Simple Minimax Optimal Byzantine Robust Algorithm for Nonconvex Objectives with Uniform Gradient Heterogeneity

Tomoya Murata,[1] Kenta Niwa,[2] Takumi Fukami,[3] Iifan Tyou[3]

1. NTT DATA Mathematical Systems, Inc. / 2. NTT Communication Science Laboratories, NTT Corporation / 3. NTT Social Information Laboratories, NTT Corporation

Overview

Byzantine tolerant nonconvex **Federated Learning** (**FL**) is focused:

- Simple Byzantine robust method combined with **Screening** and **momentum** is proposed.
- Theoretically, **minimax optimal rate** $\boldsymbol{O}(\delta^2\zeta^2)$ is achieved for objectives with ζ**-uniform gradient heterogeneity**.
- Empirically, our method enjoys better performances over various Byzantine attacks than existing methods.

Problem Settings

The following nonconvex minimization is considered:

$$\min_{x\in\mathbb{R}^d} f(x) \coloneqq \frac{1}{|G|}\sum_{i\in G} f_i(x)\,, where\ G \subset [n]\ is\ the\ set\ of\ non\ Byzantine\ clients.$$

f_i is typically the empirical or excess risk on local dataset associated with client i.

In **FL**, $f_i \neq f_j$ due to the **heterogeneity** of local datasets.

Goal of this study:

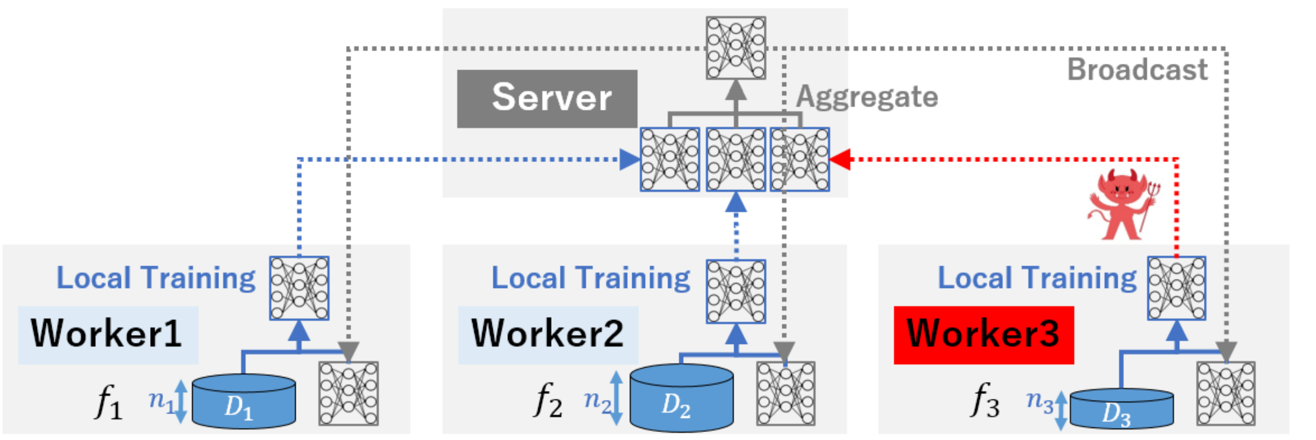
Given input $\{f_i\}_{i=1}^n$ containing Byzantine clients, find x satisfying $\|\nabla f(x)\|^2 \le \varepsilon$ with being ε as small as possible.

Motivations:

Some clients may **behave abnormally** in federated learning.

- Hardware crashes
- Message corruption
- Poisoned data
- Malicious false information

Robustness against abnormal behaviors is important!



Theoretical Assumptions:

Used in our analysis

A1. L-smoothness of f_i.

A2. Existence of global minima x_*.

A3. Sub-Gaussian tail bounds of minibatch stochastic gradient:

$$\forall x \in \mathbb{R}^d, \forall s \ge 0\colon\ \mathbb{P}(\|g_i - \nabla f_i(x)\| \ge s) \le 2\exp\left(-\frac{s^2}{2\sigma^2}\right).$$

A4. G-Lipschitzness of per-sample loss.

