalkar & Houmansadr, 2021), Min-Sum (Shejwalkar & Houmansadr, 2021), IPM (Xie et al., 2020), are listed in Table 9 below.

Table 9: The hyperparameters of six attacks. N/A indicates that the attack has no hyperparameters that need to be set.

Attacks	Hyperparameters
BitFlip	N/A
LabelFlip	N/A
LIE	$z = 1.5$
Min-Max	$\gamma_{\text{init}} = 10, \tau = 1 \times 10^{-5}$, $\boldsymbol{\delta}$: coordinate-wise standard deviation
Min-Sum	$\gamma_{\text{init}} = 10, \tau = 1 \times 10^{-5}$, $\boldsymbol{\delta}$: coordinate-wise standard deviation
IPM	# eval $= 2$

The hyperparameters of six robust AGRs: Multi-Krum (Blanchard et al., 2017), Bulyan (Guerraoui et al., 2018), Median (Yin et al., 2018), RFA (Pillutla et al., 2019), DnC (Shejwalkar & Houmansadr, 2021), RBTM (El-Mhamdi et al., 2021), are listed in Table 10 below.

Table 10: The default hyperparameters of the AGRs. N/A indicates that the robust AGR has no hyperparameters that need to be set.

AGRs	Hyperparameters
Multi-Krum	N/A
Bulyan	N/A
Median	N/A
RFA	$T = 3$
DnC	$c = 4, \text{niters} = 1, b = 10000$
RBTM	N/A

E. GAS mitigates the deviation of aggregated gradients

In Section 6, we claim that our GAS approach can reduce the deviation of aggregated gradient $\hat{\boldsymbol{g}}$ from the average of honest gradients $\boldsymbol{g}$. To verify this fact, we compare the deviation of the aggregated gradient of different defenses and their GAS variants in Figure 2. In particular, we use $\|\hat{\boldsymbol{g}} - \boldsymbol{g}\|$, the distance between the aggregated gradient $\hat{\boldsymbol{g}}$ and the average of honest gradients $\boldsymbol{g}$ to measure the deviation degree. As shown in Figure 2, the gradient deviation degree of GAS-enhanced defenses is much lower than their original versions as expected, which validates that our GAS can mitigate the gradient deviation.

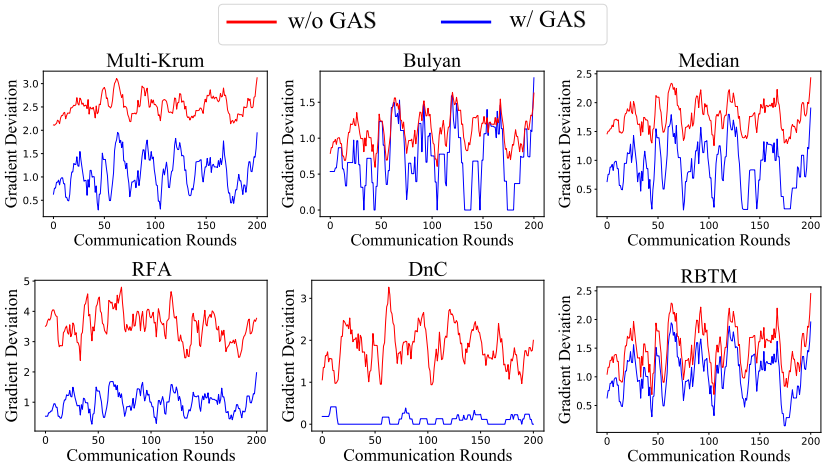


Figure 2: The gradient deviation $\|\hat{\boldsymbol{g}} - \boldsymbol{g}\|$ of six different defenses w/ and w/o GAS under LIE attack on CIFAR-10. The lower the better.