A EXPERIMENT SETUP AND ADDITIONAL EXPERIMENTS

A.1 EXPERIMENT SETUP

A.1.1 GENERAL SETUP

The default experiment setup is listed in Table 5.

Table 5: Default experimental settings for MNIST

Dataset	MNIST
Architecture	CONV-CONV-DROPOUT-FC-DROPOUT-FC
Training objective	Negative log likelihood loss
Evaluation objective	Top-1 accuracy
Batch size	$32 \times$ number of workers
Momentum	0 or 0.9
Learning rate	0.01
LR decay	No
LR warmup	No
# Iterations	600 or 4500
Weight decay	No
Repetitions	3, with varying seeds
Reported metric	Mean test accuracy over the last 150 iterations

By default the hyperparameters of the aggregators are summarized as follows

Aggregators	Hyperparameters
KRUM	N/A
CM	N/A
RFA	$T = 8$
TM	$b = q$
CCLIP	$\tau = \frac{10}{1-\beta}$

A.1.2 CONSTRUCTING DATASETS

The MNIST dataset has 10 classes each with similar amount of samples. In this part, we discuss how to process and distribute MNIST to each workers in order to achieve long-tailness and heterogeneity.

Long-tailness. The long-tailness (*-LT) is achieved by sampling class with exponentially decreasing portions $\gamma \in (0, 1]$. That is, for class $i \in [10]$, we only randomly sample γ^i portion of all samples in class i. We define α as the ratio of the largest class over the smallest class, which can be written as $\alpha = \frac{1}{\gamma^9}$. For example, if $\gamma = 1$, then all classes have same amount of samples and thus $\alpha = 1$; if $\gamma = 0.5$ then $\alpha = 2^9 = 512$. Note that the same procedure has to be applied to the test dataset.

Heterogeneity. Steps to construct IID/non-iid dataset from MNIST dataset

1. Sort the training dataset by its labels.
2. Evenly divide the sorted training dataset into chunks of same size. The number of chunks equals the number of good workers. If the last chunk has fewer samples, we augment it with samples from itself.
3. Shuffle the samples within the same worker.

Heterogeneity + Long-tailness. First transform the training dataset into long-tail dataset, then feed it to the previous procedure to introduce heterogeneity.

About dataset on Byzantine workers. The training set is divided by the number of good workers. So the good workers has to full information of training dataset. The Byzantine worker has access to the whole training dataset.

A.1.3 SETUP FOR EACH EXPERIMENT

In Table 6, we list the hyperparameters for the experiments.

Table 6: Setups for each experiment.

	n	q	momentum	Iters	LT	NonIID
Table 1	24	0	0	4500	$\alpha = 1, \alpha = 500$	iid/ non-iid
Table 2	25	5	0	600	$\alpha = 1$ (balanced)	iid/ non-iid
Table 3	24	0	0	4500	$\alpha = 1, \alpha = 500$	iid/ non-iid
Table 4	25	5	0	600	$\alpha = 1$ (balanced)	iid/ non-iid
Figure 1	25	5	0 / 0.9	600	$\alpha = 1$ (balanced)	non-iid
Figure 2	53	5	0 / 0.9	600	$\alpha = 1$ (balanced)	non-iid
Figure 3	25	5	0 / 0.5 / 0.9 / 0.99	600	$\alpha = 1$ (balanced)	non-iid
Figure 4	25	5	0 / 0.5 / 0.9 / 0.99	1200	$\alpha = 1$ (balanced)	non-iid
Figure 5	20	3	0	1200	$\alpha = 1$ (balanced)	non-iid
Figure 6	20	3	0	3000	$\alpha = 1$ (balanced)	non-iid
Figure 8	24	3	0	1200	$\alpha = 1$ (balanced)	non-iid

IPM Attack in Figure 1 and Figure 2. We set the strength of the attack $\epsilon = 0.1$.

ALIE Attack in in Figure 1. The hyperparameter z for ALIE is computed according to (Baruch et al., 2019)

$$z = \max_{z} \left(\phi(z) < \frac{n - q - s}{n - q} \right)$$

where $s = \lfloor \frac{n}{2} + 1 \rfloor - q$ and ϕ is the cumulative standard normal function. In our setup, the $z \approx 0.25$.

A.1.4 RUNNING ENVIRONMENT

We summarize the running environment of this paper as in Table 7.

Table 7: Runtime hardwares and softwares.

CPU	
Model name	Intel (R) Xeon (R) Gold 6132 CPU @ 2.60 GHz
# CPU(s)	56
NUMA node(s)	2
GPU	
Product Name	Tesla V100-SXM2-32GB
CUDA Version	11.0
PyTorch	
Version	1.7.1

A.2 ADDITIONAL EXPERIMENTS

A.2.1 CLIPPING RADIUS SCALING

The radius τ of CCLIP depends on the norm of good gradients. However, PyTorch implements SGD with momentum using the following formula

$$\boldsymbol{m}_i^t = \beta \boldsymbol{m}_i^{t-1} + \boldsymbol{g}_i(\boldsymbol{x}^{t-1}) \quad \text{for every } i \in \mathcal{G}$$

which may leads to the increase in the gradient norm.

Gradient norms. In Figure 3 we present the averaged gradient norm from all good workers. Here we use CCLIP as the aggregator and $\tau = \frac{10}{1-\beta}$. The norm of gradients are computed before aggregation. Even though the dataset on workers are non-iid, the gradient norms are roughly of same order. The gradient dissimilarity ζ^2 also increases accordingly.

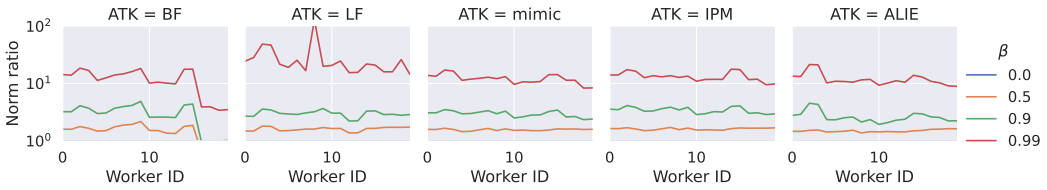

Figure 3: The ratio of norm of good gradients with momentum β over no momentum under different attacks.

Scaled clipping radius. As the gradient norm increases with momentum β, the clipping radius should increase accordingly. In Figure 4 we compare 3 schemes: 1) no scaling ($\tau = 10, \beta = 0$); 2) *linear* scaling $\frac{10}{1-\beta}$; 3) *sqrt* scaling $\frac{10}{\sqrt{1-\beta}}$. The no scaling scheme convergences but slower while with momentum. The linear scaling is usually better than *sqrt* scaling and with bucketing it becomes more stable. However, The scaled clipping radius fails for $\beta = 0.99$ under label flipping attack. This is because the gradient can be very large and ζ^2 dominates. So in general, a linear scaling of clipping radius with momentum $\beta = 0.9$ would be a good choice.

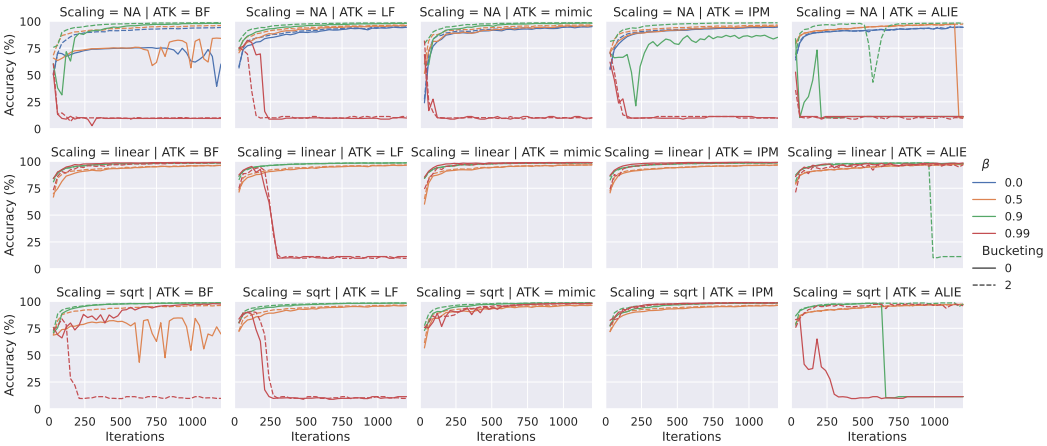

Figure 4: Convergence of CCLIP with $\tau = 10, \frac{10}{1-\beta}, \frac{10}{\sqrt{1-\beta}}$ for $\beta = 0, 0.5, 0.9, 0.99$. The s is the bucketing hyperparameter.

A.2.2 DEMONSTRATION OF EFFECTS OF BUCKETING THROUGH THE SELECTIONS OF KRUM

In the main text we have theoretically show that bucketing helps aggregators alleviate the impact of non-iid. In this section we empirically show that after bucketing aggregators can incorporate updates more evenly from good workers and therefore the problem of non-iid among good workers is less significant. Since KRUM outputs the id of the selected device, it is very convenient to record the frequency of each worker being selected. Since bucketing replicates each worker for s times, we divide their frequencies by s for normalization. From Figure 5, we can see that without bucketing KRUM basically almost always selects updates from Byzantine workers while with larger s, the selection becomes more evenly distributed.