A5. ζ-**uniform** gradient heterogeneity:

$$\max_{i\in G}\left\|\nabla f_i(x) - \nabla f_j(x)\right\|^2 \le \zeta^2$$

A5'. ζ-**mean** gradient heterogeneity:

$$\left(\frac{1}{|G|}\right)\sum_{i\in G}\left\|\nabla f_i(x) - \nabla f_j(x)\right\|^2 \le \zeta^2$$

$C_{UH}(\zeta) \coloneqq \{\{f_i\}_{i\in G} \mid A1-A4, A5\ \ hold\}$

$C_{MH}(\zeta) \coloneqq \{\{f_i\}_{i\in G} \mid A1-A4, A5'\ hold\}$

$\boldsymbol{C_{UH}(\zeta) \subset C_{MH}(\zeta)}$

Review of Existing Algorithms

Traditional Robust Aggregation:

- Coordinate Median (CM)
- Trimmed Mean
- KRUM
- Geometric Median (RFA)

Bucketing:

A wrapper technique applicable to any robust aggregation.

Given input $\{x_i\}_{i=1}^n$, create $\lceil n/s\rceil$ random buckets, and apply a robust agg. to new input $\{y_i\}_{i=1}^{\lceil n/s\rceil}$, where y_i is the average of the i-th bucket.

Centered Clipping (CClip):

Given momentum $\{m_i\}_{i=1}^n$ and initial guess v of the ideal agg., we use,

$$v + \tfrac{1}{n}\textstyle\sum_{i=1}^n \min\left\{1, \frac{\tau}{\|m_i - v\|}\right\}(m_i - v).$$

Theoretical Results:

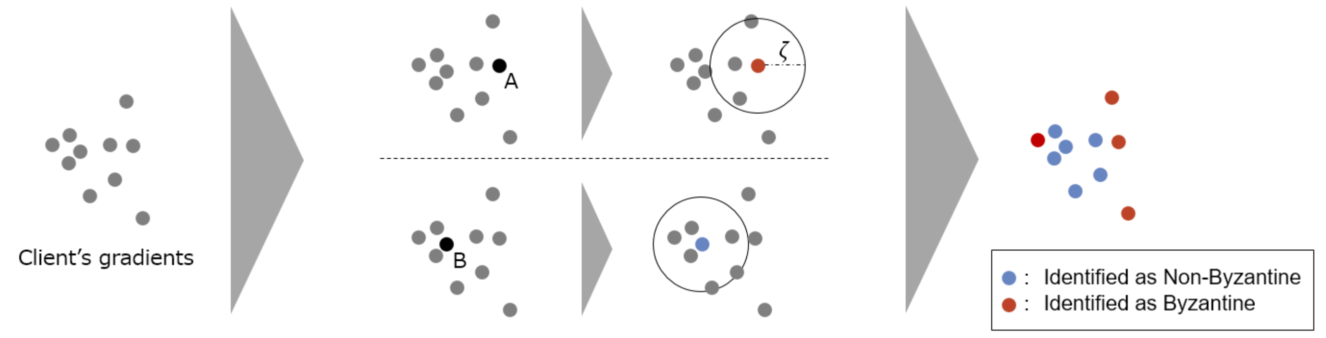
Given Byzantine fraction $\delta < 0.5$, Bucketing and CClip achieves $O(\delta\zeta^2)$ optimization error for $\{f_i\}_{i\in G} \in \boldsymbol{C_{MH}(\zeta)}$.

This rate is **minimax optimal** over $\boldsymbol{C_{MH}(\zeta)}$ [Karimireddy et al., 2022].

Proposed Algorithm

Screening (inspired by [Alistarh et al., 2018]):

The number of input $\{x_i\}_{i=1}^n$ within a hyper-sphere of radius $\Theta(\zeta)$ centered around x_i is less than half of the total number of clients
⇒ client i is identified as Byzantine and x_i is removed.



Momentum (used in [Karimireddy et al., 2022]):

To reduce the stochastic noise, momentum is introduced:

$$m_i^t = (1-\alpha)m_i^{t-1} + \alpha g_i^t,$$

where g_i^t is a minibatch stochastic gradient of client i.