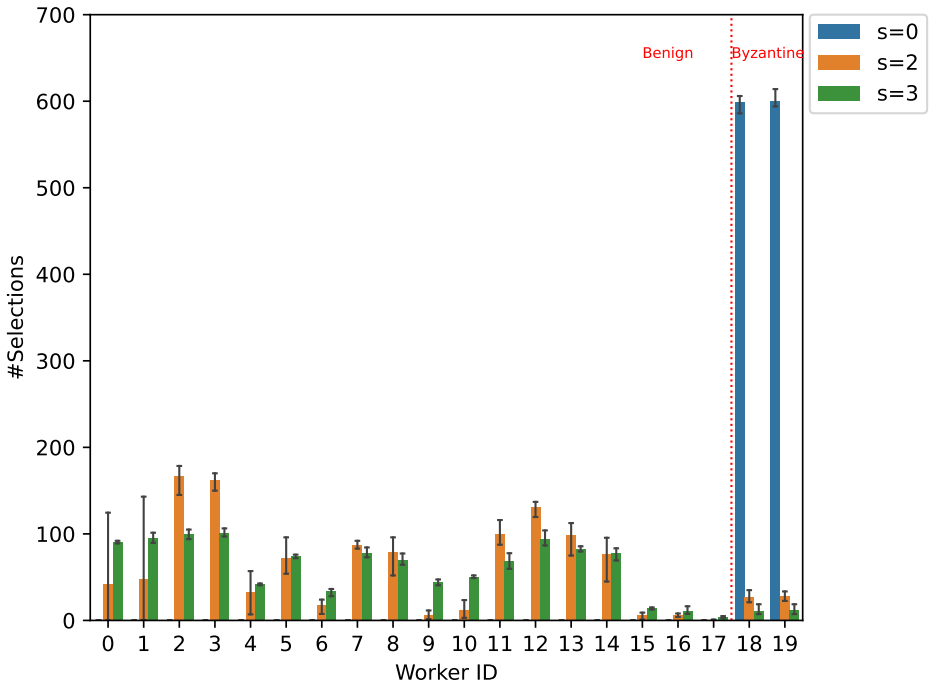

Figure 5: The selected workers of KRUM for bucketing coefficient $s = 0, 2, 3$. There are 20 workers and the last 2 workers (worker id=18,19) are Byzantine with label-flipping attack.

A.2.3 OVERPARAMETERIZATION

The architecture of the neural net used in the experiments can be scaled to make it overparameterized. We add more parameters to the model by multiplying the channels of 2D `Conv` layer and fully connected layer by a factor of 'scale'. So the original model has a scale of 1. We show the training losses decrease faster for overparameterized models in Figure 6. As we can see, the convergence behaviors are similar for different model scales with overparameterized models having smaller training loss despite the existence of Byzantine workers.

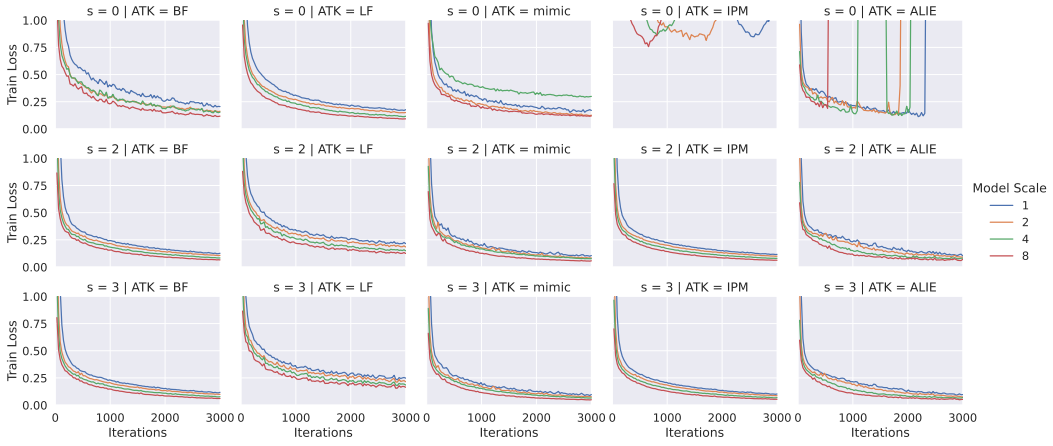

Figure 6: The training loss of models of different levels of overparameterization.

In Figure 7, we explicitly investigate the influence of overparameterization on B^2 defined in (3). As we can see, heterogeneity bound B^2 decreases with increasing level of overparameterization, showcasing how overparameterization minimizes the local objectives in the presence of Byzantine workers. It supports our theory in Section 5.4 that overparameterization can fix the convergence,

making it possible to achieve practical Byzantine-robust learning. The underlying base aggregator is RFA.

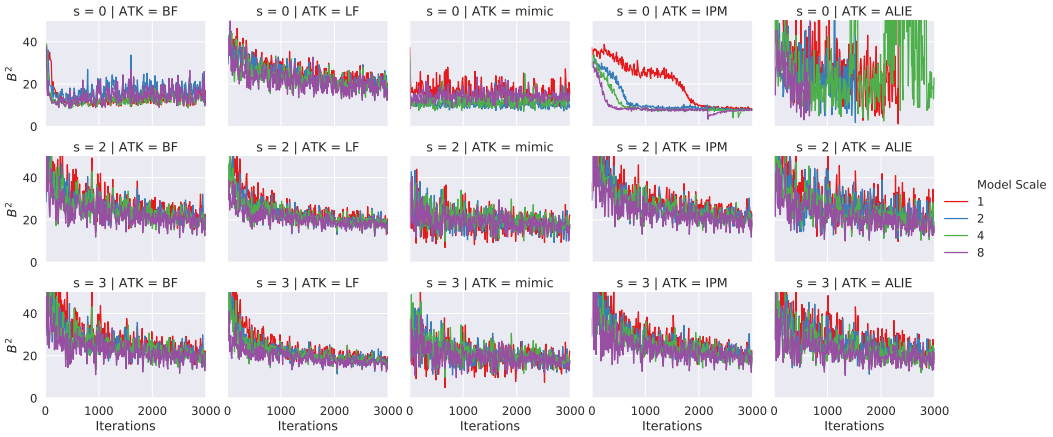

Figure 7: The B^2 in (3) for different levels of overparameterization.

A.2.4 Resampling - variant of bucketing

In the previous version of this work we repeat the gradients for s times and then put sn gradients into n buckets. The results in Figure 8 suggest that the convergence rate of bucketing and resampling is almost the same. So aggregators can benefit more from bucketing as it reduces the number of input gradients and therefore reduce the complexity.

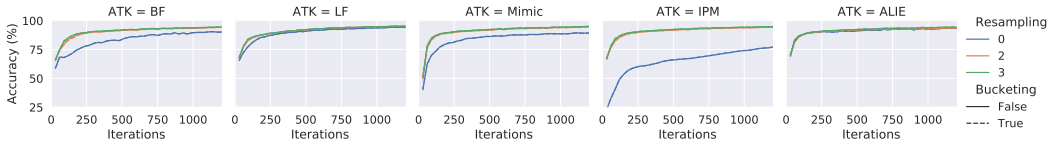

Figure 8: The convergence SGD with bucketing and resampling under different attacks. The underlying aggregator is RFA.

B Implementing the mimic attack

The Section 3.2 describes the idea and formulation of the mimic attack. In this section, we discuss how to pick $i_\star$ and implement the mimic attack efficiently. To pick $i_\star$, we use an initial phase ($\mathcal{I}^0 \approx$ 1 epoch) to compute a direction $\boldsymbol{z}$ of maximum variance of the outputs of the good workers:

$$\boldsymbol{z} = \operatorname*{argmax}_{\|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \Big(\sum_{t\in\mathcal{I}_0} \sum_{i\in\mathcal{G}} (\boldsymbol{x}_i^t - \boldsymbol{\mu})(\boldsymbol{x}_i^t - \boldsymbol{\mu})^\top \Big) \boldsymbol{z} \quad \text{where} \quad \boldsymbol{\mu} = \frac{1}{|\mathcal{G}||\mathcal{I}_0|} \sum_{i\in\mathcal{G}, t\in\mathcal{I}_0} \boldsymbol{x}_i^t \,.$$

Then we pick a worker $i^\star$ to mimic by computing

$$i_\star = \operatorname*{argmax}_{i\in\mathcal{G}} \Big| \sum_{t\in\mathcal{I}_0} \boldsymbol{z}^\top \boldsymbol{x}_i^t \Big| \,.$$

In the following steps, we show how to solve the optimization problem.

First, rewrite the mimic attack in its online version at time $t \in \mathcal{I}_0$

$$\boldsymbol{z}^t = \operatorname*{argmax}_{\|\boldsymbol{z}\|=1} h^t(\boldsymbol{z})$$

where $\boldsymbol{\mu}^t = \frac{1}{|\mathcal{G}|t} \sum_{\tau\le t} \sum_{i\in\mathcal{G}} \boldsymbol{x}_i^\tau$ and

$$h^t(\boldsymbol{z}) = \boldsymbol{z}^\top \left(\sum_{\tau\le t} \sum_{i\in\mathcal{G}} (\boldsymbol{x}_i^\tau - \boldsymbol{\mu}^t)(\boldsymbol{x}_i^\tau - \boldsymbol{\mu}^t)^\top \right) \boldsymbol{z}.$$

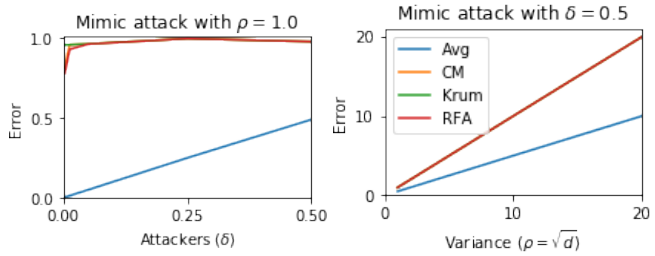

Figure 9: Error with random vectors with variance $\rho^2 = d$ and δ fraction of Byzantine workers imitating a fixed good worker (say worker $1 \in \mathcal{G}$). RFA performs slightly better than CM and KRUM, but all have *higher error* than simply averaging across various settings of δ and ρ.

Thus we can iteratively update $\boldsymbol{\mu}^t$ by

$$\boldsymbol{\mu}^{t+1} = \frac{t}{1+t}\boldsymbol{\mu}^t + \frac{1}{1+t}\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}} \boldsymbol{x}_i^{t+1},$$

and then

$$\begin{aligned}\operatorname*{argmax}_{\|\boldsymbol{z}\|=1} h^{t+1}(\boldsymbol{z}) \approx& \frac{t}{1+t}\boldsymbol{z}^t + \frac{1}{1+t}\operatorname*{argmax}_{\|\boldsymbol{z}\|=1} \boldsymbol{z}^\top \left(\sum_{i\in\mathcal{G}} (\boldsymbol{x}_i^{t+1} - \boldsymbol{\mu}^{t+1})(\boldsymbol{x}_i^{t+1} - \boldsymbol{\mu}^{t+1})^\top\right)\boldsymbol{z} \\ \approx& \frac{t}{1+t}\boldsymbol{z}^t + \frac{1}{1+t}\left(\sum_{i\in\mathcal{G}} (\boldsymbol{x}_i^{t+1} - \boldsymbol{\mu}^{t+1})(\boldsymbol{x}_i^{t+1} - \boldsymbol{\mu}^{t+1})^\top\right)\boldsymbol{z}^t.\end{aligned}$$

The above algorithm corresponds to Oja's method for computing the top eigenvector in a streaming fashion (Oja, 1982). Then, in each subsequent iteration t, we pick

$$i_\star^t = \operatorname*{argmax}_{i\in\mathcal{G}} \boldsymbol{z}^\top \boldsymbol{x}_i^t\,.$$

Example. Each of the good workers $i \in \mathcal{G} \subseteq [n]$ has an input a $\boldsymbol{x}_i \in \{\pm 1\}^d$ where each coordinate is an independent Rademacher random variable. The inputs then have mean $\mathbf{0}$ and variance $\mathbb{E}\|\boldsymbol{x}_i\|^2 = \rho^2 = d$. Now, the Byzantine attackers $j \in \mathcal{B}$ have dual goals: i) escape detection, and ii) increase data imbalance. For this, we propose the following simple passive attack: pick some fixed worker $i_\star \in \mathcal{G}$ (say 1) and every Byzantine worker $j \in \mathcal{B}$ outputs $\boldsymbol{x}_j = \boldsymbol{x}_1$. The attackers cannot be filtered as they imitate an existing good worker, but still can cause imbalance in the data distribution. This serves as the intuition for our attack.

C CONSTRUCTING A ROBUST AGGREGATOR USING BUCKETING

C.1 SUPPORTING LEMMAS

We first start with proving the main bucketing Lemma 1 restated below.

Lemma' 1. *Suppose we are given n independent (but not identical) random vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ such that a good subset $\mathcal{G} \subseteq [n]$ of size at least $|\mathcal{G}| \geq n(1-\delta)$ satisfies:*

$$\mathbb{E}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \leq \rho^2\,, \quad \text{for any fixed } i, j \in \mathcal{G}\,.$$

Define $\bar{\boldsymbol{x}} := \frac{1}{|\mathcal{G}|}\sum_{j\in\mathcal{G}} \boldsymbol{x}_j$ and $m = \lceil n/s \rceil$. Let the outputs after s-bucketing be $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\}$. Then, a subset of the outputs $\tilde{\mathcal{G}} \subseteq \{1, \ldots, m\}$ of size at least $|\tilde{\mathcal{G}}| \geq m(1-\delta s)$ satisfies

$$\mathbb{E}[\boldsymbol{y}_i] = \mathbb{E}[\bar{\boldsymbol{x}}] \quad \text{and} \quad \mathbb{E}\|\boldsymbol{y}_i - \boldsymbol{y}_j\| \leq \rho^2/s \quad \text{for any fixed } i, j \in \tilde{\mathcal{G}}\,.$$

Proof. Let us define the buckets used to compute $\boldsymbol{y}_i$ as

$$B_i := \{\pi(s(i-1)+1), \ldots, \pi(\min\{s\cdot i, n\})\}\,.$$

Recall that for some permutation π over $[n]$ and for every $i = \{1, \dots, m\}$, we defined $m = \lceil n/s \rceil$ and

$$\boldsymbol{y}_i \leftarrow \frac{1}{|B_i|} \sum_{k=(i-1)\cdot s+1}^{\min(n\,,\,i\cdot s)} \boldsymbol{x}_{\pi(k)}\,.$$

Then, define the *new* good set

$$\tilde{\mathcal{G}} = \{i \in [m] \mid B_i \subseteq \mathcal{G}\}$$

$\tilde{\mathcal{G}}$ contains the set of all the resampled vectors which are made up of only good vectors i.e. are uninfluenced by any Byzantine vector. Since $|\mathcal{B}| \leq \delta n$ and each can belong to only 1 bucket, we have that $|\tilde{\mathcal{G}}| \geq (1 - \delta s)m$. Now, for any fixed $i \in \tilde{\mathcal{G}}$, let us look at the conditional expectation over the random permutation π we have

$$\mathbb{E}_\pi[\boldsymbol{y}_i | i \in \tilde{\mathcal{G}}] = \frac{1}{|B_i|} \sum_{k=(i-1)\cdot s+1}^{\min(n\,,\,i\cdot s)} \mathbb{E}_\pi[\boldsymbol{x}_{\pi(k)} | \pi(k) \in \mathcal{G}] = \frac{1}{|\mathcal{G}|} \sum_{j \in \mathcal{G}} \boldsymbol{x}_j = \bar{\boldsymbol{x}}\,.$$

This yields the first part of the lemma. Now we analyze the variance. Thus, we can write $\boldsymbol{y}_i = \frac{1}{s}\sum_{k \in B_i} \boldsymbol{x}_k$. Further, $|B_i| = s$ for any i, and $B_i \subseteq \mathcal{G}$ if $i \in \tilde{\mathcal{G}}$. With this, for any fixed $i, j \in \tilde{\mathcal{G}}$ the variance can be written as

$$\begin{aligned}\mathbb{E}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 &= \mathbb{E}\left\|\frac{1}{s}\sum_{k \in B_i} \boldsymbol{x}_k - \frac{1}{s}\sum_{l \in B_j} \boldsymbol{x}_l\right\|^2 \\ &= \frac{\rho^2}{s}\,.\end{aligned}$$

□

This additional lemma about the maximum expected distance between good workers will also be useful later.

Lemma 7 (maximum good distance)**.** *Suppose we are given the output of bucketing $\boldsymbol{y}_1, \dots, \boldsymbol{y}_m$ which for $m = \lceil n/s \rceil$ satisfy for any fixed $i \in \tilde{\mathcal{G}}$, $\mathbb{E}[\boldsymbol{y}_i] = \boldsymbol{\mu}$ and $\mathbb{E}\|\boldsymbol{y}_i - \boldsymbol{\mu}\|^2 \leq \rho^2/s$. Then, we have*

$$\mathbb{E}\left[\max_{i \in \tilde{\mathcal{G}}} \|\boldsymbol{y}_i - \boldsymbol{\mu}\|^2\right] \leq n\rho^2/s^2\,.$$

Further, there exist instances where

$$\mathbb{E}\left[\max_{i \in \tilde{\mathcal{G}}} \|\boldsymbol{y}_i - \boldsymbol{\mu}\|^2\right] \geq \Omega(n\rho^2/s^2)\,.$$

Proof. For the upper bound, we simply use

$$\mathbb{E}\left[\max_{i \in \tilde{\mathcal{G}}} \|\boldsymbol{y}_i - \boldsymbol{\mu}\|^2\right] \leq \sum_{i \in \tilde{\mathcal{G}}} \mathbb{E}\|\boldsymbol{y}_i - \boldsymbol{\mu}\|^2 \leq m\rho^2/s\,.$$

For the lower bound, let $\tilde{\mathcal{G}} = [m]$ and consider $\boldsymbol{y}_i \sim \tilde{\rho}\sqrt{m}\text{Bern}(p = \frac{1}{m})$. This means $\boldsymbol{y}_i$ is either 0 or $\tilde{\rho}\sqrt{m}$. Further, its variance is clearly bounded by $\tilde{\rho}^2$. Upon drawing m samples, the probability of seeing at least 1 $\boldsymbol{y}_j = \tilde{\rho}\sqrt{m}$ is

$$1 - \Pr(\boldsymbol{y}_i = 0\ \forall i \in [m]) = 1 - (1 - \tfrac{1}{m})^m \geq 1 - 1/e \geq 0.5\,.$$

Thus, with probability at least 0.5 we have

$$\max_{i \in [n]} \|\boldsymbol{y}_i - \boldsymbol{\mu}\|^2 \geq m\tilde{\rho}^2/2\,.$$

This directly proves our lower bound by defining $\tilde{\rho}^2 := \rho^2/s$ and recalling that $m = \lceil n/s \rceil$. Note that this lemma can be tightened if we make stronger assumptions on the noise such as $\mathbb{E}\|\boldsymbol{y}_i - \boldsymbol{\mu}\|^r \leq (\rho/\sqrt{s})^r$ for some large $r \geq 2$. However, we focus on using standard stochastic assumptions ($r = 2$) in this work. □

C.2 PROOFS OF ROBUSTNESS

Let $\{\boldsymbol{y}_1 \ldots, \boldsymbol{y}_m\}$ be the resampled vectors with bucketing using $s = \frac{\delta_{\max}}{\delta}$. By Lemma 1, we have that there is a $\tilde{\mathcal{G}} \subseteq [m]$ of size $|\tilde{\mathcal{G}}| > m(1-\delta_{\max})$ which satisfies for any fixed $i, j \in \tilde{\mathcal{G}}$

$$\mathbb{E}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \leq \frac{\delta\rho^2}{\delta_{\max}} =: \tilde{\rho}^2 \,.$$

This observation will be combined with each of the algorithms to obtain robustness guarantees.