⇒ **Screening is applied to momentum** $\{\boldsymbol{m_i^t}\}_{i=1}^n$ **for each round.**

Concrete Algorithm:

```
Momentum Screening $(x^0, \eta, \alpha, \tau)$:
For round $t = 1$ to $T$ do:
  For client $i \in \{1, \dots, n\}$ in parallel do:
    If $i \in G$ then:
      Compute minibatch stochastic gradient $g_i^t$ at $x^{t-1}$.
      Send $\boldsymbol{m_i^t = (1-\alpha)m_i^{t-1} + \alpha g_i^t}$ $(m_i^0 = g_i^1)$ to the server.
    Else:
      Send arbitrary vector to the server.  # Client $i$ is Byzantine
  $\boldsymbol{\widehat{G} \coloneqq \{i \in [n]\colon |\{j \in [n]\colon \|m_i - m_j\| \le \tau\}| \ge 0.5n\}}$. # Screened clients
  $x^t = x^{t-1} - \eta(1/|\widehat{G}|)\sum_{i\in\widehat{G}} m_i$.
```

Theoretical Results

Theorem (Convergence Rate):

Let $\eta \le \frac{1}{8\sqrt{6}}$, $\alpha \coloneqq 4\sqrt{6}\eta L$ (≤ 0.5). For any $\{f_i\}_{i\in G} \in \boldsymbol{C_{UH}(\zeta)}$,

Momentum Screening with appropriate $\tau = \Theta(\zeta)$ satisfies

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x^{t-1})\|^2 \le O\left(\frac{\Delta_{0,*}}{\eta T}\right) + O(\delta^2\zeta^2) + \tilde{O}\left(\left(\frac{1}{\eta L T} + \eta L\right)\left(\delta^2 + \frac{1}{|G|}\right)\sigma^2\right)$$

with high probability, where $\Delta_{0,*} \coloneqq f(x^0) - f(x^*)$.

In particular, $\eta \coloneqq \frac{1}{8\sqrt{6}L} \wedge \left(\frac{1}{\sqrt{TL}}\right)$ yields

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(x^{t-1})\|^2 \le \boldsymbol{O(\delta^2\zeta^2)}$$

for sufficiently large T.

⇒ The rate is better than the previous optimal rate $\boldsymbol{O(\delta\zeta^2)}$ for $\boldsymbol{C_{MH}(\zeta)}$.

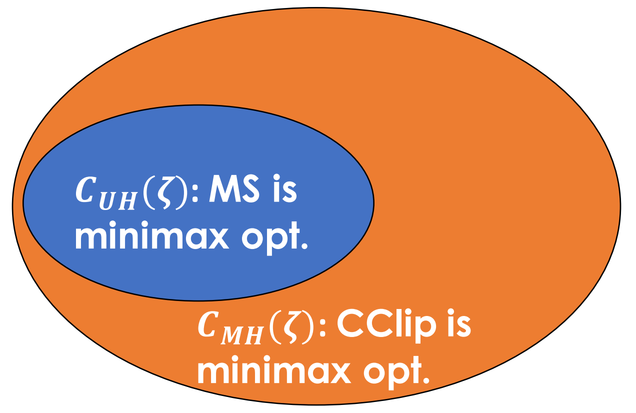
Theorem (Lower Bound for $\boldsymbol{C_{UH}(\zeta)}$):

For any opt. alg. $\boldsymbol{A}$, there exists $\{f_i\}_{i=1}^{(1-\delta)n} \in \boldsymbol{C_{UH}(\zeta)}$ and $\{f_i\}_{i=(1-\delta)n+1}^n$ s.t.

$$\mathbb{E}_\pi\left\|\nabla f\left(A\left(\{f_{\pi(i)}\}_{i=1}^n\right)\right)\right\|^2 \ge \boldsymbol{\Omega(\delta^2\zeta^2)},$$

where π is a random permutation on $[n]$.

This implies **minimax optimality** of MS on $\boldsymbol{C_{UH}(\zeta)}$!



$C_{UH}(\zeta)$: MS is minimax opt.

$C_{MH}(\zeta)$: CClip is minimax opt.

Empirical Validation of A5

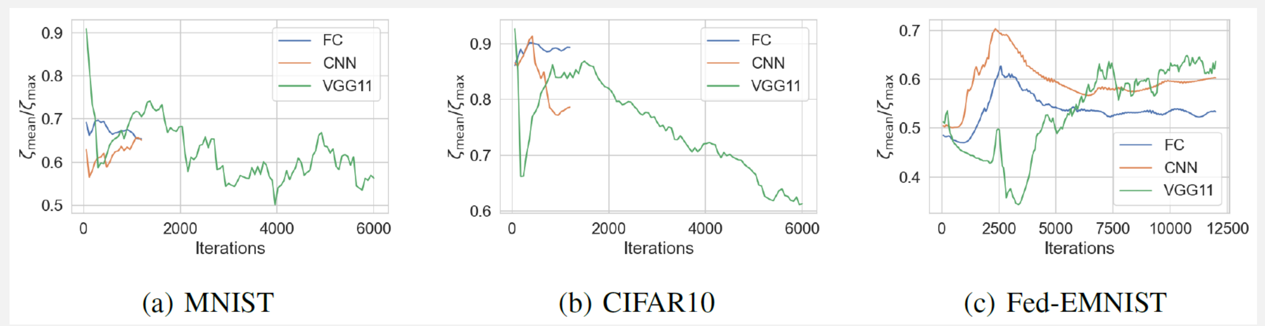
Empirical Comparison of A5 and A5':

Given $\{f_i\}_{i\in G}$, ζ_{max} and ζ_{mean} denote ζ defined in **A5** and **A5'** resp.