Robustness of KRUM. We now prove that KRUM when combined with bucketing is a robust aggregator. We can rewrite the output of KRUM as the following for $\delta_{\max} = 1/4 - \nu$ for some arbitrarily small positive number $\nu \in (0, 1/4)$:

$$\text{KRUM}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) = \operatorname*{argmin}_{\boldsymbol{y}_i} \min_{|\mathcal{S}|=3m/4} \sum_{j\in\mathcal{S}} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \,.$$

Let $\mathcal{S}^\star$ and $k^\star$ be the quantities which minimize the optimization problem solved by KRUM.

The main difficulty of analyzing KRUM is that even if we succeed in selecting a $k^\star \in \tilde{\mathcal{G}}$, $k^\star$ depends on the sampling. Hence, we **cannot** claim that the error is bounded by $\tilde{\rho}^2$ i.e [3]

$$\mathbb{E}\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_j\|^2 \not\leq \tilde{\rho}^2 \text{ for some fixed } j \in \tilde{\mathcal{G}} \,.$$

This is because the variance is bounded by $\tilde{\rho}^2$ only for a *fixed* i, and not a data dependent $k^\star$. Instead, we will have to rely on Lemma 7 that

$$\mathbb{E}\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_j\|^2 \leq \mathbb{E}\max_{i\in\tilde{\mathcal{G}}} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 \leq m\tilde{\rho}^2 \,.$$

Lemma 7 shows that this inequality is essentially tight and hence relying on it necessarily incurs an extra factor of m which can be very large. Instead, we show an alternate analysis which works for a smaller breakdown point of $\delta_{\max} = 1/4$, but *does not* incur the extra m factor.

For any good input $i \in \tilde{\mathcal{G}}$, we have

$$\begin{aligned} \|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 &\leq 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_i\|^2 + 2\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2 \\ \Rightarrow 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_i\|^2 &\geq \|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 - 2\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2 \,. \end{aligned}$$

Further, for a bad worker $j \in \tilde{\mathcal{B}}$ we can write

$$2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_j\|^2 \geq \|\boldsymbol{y}_j - \bar{\boldsymbol{x}}\|^2 - 2\|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 \,.$$

Combining both and summing over $\mathcal{S}^\star$,

$$\begin{aligned} \sum_{i\in\mathcal{S}^\star} 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_i\|^2 &= \sum_{i\in\tilde{\mathcal{G}}\cap\mathcal{S}^\star} 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_i\|^2 + \sum_{j\in\tilde{\mathcal{B}}\cap\mathcal{S}^\star} 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_j\|^2 \\ &\geq \sum_{j\in\tilde{\mathcal{B}}\cap\mathcal{S}^\star} \|\boldsymbol{y}_j - \bar{\boldsymbol{x}}\|^2 - 2\sum_{i\in\tilde{\mathcal{G}}\cap\mathcal{S}^\star} \|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2 \\ &\qquad + (|\tilde{\mathcal{G}}\cap\mathcal{S}^\star| - 2|\tilde{\mathcal{B}}\cap\mathcal{S}^\star|)\|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 \,. \end{aligned}$$

We can rearrange the above equation as

$$\begin{aligned} \|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 &\leq \frac{1}{(|\tilde{\mathcal{G}}\cap\mathcal{S}^\star| - 2|\tilde{\mathcal{B}}\cap\mathcal{S}^\star|)} \Big(\sum_{i\in\mathcal{S}^\star} 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_i\|^2 + \sum_{i\in\tilde{\mathcal{G}}\cap\mathcal{S}^\star} 2\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2\Big) \\ &\leq \frac{1}{(|\mathcal{S}^\star| - 3|\tilde{\mathcal{B}}|)} \Big(\sum_{i\in\mathcal{S}^\star} 2\|\boldsymbol{y}_{k^\star} - \boldsymbol{y}_i\|^2 + \sum_{i\in\tilde{\mathcal{G}}\cap\mathcal{S}^\star} 2\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2\Big) \\ &\leq \frac{1}{(|\mathcal{S}^\star| - 3|\tilde{\mathcal{B}}|)} \Big(2\min_{k,|\mathcal{S}|=3m/4} \sum_{i\in\mathcal{S}} \|\boldsymbol{y}_k - \boldsymbol{y}_i\|^2 + \sum_{i\in\tilde{\mathcal{G}}} 2\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2\Big). \end{aligned}$$

[3] This issue was incorrectly overlooked in the original analysis of KRUM (Blanchard et al., 2017)

Taking expectation now on both sides yields

$$\mathbb{E}\|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 \leq \frac{4m\tilde{\rho}^2}{|\mathcal{S}^\star| - 3|\tilde{\mathcal{B}}|}\,.$$

Now, recall that we used a bucketing value of $s = \delta_{\max}/\delta$ where for KRUM we have $\delta_{\max} = 1/4 - \nu$. Then, the number of Byzantine workers can be bounded as $|\tilde{\mathcal{B}}| \leq m(1/4 - \nu)$. This gives the final result that

$$\mathbb{E}\|\boldsymbol{y}_{k^\star} - \bar{\boldsymbol{x}}\|^2 \leq \frac{4m\tilde{\rho}^2}{3m/4 - 3(m/4 - \nu m)} = \frac{4\tilde{\rho}^2}{3\nu} \leq \frac{4}{3\nu(1/4 - \nu)}\delta\rho^2\,.$$

Thus, KRUM with bucketing indeed satisfies Definition A with $\delta_{\max} = (1/4 - \nu)$ and $c = 4/(3\nu(1/4 - \nu))$.

Robustness of Geometric median. Geometric median computes the minimum of the following optimization problem

$$\boldsymbol{y}^\star = \operatorname*{argmin}_{\boldsymbol{y}} \sum_{i\in[m]} \|\boldsymbol{y} - \boldsymbol{y}_i\|_2\,.$$

We will adapt Lemma 24 of Cohen et al. (2016), which itself is based on (Minsker et al., 2015). For a good bucket $i \in \tilde{\mathcal{G}}$ and bad bucket $j \in \tilde{\mathcal{B}}$:

$$\|\boldsymbol{y}^\star - \boldsymbol{y}_i\|_2 \geq \|\boldsymbol{y}^\star - \bar{\boldsymbol{x}}\|_2 - \|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|_2 \text{ for } i \in \tilde{\mathcal{G}}, \text{ and}$$
$$\|\boldsymbol{y}^\star - \boldsymbol{y}_j\|_2 \geq \|\boldsymbol{y}_j - \bar{\boldsymbol{x}}\|_2 - \|\boldsymbol{y}^\star - \bar{\boldsymbol{x}}\|_2 \text{ for } j \in \tilde{\mathcal{B}}\,.$$

Summing this over all buckets we have

$$\begin{aligned}
&\sum_{i\in[n]} \|\boldsymbol{y}^\star - \boldsymbol{y}_i\| \geq (|\tilde{\mathcal{G}}| - |\tilde{\mathcal{B}}|)\|\boldsymbol{y}^\star - \bar{\boldsymbol{x}}\| + \sum_{j\in\tilde{\mathcal{B}}}\|\boldsymbol{y}_j - \bar{\boldsymbol{x}}\| - \sum_{i\in\tilde{\mathcal{G}}}\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\| \\
&\Rightarrow \|\boldsymbol{y}^\star - \bar{\boldsymbol{x}}\| \leq \frac{1}{(|\tilde{\mathcal{G}}| - |\tilde{\mathcal{B}}|)}\left(\sum_{i\in[n]}\|\boldsymbol{y}^\star - \boldsymbol{y}_i\| - \sum_{j\in\tilde{\mathcal{B}}}\|\boldsymbol{y}_j - \bar{\boldsymbol{x}}\| + \sum_{i\in\tilde{\mathcal{G}}}\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|\right) \\
&\qquad = \frac{1}{(|\tilde{\mathcal{G}}| - |\tilde{\mathcal{B}}|)}\left(\min_{\boldsymbol{y}}\sum_{i\in[n]}\|\boldsymbol{y} - \boldsymbol{y}_i\| - \sum_{j\in\tilde{\mathcal{B}}}\|\boldsymbol{y}_j - \bar{\boldsymbol{x}}\| + \sum_{i\in\tilde{\mathcal{G}}}\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|\right) \\
&\qquad \leq \frac{2}{(|\tilde{\mathcal{G}}| - |\tilde{\mathcal{B}}|)}\left(\sum_{i\in\tilde{\mathcal{G}}}\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|\right).
\end{aligned}$$

The last step we substituted $\boldsymbol{y} = \bar{\boldsymbol{x}}$. Squaring both sides, expanding, and then taking expectation gives

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{y}^\star - \bar{\boldsymbol{x}}\|^2 &\leq \frac{4}{(|\tilde{\mathcal{G}}| - |\tilde{\mathcal{B}}|)^2}\,\mathbb{E}\left(\sum_{i\in\tilde{\mathcal{G}}}\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|\right)^2 \\
&\leq \frac{4}{(|\tilde{\mathcal{G}}| - |\tilde{\mathcal{B}}|)^2}\left(|\tilde{\mathcal{G}}|\sum_{i\in\tilde{\mathcal{G}}}\mathbb{E}\|\boldsymbol{y}_i - \bar{\boldsymbol{x}}\|^2\right) \\
&\leq \frac{4|\tilde{\mathcal{G}}|^2}{(n - 2|\tilde{\mathcal{B}}|)^2}\tilde{\rho}^2\,.
\end{aligned}$$

Now, recall that we used a bucketing value of $s = \delta_{\max}/\delta$ where for KRUM we have $\delta_{\max} = 1/2 - \nu$. Then, the number of Byzantine workers can be bounded as $|\tilde{\mathcal{B}}| \leq n(1/2 - \nu)$. This gives the final result that

$$\mathbb{E}\|\boldsymbol{y}^\star - \bar{\boldsymbol{x}}\|^2 \leq \frac{4n^2}{4n^2\nu^2}\tilde{\rho}^2 \leq \frac{\tilde{\rho}^2}{\nu^2} \leq \frac{1}{\nu(1/2 - \nu)}\delta\rho^2\,.$$

Thus, geometric median with bucketing indeed satisfies Definition A with $\delta_{\max} = (1/2 - \nu)$ and $c = 1/(\nu(1/2 - \nu))$. Note that geometric median has better theoretical performance than KRUM.

Robustness of Coordinate-wise median. The proof of coordinate-wise median largely follows that of the geometric median. First, we note that we can separate out the objective by coordinates

$$\mathbb{E}\|\text{CM}(\boldsymbol{y}_1,\dots,\boldsymbol{y}_m)-\bar{\boldsymbol{x}}\|^2=\sum_{l=1}^{d}\mathbb{E}(\text{CM}([\boldsymbol{y}_1]_l,\dots,[\boldsymbol{y}_m]_l)-[\bar{\boldsymbol{x}}]_l)^2\,.$$

Then note that, for any fixed coordinate $l\in[d]$ and fixed good worker $i\in\mathcal{G}$, we have $\mathbb{E}([\boldsymbol{y}_i]_l-[\bar{\boldsymbol{x}}]_l)^2\le\mathbb{E}\|\boldsymbol{y}_i-\bar{\boldsymbol{x}}\|^2\le\tilde{\rho}^2$. Thus, we can simply analyze coordinate-wise median as d separate (geometric) median problems on scalars. Thus for any fixed coordinate $l\in[d]$, we have

$$\mathbb{E}(\text{CM}([\boldsymbol{y}_1]_l,\dots,[\boldsymbol{y}_m]_l)-[\bar{\boldsymbol{x}}]_l)^2\le\frac{\tilde{\rho}^2}{\nu^2}\Rightarrow\mathbb{E}\|\text{CM}(\boldsymbol{y}_1,\dots,\boldsymbol{y}_m)-\bar{\boldsymbol{x}}\|^2\le\frac{d\tilde{\rho}^2}{\nu^2}\le\frac{d}{\nu(1/2-\nu)}\delta\rho^2\,.$$

Thus, coordinate-wise median with bucketing indeed satisfies Definition A with $\delta_{\max}=(1/2-\nu)$ and $c=d/(\nu(1/2-\nu))$.

D Lower bounds on non-iid data (Proof of Theorem III)

Our proof builds two sets of functions $\{f_i^1(\boldsymbol{x})\,|\,i\in\mathcal{G}^1\}$ and $\{f_i^2(\boldsymbol{x})\,|\,i\in\mathcal{G}^2\}$ and will show that in the presence of δ-fraction of Byzantine workers, no algorithm can distinguish between them. Since the problems have different optima, this means that the algorithm necessarily has an error on at least one of them.

For the first set of functions, let there be *no* bad workers and hence $\mathcal{G}^1=[n]$. Then, we define the following functions for any $i\in[n]$:

$$f_i^1(x)=\begin{cases}\frac{\mu}{2}x^2-\zeta\delta^{-1/2}x & \text{for } i\in\{1,\dots,\delta n\}\\ \frac{\mu}{2}x^2 & \text{for } i\in\{\delta n+1,\dots,n\}\,.\end{cases}$$

Defining $G:=\zeta\delta^{1/2}$, the average function which is our objective is

$$f^1(x)=\frac{1}{n}\sum_{i=1}^{n}f_i^1(x)=\frac{\mu}{2}x^2-Gx\,.$$

The optimum of our $f^1(x)$ is at $x=\frac{G}{\mu}$. Note that the gradient heterogeneity amongst these workers is bounded as

$$\begin{aligned}\mathbb{E}_{i\sim[n]}\|\nabla f_i^1(x)-\nabla f^1(x)\|^2=&\delta(\zeta\delta^{-1/2}-\zeta\delta^{1/2})^2+(1-\delta)(\zeta\delta^{1/2})^2\\ =&\zeta^2(1-\delta)^2+\zeta^2(1-\delta)\delta=\zeta^2(1-\delta)\le\zeta^2.\end{aligned}$$

Now, we define the second set of functions. Here, suppose that we have δn Byzantine attackers with $\mathcal{B}^2=\{1,\dots,\delta n\}$. Then, the functions of the good workers are defined as

$$f_i^2(x)=\frac{\mu}{2}x^2\text{ for }i\in\mathcal{G}^2=\{\delta n+1,\dots,n\}\,.$$

We then have that the second average objective is

$$f^2(x)=\frac{1}{|\mathcal{G}^2|}\sum_{i\in\mathcal{G}^2}f_i^2(x)=\frac{\mu}{2}x^2\,.$$

Here, we have gradient heterogeneity of 0 and hence is smaller than ζ^2. The optimum of $f^2(x)$ is at $x=0$. The Byzantine attackers simply imitate as if they have the functions:

$$f_j^2(x)=\frac{\mu}{2}x^2-\zeta\delta^{-1/2}x\text{ for }j\in\mathcal{B}^2=\{1,\dots,\delta n\}\,.$$

Note that the set of functions, $\{f_1^1(\boldsymbol{x}),\dots,f_n^1(x)\}$ is exactly identical to the set $\{f_1^2(\boldsymbol{x}),\dots,f_n^2(x)\}$. Only the identity of the good workers $\mathcal{G}^1$ and $\mathcal{G}^2$ are different, leading to different objective functions $f^1(x)$ and $f^2(x)$. However, since the algorithm does not have access to $\mathcal{G}$, its output on each of them is identical i.e.

$$x^{\text{out}}=\text{ALG}(f_1^1(\boldsymbol{x}),\dots,f_n^1(x))=\text{ALG}(f_1^2(\boldsymbol{x}),\dots,f_n^2(x))\,.$$

However, the leads to making a large error in least one of f^1 and f^2 since they have different optimum. This proves a lower bound error of

$$\max_{k\in\{1,2\}} f^k(x^{\text{out}}) - f^k(x^\star) \geq \mu\left(\frac{G}{2\mu}\right)^2 = \frac{\delta\zeta^2}{4\mu}\,.$$

Similarly, we can also bound the gradient norm error bound as

$$\max_{k\in\{1,2\}} \|\nabla f^k(x^{\text{out}})\|^2 \geq \mu^2\left(\frac{G}{2\mu}\right)^2 = \frac{\delta\zeta^2}{4}\,.$$

□

E Convergence of robust optimization on non-iid data (Theorems II and IV)

We will prove a more general convergence theorem which generalizes Theorems II and IV.

Theorem V. *Suppose we are given a $(\delta_{\max}, c)$-ARAgg satisfying Definition A, and n workers of which a subset $\mathcal{G}$ of size at least $|\mathcal{G}| \geq n(1-\delta)$ faithfully follow the algorithm for $\delta \leq \delta_{\max}$. Further, for any good worker $i \in \mathcal{G}$ let f_i be a possibly non-convex function with L-Lipschitz gradients, and the stochastic gradients on each worker be independent, unbiased and satisfy*

$$\mathbb{E}_{\boldsymbol{\xi}_i}\|\boldsymbol{g}_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{x})\|^2 \leq \sigma^2 \text{ and } \mathbb{E}_{j\sim\mathcal{G}}\|\nabla f_j(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \zeta^2 + B^2\|\nabla f(\boldsymbol{x})\|^2\,, \quad \forall \boldsymbol{x}\,,$$

where $\delta \leq 1/(60cB^2)$. Then, for $F^0 := f(\boldsymbol{x}^0) - f^\star$, the output of Algorithm 2 with step-size $\eta = \min\left(\mathcal{O}\left(\sqrt{\frac{LF^0 + c\delta(\zeta^2+\sigma^2)}{TL^2\sigma^2(n^{-1}+c\delta)}}\right), \frac{1}{8L}\right)$ and momentum parameter $\beta = (1-8L\eta)$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 \leq \mathcal{O}\Big(\frac{1}{1-60c\delta B^2}\cdot\Big(c\delta\zeta^2 + \sigma\sqrt{\frac{LF^0}{T}(c\delta + 1/n)} + \frac{LF^0}{T}\Big)\Big)\,.$$

Notes on $\delta \leq 1/(60cB^2)$. In practice the upper bound $\delta \leq 1/(60cB^2)$ does not put an extra strict constraint on δ. This is because one can always decrease B^2 and increase ζ^2 such that $\mathbb{E}_{j\sim\mathcal{G}}\|\nabla f_j(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \zeta^2 + B^2\|\nabla f(\boldsymbol{x})\|^2$ holds for a sufficiently large domain of $\boldsymbol{x}$.