Q. Is ζ_{max} much larger than ζ_{mean} practically?

A. No! $\zeta_{mean}/\zeta_{max} \approx 0.3{\sim}0.9$ in our experiments.

⇒ $\boldsymbol{C_{UH}(\zeta)}$ is not so small compared to $\boldsymbol{C_{MH}(\zeta)}$ empirically.



(a) MNIST (b) CIFAR10 (c) Fed-EMNIST

Empirical values of ζ_{mean}/ζ_{max} along the trajectories of momentum SGD ($\alpha = 0.1$) without Byzantine clients for FC, CNN, and VGG11 on MNIST, CIFAR10, and Fed-EMNIST.

Numerical Results

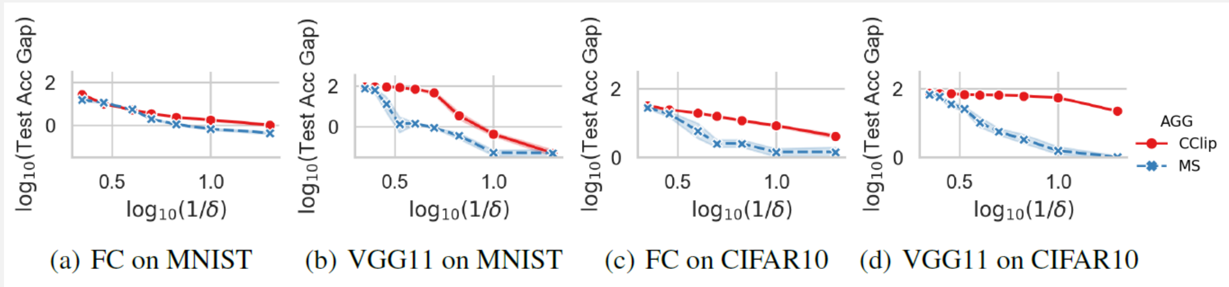
Experiment1: Investigating robustness to various attacks

- **Models**: Fully Connected MLP (FC), VGG11
- **Datasets**: MNIST, CIFAR10 with non IID allocation
- **Attacks**: Bit Flipping (BF), Label Flipping (LF), Mimic, IPM, ALIE
- **Methods**: Average (Avg), $\underbrace{\text{CM, KRUM, RFA}}_{\text{Bucketing was applied.}}$, CClip, MS
- **Byzantine frac.**: $\delta = 3/20$