Definitions. Recall our algorithm which performs for $t \geq 2$ the following update with $(1-\beta) = \alpha$:

$$\begin{aligned}\boldsymbol{m}_i^t &= (1-\alpha)\boldsymbol{m}_i^{t-1} + \alpha\boldsymbol{g}_i(\boldsymbol{x}^{t-1}) \quad \text{for every } i \in \mathcal{G}\,,\\ \boldsymbol{x}^t &= \boldsymbol{x}^{t-1} - \eta\text{ARAgg}(\boldsymbol{m}_1^t, \ldots, \boldsymbol{m}_n^t)\,.\end{aligned}$$

For $t = 1$, we use $\alpha = 0$ i.e. $\boldsymbol{m}_i^1 = \boldsymbol{g}_i(\boldsymbol{x}^0)$. Let us also define the actual and ideal momentum aggregates as

$$\boldsymbol{m}^t := \text{ARAgg}(\boldsymbol{m}_1^t, \ldots, \boldsymbol{m}_n^t) \quad \text{and} \quad \bar{\boldsymbol{m}}^t := \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\boldsymbol{m}_i^t\,.$$

We state several supporting lemmas before proving our main Theorem V. We will loosely follow the proof of Byzantine robustness in the iid case by Karimireddy et al. (2021), with the key difference of Lemma 8 which accounts for the non-iid error.

Lemma 8 (Aggregation error)**.** *Given that* ARAgg *satisfies Definition A holds, the error between the ideal average momentum $\bar{\boldsymbol{m}}^t$ and the output of the robust aggregation rule $\boldsymbol{m}^t$ for any $t \geq 2$ can be bounded as*

$$\mathbb{E}\|\boldsymbol{m}^t - \bar{\boldsymbol{m}}^t\|^2 \leq c\delta\rho_t^2\,,$$

where we define for $t \geq 2$

$$\rho_t^2 := 4(6\alpha\sigma^2 + 3\zeta^2) + 4(6\sigma^2 - 3\zeta^2)(1-\alpha)^t + 12\sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2\,.$$

For $t = 1$ we can simplify the bound as $\rho_1^2 := 24c\delta\sigma^2 + 12c\delta\zeta^2 + 12c\delta B^2\|\nabla f(\boldsymbol{x}^0)\|^2$.

Proof. Let $\mathbb{E}_{\xi^t} := \mathbb{E}_{\xi_1^t,\ldots,\xi_n^t,\xi_1^{t-1},\ldots,\xi_n^{t-1},\ldots,\xi_1^0,\ldots,\xi_n^0}$ be the expectation with respect to all of the randomness until time t and let $\mathbb{E}_i := \mathbb{E}_{i\in\mathcal{G}}$ and $\mathbb{E} = \mathbb{E}_{\xi^t}\,\mathbb{E}_i$. Expanding the definition of the worker momentum for a fixed good workers $i \in \mathcal{G}$,

$$\begin{aligned}\mathbb{E}_{\xi^t}\|\boldsymbol{m}_i^t - \mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t]\|^2 &= \mathbb{E}_{\xi^t}\|\alpha(\boldsymbol{g}_i(\boldsymbol{x}^{t-1}) - \nabla f_i(\boldsymbol{x}^{t-1})) + (1-\alpha)(\boldsymbol{m}_i^{t-1} - \mathbb{E}_{\xi^t}[\boldsymbol{m}_i^{t-1}])\|^2\\ &\le \mathbb{E}_{\xi^{t-1}}\|(1-\alpha)(\boldsymbol{m}_i^{t-1} - \mathbb{E}[\boldsymbol{m}_i^{t-1}])\|^2 + \alpha^2\sigma^2\\ &\le (1-\alpha)\,\mathbb{E}_{\xi^{t-1}}\|\boldsymbol{m}_i^{t-1} - \mathbb{E}[\boldsymbol{m}_i^{t-1}]\|^2 + \alpha^2\sigma^2\,.\end{aligned}$$

Unrolling the recursion above yields

$$\mathbb{E}_{\xi^t}\|\boldsymbol{m}_i^t - \mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t]\|^2 \le \left(\sum_{\ell=2}^{t}(1-\alpha)^{t-\ell}\right)\alpha^2\sigma^2 + (1-\alpha)^{t-1}\sigma^2 \le \sigma^2(\alpha + (1-\alpha)^{t-1})\,.$$

Similar computation also shows

$$\mathbb{E}_{\xi^t}\|\bar{\boldsymbol{m}}^t - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\|^2 \le \frac{\sigma^2}{n}(\alpha + (1-\alpha)^{t-1})\,.$$

So far, the expectation was only over the stochasticity of the gradients of worker i. Note that we have $\mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t] = \mathbb{E}_{\xi^{t-1}}[\alpha\nabla f_i(\boldsymbol{x}^{t-1}) + (1-\alpha)\boldsymbol{m}_i^{t-1}]$. Now, suppose we sample i uniformly at random from $\mathcal{G}$. Then,

$$\begin{aligned}&\mathbb{E}_i\left\|\mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t] - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\right\|^2\\ =&\mathbb{E}_i\|\alpha\,\mathbb{E}_{\xi^{t-1}}[\nabla f_i(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^{t-1})] + (1-\alpha)(\mathbb{E}_{\xi^{t-1}}[\boldsymbol{m}_i^{t-1}] - \mathbb{E}_{\xi^{t-1}}[\bar{\boldsymbol{m}}^{t-1}])\|^2\\ \le&(1-\alpha)\,\mathbb{E}_i\|\mathbb{E}_{\xi^{t-1}}[\boldsymbol{m}_i^{t-1}] - \mathbb{E}_{\xi^{t-1}}[\bar{\boldsymbol{m}}^{t-1}]\|^2 + \alpha\,\mathbb{E}_i\|\mathbb{E}_{\xi^{t-1}}[\nabla f_i(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^{t-1})]\|^2\\ \le&(1-\alpha)\,\mathbb{E}_i\|\mathbb{E}_{\xi^{t-1}}[\boldsymbol{m}_i^{t-1}] - \mathbb{E}_{\xi^{t-1}}[\bar{\boldsymbol{m}}^{t-1}]\|^2 + \alpha\,\mathbb{E}_i\,\mathbb{E}_{\xi^{t-1}}\|\nabla f_i(\boldsymbol{x}^{t-1}) - \nabla f(\boldsymbol{x}^{t-1})\|^2\\ \le&(1-\alpha)\,\mathbb{E}_i\|\mathbb{E}_{\xi^{t-1}}[\boldsymbol{m}_i^{t-1}] - \mathbb{E}_{\xi^{t-1}}[\bar{\boldsymbol{m}}^{t-1}]\|^2 + \alpha\zeta^2 + \alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2\end{aligned}$$

where the second inequality uses the probabilistic Jensen's inequality. Note that here we only get α instead of α^2 as before. This is because the randomness in the sampling i of $\nabla f_i(\boldsymbol{x}^{t-1})$ is not independent of the second term $\mathbb{E}_{\xi^{t-1}}[\boldsymbol{m}_i^{t-1}] - \mathbb{E}_{\xi^{t-1}}[\bar{\boldsymbol{m}}^{t-1}]$. Expanding this we get,

$$\mathbb{E}_i\left\|\mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t] - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\right\|^2 \le \zeta^2(1-(1-\alpha)^t) + \sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2\,.$$

We can combine all three bounds above as

$$\begin{aligned}\mathbb{E}\|\boldsymbol{m}_i^t - \bar{\boldsymbol{m}}^t\|^2 &\le 3\,\mathbb{E}\|\boldsymbol{m}_i^t - \mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t]\|^2 + 3\,\mathbb{E}\|\bar{\boldsymbol{m}}^t - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\|^2 + 3\,\mathbb{E}_i\|\mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t] - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\|^2\\ &= 3\,\mathbb{E}_i\,\mathbb{E}_{\xi^t}\|\boldsymbol{m}_i^t - \mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t]\|^2 + 3\,\mathbb{E}_{\xi^t}\|\bar{\boldsymbol{m}}^t - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\|^2 + 3\,\mathbb{E}_i\|\mathbb{E}_{\xi^t}[\boldsymbol{m}_i^t] - \mathbb{E}_{\xi^t}[\bar{\boldsymbol{m}}^t]\|^2\\ &\le (6\alpha\sigma^2 + 3\zeta^2) + (6\sigma^2 - 3\zeta^2)(1-\alpha)^t + 3\sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2\,.\end{aligned}$$

Therefore for $i, j \in \mathcal{G}$

$$\begin{aligned}\mathbb{E}\|\boldsymbol{m}_i^t - \boldsymbol{m}_j^t\|^2 &\le 2\,\mathbb{E}\|\boldsymbol{m}_i^t - \bar{\boldsymbol{m}}^t\|^2 + 2\,\mathbb{E}\|\boldsymbol{m}_j^t - \bar{\boldsymbol{m}}^t\|^2\\ &\le 4(6\alpha\sigma^2 + 3\zeta^2) + 4(6\sigma^2 - 3\zeta^2)(1-\alpha)^t + 12\sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2\,.\end{aligned}$$

Recall that the right hand side was defined to be ρ_t^2. Using Definition A, we can show that the output of the aggregation rule ARAGG satisfies the condition in the lemma. □

One major caveat in the above lemma is that here ρ^2 cannot be known to the robust aggregation since it involves $\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2$ whose value we do not have access to. However, this does not present a hurdle to *agnostic* aggregation rules which are automatically adaptive to the value of ρ^2. Deriving a similarly provable adaptive clipping method is a very important open problem.

Lemma 9 (Descent bound)**.** *For any $\alpha \in [0,1]$ for $t \geq 2$, $\eta \leq \frac{1}{L}$, and an L-smooth function f we have for any $t \geq 1$*

$$\mathbb{E}[f(\boldsymbol{x}^t)] \leq f(\boldsymbol{x}^{t-1}) - \frac{\eta}{2}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 + \eta\,\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 + \eta\,\mathbb{E}\|\boldsymbol{m}^t - \bar{\boldsymbol{m}}^t\|^2\,.$$

where $\bar{\boldsymbol{e}}^t := \bar{\boldsymbol{m}}^t - \nabla f(\boldsymbol{x}^{t-1})$.

Proof. By the smoothness of the function f and the server update,

$$\begin{aligned}
f(\boldsymbol{x}^t) &\leq f(\boldsymbol{x}^{t-1}) - \eta\langle\nabla f(\boldsymbol{x}^{t-1}), \boldsymbol{m}^t\rangle + \frac{L\eta^2}{2}\|\boldsymbol{m}^t\|^2 \\
&\leq f(\boldsymbol{x}^{t-1}) - \eta\langle\nabla f(\boldsymbol{x}^{t-1}), \boldsymbol{m}^t\rangle + \frac{\eta}{2}\|\boldsymbol{m}^t\|^2 \\
&= f(\boldsymbol{x}^{t-1}) + \frac{\eta}{2}\|\boldsymbol{m}^t - \nabla f(\boldsymbol{x}^{t-1})\|^2 - \frac{\eta}{2}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 \\
&= f(\boldsymbol{x}^{t-1}) + \frac{\eta}{2}\|\boldsymbol{m}^t \pm \bar{\boldsymbol{m}}^t - \nabla f(\boldsymbol{x}^{t-1})\|^2 - \frac{\eta}{2}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 \\
&\leq f(\boldsymbol{x}^{t-1}) + \eta\|\bar{\boldsymbol{e}}^t\|^2 + \eta\|\boldsymbol{m}^t - \bar{\boldsymbol{m}}^t\|^2 - \frac{\eta}{2}\|\nabla f(\boldsymbol{x}^{t-1})\|^2\,.
\end{aligned}$$

Here we used the identities that $-2ab = (a-b)^2 - a^2 - b^2$, and Young's inequality that $(a+b)^2 \leq (1+\gamma)a^2 + (1+\frac{1}{\gamma})b^2$ for any positive constant $\gamma \geq 0$ (here we used $\gamma = 1$). Taking conditional expectation on both sides yields the lemma. □

Lemma 10 (Error bound)**.** *Using any constant momentum and step-sizes such that $1 \geq \alpha \geq 8L\eta$ for $t \geq 2$, we have for an L-smooth function f that $\mathbb{E}\|\bar{\boldsymbol{e}}^1\|^2 \leq \frac{2\sigma^2}{n}$ and for $t \geq 2$*

$$\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 \leq (1 - \tfrac{2\alpha}{5})\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \tfrac{\alpha}{10}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-2})\|^2 + \tfrac{\alpha}{10}\,\mathbb{E}\|\boldsymbol{m}^{t-1} - \bar{\boldsymbol{m}}^{t-1}\|^2 + \alpha^2\tfrac{2\sigma^2}{n}\,.$$

Proof. Let us define $\bar{\boldsymbol{g}}(\boldsymbol{x}) := \frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{G}}\boldsymbol{g}_i(\boldsymbol{x})$. This implies that

$$\mathbb{E}\|\bar{\boldsymbol{g}}(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 \leq \frac{\sigma^2}{|\mathcal{G}|} \leq \frac{2\sigma^2}{n}\,.$$

Then by definition of $\bar{\boldsymbol{m}}$, we can expand the error as:

$$\begin{aligned}
\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 &= \mathbb{E}\|\bar{\boldsymbol{m}}^t - \nabla f(\boldsymbol{x}^{t-1})\|^2 \\
&= \mathbb{E}\|\alpha\bar{\boldsymbol{g}}(\boldsymbol{x}^{t-1}) + (1-\alpha)\bar{\boldsymbol{m}}^{t-1} - \nabla f(\boldsymbol{x}^{t-1})\|^2 \\
&\leq \mathbb{E}\|\alpha\nabla f(\boldsymbol{x}^{t-1}) + (1-\alpha)\bar{\boldsymbol{m}}^{t-1} - \nabla f(\boldsymbol{x}^{t-1})\|^2 + \frac{2\alpha^2\sigma^2}{n} \\
&= (1-\alpha)^2\,\mathbb{E}\|(\bar{\boldsymbol{m}}^{t-1} - \nabla f(\boldsymbol{x}^{t-2})) + (\nabla f(\boldsymbol{x}^{t-2}) - \nabla f(\boldsymbol{x}^{t-1}))\|^2 + \frac{2\alpha^2\sigma^2}{n} \\
&\leq (1-\alpha)(1+\tfrac{\alpha}{2})\,\mathbb{E}\|(\bar{\boldsymbol{m}}^{t-1} - \nabla f(\boldsymbol{x}^{t-2}))\|^2 \\
&\qquad + (1-\alpha)(1+\tfrac{2}{\alpha})\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-2}) - \nabla f(\boldsymbol{x}^{t-1})\|^2 + \frac{2\alpha^2\sigma^2}{n} \\
&\leq (1-\tfrac{\alpha}{2})\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \tfrac{2L^2}{\alpha}\,\mathbb{E}\|\boldsymbol{x}^{t-2} - \boldsymbol{x}^{t-1}\|^2 + \frac{2\alpha^2\sigma^2}{n} \\
&= (1-\tfrac{\alpha}{2})\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \tfrac{2L^2\eta^2}{\alpha}\,\mathbb{E}\|\boldsymbol{m}^{t-1}\|^2 + \frac{2\alpha^2\sigma^2}{n} \\
&\leq (1-\tfrac{\alpha}{2})\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \tfrac{6L^2\eta^2}{\alpha}\|\bar{\boldsymbol{e}}^{t-1}\|^2 \\
&\qquad + \tfrac{6L^2\eta^2}{\alpha}\,\mathbb{E}\|\boldsymbol{m}^{t-1} - \bar{\boldsymbol{m}}^{t-1}\|^2 + \tfrac{6L^2\eta^2}{\alpha}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-2})\|^2 + \frac{2\alpha^2\sigma^2}{n}\,.
\end{aligned}$$

Our choice of the momentum parameter α implies $64L^2\eta^2 \leq \alpha^2$ and yields the lemma statement.

□

Proof of Theorem V. Scale the error bound Lemma 10 by $\frac{5\eta}{2\alpha}$ and add it to the descent bound Lemma 9 taking expectations on both sides to get for $t \geq 2$

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}^t)] + \tfrac{5\eta}{2\alpha}\,\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 \leq \mathbb{E}[f(\boldsymbol{x}^{t-1})] - \tfrac{\eta}{2}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 + \eta\,\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 + \eta\,\mathbb{E}\|\boldsymbol{m}^t - \bar{\boldsymbol{m}}^t\|^2 + \\ \tfrac{5\eta}{2\alpha}\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 - \eta\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \tfrac{\eta}{4}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-2})\|^2 \\ + \tfrac{\eta}{4}\,\mathbb{E}\|\boldsymbol{m}^{t-1} - \bar{\boldsymbol{m}}^{t-1}\|^2 + 5\eta\alpha\frac{\sigma^2}{n}.\end{aligned}$$

Now, let use the aggregation error Lemma 8 to bound $\mathbb{E}\|\boldsymbol{m}^{t-1} - \bar{\boldsymbol{m}}^{t-1}\|^2$ and $\mathbb{E}\|\boldsymbol{m}^t - \bar{\boldsymbol{m}}^t\|^2$ in the above expression to get

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{x}^t)] + \tfrac{5\eta}{2\alpha}\,\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 \leq \mathbb{E}[f(\boldsymbol{x}^{t-1})] - \tfrac{\eta}{2}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 + \eta\,\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 \\ + \tfrac{5\eta}{2\alpha}\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 - \eta\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \tfrac{\eta}{4}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-2})\|^2 + 5\eta\alpha\tfrac{\sigma^2}{n} \\ + 5\eta c\delta((6\alpha\sigma^2 + 3\zeta^2) + 6\sigma^2(1-\alpha)^{t-2}) \\ + \eta c\delta\left(3\sum_{k=1}^{t-1}(1-\alpha)^{t-1-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2\right) \\ + 4\eta c\delta\left(3\sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2\right).\end{aligned}$$

Let us define $S_t := \sum_{k=1}^{t}(1-\alpha)^{t-k}\alpha B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{k-1})\|^2$. Then, S_t satisfies the recursion:

$$\tfrac{1}{\alpha}S_t = (\tfrac{1}{\alpha} - 1)S_{t-1} + B^2\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2\,.$$