Model/Data	AGG	BF	LF	Mimic	IPM	ALIE	Worst
FC/ MNIST	Avg	95.1 ± 0.2	**95.5 ± 0.3**	**95.5 ± 0.3**	**94.8 ± 0.1**	89.3 ± 0.7	89.3 ± 0.7
	CM	93.1 ± 0.6	93.3 ± 0.2	94.1 ± 0.6	91.4 ± 0.6	88.2 ± 3.2	88.2 ± 3.2
	KRUM	93.0 ± 0.3	94.0 ± 0.4	94.5 ± 1.0	92.8 ± 0.4	**95.1 ± 0.1**	92.8 ± 0.3
	RFA	94.7 ± 0.2	95.3 ± 0.3	95.3 ± 0.4	93.7 ± 0.2	90.2 ± 0.5	90.2 ± 0.5
	CClip	94.8 ± 0.2	95.2 ± 0.3	95.4 ± 0.3	93.7 ± 0.2	93.2 ± 0.4	93.2 ± 0.4
	MS (ours)	**95.2 ± 0.2**	95.4 ± 0.3	**95.5 ± 0.3**	94.5 ± 0.1	94.9 ± 0.2	**94.5 ± 0.1**
VGG11/ MNIST	Avg	**99.3 ± 0.1**	**99.3 ± 0.1**	**99.4 ± 0.1**	**99.3 ± 0.1**	30.8 ± 15.1	30.8 ± 15.1
	CM	99.2 ± 0.1	99.1 ± 0.1	99.3 ± 0.1	99.1 ± 0.0	67.0 ± 10.5	67.0 ± 10.5
	KRUM	98.9 ± 0.1	99.2 ± 0.1	99.0 ± 0.1	98.7 ± 0.1	99.2 ± 0.1	98.7 ± 0.1
	RFA	**99.3 ± 0.1**	**99.3 ± 0.1**	99.3 ± 0.1	**99.3 ± 0.1**	72.8 ± 34.7	72.8 ± 34.7
	CClip	**99.3 ± 0.1**	**99.3 ± 0.1**	99.3 ± 0.1	**99.3 ± 0.1**	95.3 ± 2.8	95.3 ± 2.8
	MS (ours)	**99.3 ± 0.1**	**99.3 ± 0.0**	99.3 ± 0.1	99.0 ± 0.3	**99.3 ± 0.0**	**99.0 ± 0.3**
FC/ CIFAR10	Avg	**46.7 ± 1.3**	**46.9 ± 1.4**	**46.1 ± 1.2**	**46.7 ± 1.3**	25.2 ± 3.3	25.2 ± 3.3
	CM	39.6 ± 2.2	39.6 ± 0.9	40.2 ± 1.6	37.6 ± 1.3	27.4 ± 1.7	27.4 ± 1.7
	KRUM	35.6 ± 1.9	38.6 ± 1.2	38.2 ± 3.4	33.3 ± 1.4	37.7 ± 2.5	33.7 ± 2.1
	RFA	46.2 ± 0.7	46.7 ± 0.8	45.9 ± 2.0	45.8 ± 1.0	29.0 ± 3.7	29.0 ± 3.7
	CClip	44.5 ± 1.2	45.7 ± 0.6	44.0 ± 3.5	40.9 ± 1.0	35.4 ± 0.8	35.4 ± 0.8
	MS (ours)	46.3 ± 1.1	46.2 ± 1.3	45.2 ± 1.6	45.8 ± 1.9	**45.0 ± 2.5**	**44.6 ± 2.0**
VGG11/ CIFAR10	Avg	**84.3 ± 0.9**	**85.0 ± 0.4**	**85.1 ± 0.8**	**84.5 ± 0.3**	19.2 ± 1.3	19.2 ± 1.3
	CM	45.6 ± 2.5	43.7 ± 4.3	57.2 ± 9.2	34.9 ± 3.7	19.1 ± 1.9	19.1 ± 1.9
	KRUM	55.8 ± 2.5	64.2 ± 1.8	70.3 ± 2.2	40.6 ± 4.8	71.9 ± 8.3	40.6 ± 4.8
	RFA	82.7 ± 0.3	83.9 ± 0.2	84.2 ± 0.4	81.5 ± 0.6	20.3 ± 1.3	20.3 ± 1.3
	CClip	77.9 ± 0.7	81.3 ± 0.6	81.3 ± 0.6	64.2 ± 18.3	22.7 ± 2.3	22.7 ± 2.3
	MS (ours)	84.2 ± 0.4	84.6 ± 0.6	84.8 ± 0.9	83.5 ± 0.8	**83.3 ± 3.4**	**82.8 ± 2.5**

Experiment2: Investigating test acc gap for Byzantine frac. changes

- **Models**: Fully Connected MLP (FC), VGG11
- **Datasets**: MNIST, CIFAR10 with non IID allocation
- **Attacks**: Bit Flipping (BF), Label Flipping (LF), Mimic, IPM, ALIE
- **Methods**: CClip, MS
- **Byzantine frac.**: $\delta \in \{1/20, 2/20, 3/20, 4/20, 5/20, 7/20, 9/20\}$



(a) FC on MNIST (b) VGG11 on MNIST (c) FC on CIFAR10 (d) VGG11 on CIFAR10

Y-axis shows the gap between the best test acc of momentum SGD without Byzantine clients and the worst best test acc against 5 attacks and in log scale (smaller is better).

Results:

Both on Experiments1 and 2, **MS outperformed** the other methods including **CClip and Bucketing** in terms of the **worst best test acc against 5 attacks**.

⇒ **MS** is empirically **robust** compared with the existing methods!

References

[Karimireddy et al., 2022]: Byzantine-Robust Learning on Heterogeneous Datasets via Bucketing.

[Alistarh et al., 2018]: Byzantine Stochastic Gradient Descent.