Adding $\frac{3\eta c\delta(\frac{5}{\alpha}-4)}{\alpha}S_t$ on both sides to the bound above and rearranging gives the following for $t \geq 2$

$$\begin{aligned}&\underbrace{\mathbb{E}\ f(\boldsymbol{x}^t) - f^\star + (\tfrac{5\eta}{2\alpha} - \eta)\,\mathbb{E}\|\bar{\boldsymbol{e}}^t\|^2 + \frac{\eta}{4}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 + \frac{3\eta c\delta(\frac{5}{\alpha}-4)}{\alpha}S_t}_{=:\mathcal{E}_t} \\ &\quad\leq \underbrace{\mathbb{E}\ f(\boldsymbol{x}^{t-1}) - f^\star + (\tfrac{5\eta}{2\alpha} - \eta)\,\mathbb{E}\|\bar{\boldsymbol{e}}^{t-1}\|^2 + \frac{\eta}{4}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-2})\|^2 + \frac{3\eta c\delta(\frac{5}{\alpha}-4)}{\alpha}S_{t-1}}_{=:\mathcal{E}_{t-1}} \\ &\qquad(-\tfrac{\eta}{4} + 15\eta c\delta B^2)\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 \\ &\qquad+ \frac{5\eta\alpha}{n}\sigma^2 + 5\eta c\delta\big((6\alpha\sigma^2 + 3\zeta^2) + 6\sigma^2(1-\alpha)^{t-2}\big) \\ &\quad\leq \mathcal{E}_{t-1} - \frac{\eta}{4}(1 - 60c\delta B^2)\,\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 \\ &\qquad+ \underbrace{5\eta\alpha\sigma^2\big(\tfrac{1}{n} + 6c\delta(1 + \tfrac{1}{\alpha}(1-\alpha)^{t-2})\big) + 15\eta c\delta\zeta^2}_{=:\eta\xi_{t-1}^2}\,.\end{aligned}$$

Further, specializing the descent bound Lemma 9 and error bound Lemma 10 for $t = 1$ we have

$$\begin{aligned}\mathcal{E}_1 &= \mathbb{E}\ f(\boldsymbol{x}^1) - f^\star + \frac{3\eta}{2}\,\mathbb{E}\|\bar{\boldsymbol{e}}^1\|^2 + \frac{\eta}{4}\,\mathbb{E}\|\nabla f(\boldsymbol{x}^0)\|^2 + 3\eta c\delta B^2(\frac{5}{\alpha} - 4)\|\nabla f(\boldsymbol{x}^0)\|^2 \\ &\leq f(\boldsymbol{x}^0) - f^\star + \frac{5\eta}{2}\,\mathbb{E}\|\bar{\boldsymbol{e}}^1\|^2 - \frac{\eta}{4}(1 - 60c\delta B^2)\,\mathbb{E}\|\nabla f(\boldsymbol{x}^0)\|^2 + \eta\,\mathbb{E}\|\boldsymbol{m}_1 - \bar{\boldsymbol{m}}_1\|^2 \\ &\leq f(\boldsymbol{x}^0) - f^\star - \frac{\eta}{4}(1 - 60c\delta B^2)\,\mathbb{E}\|\nabla f(\boldsymbol{x}^0)\|^2 + \frac{5\eta\sigma^2}{n} + 12c\delta\eta(2\sigma^2 + \zeta^2 + B^2\|\nabla f(\boldsymbol{x}^0)\|^2) \\ &= f(\boldsymbol{x}^0) - f^\star - \frac{\eta}{4}(1 - 60c\delta B^2)\,\mathbb{E}\|\nabla f(\boldsymbol{x}^0)\|^2 + \eta\xi_0^2\,.\end{aligned}$$

Above, we defined $\xi_0^2 := \frac{5\sigma^2}{n} + 12c\delta(2\sigma^2 + \zeta^2 + B^2\|\nabla f(\boldsymbol{x}^0)\|^2)$. Summing over t from 2 until T, again rearranging our recursion for $\mathcal{E}_t$, and adding $(1 - 3c\delta B^2)\,\mathbb{E}\|\nabla f(\boldsymbol{x}^0)\|^2$ on both sides gives

$$
\begin{aligned}
(1 - 60c\delta B^2)\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 &\leq \frac{4(f(\boldsymbol{x}^0) - f^\star)}{\eta T} + \frac{1}{T}\sum_{t=1}^{T} 4\xi_{t-1}^2 \\
&= \frac{4(f(\boldsymbol{x}^0) - f^\star)}{\eta T} + \frac{4\xi_0^2}{T} \\
&\quad + \frac{1}{T}\sum_{t=2}^{T} 20\alpha\sigma^2\left(\tfrac{1}{n} + 6c\delta(1 + \tfrac{1}{\alpha}(1-\alpha)^{t-2})\right) \\
&\quad + \frac{1}{T}\sum_{t=2}^{T} 60c\delta\zeta^2 \\
&\leq \frac{4(f(\boldsymbol{x}^0) - f^\star)}{\eta T} + \frac{4\xi_0^2}{T} + 60c\delta\zeta^2 \\
&\quad + 20\alpha\sigma^2\left(\tfrac{1}{n} + 6c\delta\right) + \frac{120c\delta\sigma^2}{\alpha T} \\
&= \frac{4(f(\boldsymbol{x}^0) - f^\star)}{\eta T} + \frac{120c\delta\sigma^2}{\eta 8LT} + \eta 160L\sigma^2\left(\tfrac{1}{n} + 6c\delta\right) \\
&\quad + \frac{4\xi_0^2}{T} + 60c\delta\zeta^2.
\end{aligned}
$$

The last equality substituted the value of $\alpha = 8L\eta$. Next, let us use the appropriate step-size of

$$
\eta = \min\left(\sqrt{\frac{4(f(\boldsymbol{x}_0) - f^\star) + \frac{15c\delta}{L}(\zeta^2 + 2\sigma^2)}{T(160L\sigma^2)\left(\frac{1}{n} + 6c\delta\right)}}, \frac{1}{8L}\right).
$$

This gives the following final rate of convergence:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}^{t-1})\|^2 \\
&\qquad \leq \frac{1}{1 - 60c\delta B^2}\cdot\left(60c\delta\zeta^2 + \sqrt{\frac{160L\sigma^2\left(\frac{1}{n} + 6c\delta\right)}{T}}\cdot\sqrt{4(f(\boldsymbol{x}_0) - f^\star) + \tfrac{15c\delta}{L}(\zeta^2 + 2\sigma^2)}\right. \\
&\qquad\qquad + \frac{32L(f(\boldsymbol{x}^0) - f^\star)}{T} + \frac{15c\delta\sigma^2}{T} \\
&\qquad\qquad \left. + \frac{\frac{20\sigma^2}{n} + 12c\delta(2\sigma^2 + \zeta^2 + B^2\|\nabla f(\boldsymbol{x}^0)\|^2)}{T}\right).
\end{aligned}
$$

□

F UPDATES WITH RESPECT TO REVIEWS.

F.1 ADDITIONAL RELATED WORK

In this section, we add comments on works which are very close to this paper.

- Li et al. (2019) propose RSA for Byzantine-resilient distributed learning on heterogeneous data. They introduce an additional ℓ_p-norm regularized term to the objective to penalize the difference between local iterates and server iterate and show convergence of RSA for strongly convex local objectives and penalized term. However, RSA cannot defend the state-of-the-art attacks like (Baruch et al., 2019; Xie et al., 2020) as they didn't utilize the temporal information. Compared to RSA, our method does not assume strongly convexity and we consider more general cost functions with no explicit regularized term. In addition, our method is shown to defend the state-of-the-art attacks.

- (Yang & Li, 2021) is a parallel work which uses buffer for asynchronous Byzantine-resilient training (BASGD). The buffer and bucketing are similar techniques with vastly different motivations. The key difference between buffer and bucketing is that buffer is only reassigned when timer exceeds a threshold while bucketing reshuffles in each iteration. Therefore, buffer does not guarantee that partial aggregated gradients are identically distributed while bucketing does. In addition, our theoretical analysis does not require bounded gradient assumption $\|\nabla f(\boldsymbol{x})\| \leq D$ for all $\boldsymbol{x}$.
- Wu et al. (2020) uses ByrdSAGA for Byzantine-resilient SAGA approach for distributed learning. The key differences between ByrdSAGA and our work are as follows. In our setting, there are two sources of variances of the gradients - intra-worker variance σ^2 and inter-client variance ζ^2. We show that simply using worker momentum suffices to tackle the former and handling the latter ζ^2 is the main challenge. ByrdSAGA assumes that each worker only has finite data points as opposed to the stochastic setting we consider. Hence they can use SAGA on the worker in place of worker momentum to reduce the intra-client variance σ^2. The effect of ζ^2 (which is our main focus) remains unaffected.

  Further, they consider the strongly convex setting whereas we analyze non-convex functions. Ignoring μ for sake of comparison, their Theorem 1 proves convergence to a radius of $\Delta_1 = O(\zeta^2)$ since always $C_\alpha \geq 2$. Thus, their rates are similar to (Acharya et al., 2021) and do not converge to the optimum even when $\delta = 0$. In contrast, our Theorem II proves convergence to a radius of $O(\delta\zeta^2)$. We believe our improved handling of ζ^2 can be combined with their usage of SAGA/variance reduction to yield even faster rates. This we leave for future work.