Best Effort Voting Power Control for Byzantine-resilient Federated Learning Over the Air

Xin Fan[1] , Prof. Yue Wang[2] , Prof. Yan Huo[1] , and Prof. Zhi Tian[2]
[1]Beijing Jiaotong University, China
[2]George Mason University, USA
E-mail: {yhuo, fanxin}@bjtu.edu.cn, {ywang56,ztian1}@gmu.edu

IEEE ICC®
IEEE International Conference on Communications
16–20 May 2022 // Seoul, South Korea // Hybrid: In-Person and Virtual Conference
Intelligent Connectivity for Smart World

Outline

- Introduction
- Algorithm
- Performance Analysis
- Simulation Results
- Conclusion

Outline

- **Introduction**
- Algorithm
- Performance Analysis
- Simulation Results
- Conclusion

Introduction

--Background

Conventional centralized machine learning

- **Characteristics:**
  - Collect data for training
- **Cons:**
  - **Big data**
  - **Privacy**

Federated learning (FL)

- **Characteristics:**
  - Keep data locally
  - Send weights or gradients to a center
- **Cons:**
  - **Massive devices**

FL over the air

- **Characteristics:**
  - Uncoding
  - Linear analog modulation
  - Analog aggregation
- **Pros:**
  - **Transmit updates using the same time-frequency resources.**

Introduction

--Motivation

Challenges

- The individual local updates are unavailable
- Existing screening methods (such as geometric median, coordinate-wise median/trimmed mean) cannot work

Contributions

- Power control policy
  - Best effort voting (BEV)
- Convergence analysis
  - Strongest attack
  - Existing power control policy
  - Our BEV

Outline

- Introduction
- **Algorithm**
- Performance Analysis
- Simulation Results
- Conclusion

Algorithm

Federated learning (FL)

- Local devices (workers)
  - Receive $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ from a parameter server (PS)
  - Train to get the updates (local gradients, $\mathbf{g}_i$)
  - Send $\mathbf{g}_i$ to the PS
- PS
  - Receive $\mathbf{g}_i$ and average them

$$\mathbf{g} = \frac{\sum_{i=1}^{U} \mathbf{g}_{i,t}}{U}$$

  - Update the sharing model

$$\mathbf{w} = \mathbf{w} - \alpha \mathbf{g}$$

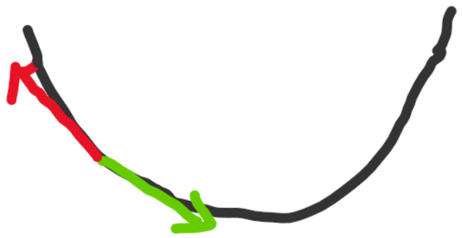
  - Broadcast $\mathbf{w}$ to local workers

Algorithm --Channel Inversion Power Control

❑ FL over the air

- *N* out of *U* workers are Byzantine attackers and *M* =*U*-*N* normal workers

$$\mathbf{y}_t = \sum_{m=1}^{M} p_{m,t}|h_{m,t}|\tilde{\mathbf{g}}_{m,t} + \sum_{n=1}^{N} \hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t} + \mathbf{z}_t$$



❑ The existing channel inversion (CI) power control

- The power allocation factor

$$p_{i,t} = \frac{b_0}{|h_{i,t}|}, \quad \forall i \qquad p_{i,t}^2 \le p_i^{\max}, \quad \forall i \qquad b_0 = \min\{|h_{i,t}|\sqrt{p_i^{\max}}\}_i^U$$

- When N=0, then

$$\mathbf{y}_t = \sum_{m=1}^{U} b_0\tilde{\mathbf{g}}_{m,t} + \mathbf{z}_t \longrightarrow \hat{\mathbf{g}} = \frac{\mathbf{y}_t}{Ub_0} = \frac{\sum_{m=1}^{U}\tilde{\mathbf{g}}_{m,t}}{U} + \frac{\mathbf{z}_t}{Ub_0}$$

Voting: [1 1 1 1 -5]→-1

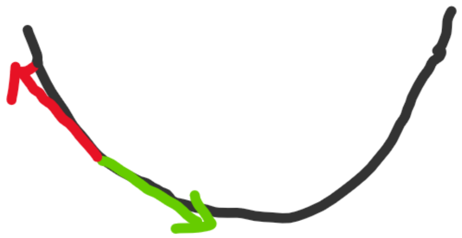
$$\mathbf{g} = \frac{\sum_{i=1}^{U}\mathbf{g}_{i,t}}{U}$$

Algorithm

--BEV-SGD

- FL over the air
  - N out of U workers are Byzantine attackers and M =U-N normal workers

$$\mathbf{y}_t = \sum_{m=1}^{M} p_{m,t}|h_{m,t}|\tilde{\mathbf{g}}_{m,t} + \sum_{n=1}^{N} \hat{p}_{n,t}|h_{n,t}|\hat{\mathbf{g}}_{n,t} + \mathbf{z}_t$$



- Best effort voting SGD (BEV- SGD)
  - Transmission with the maximum power.
  - Byzantine attackers can send anything under the power constraints.

Voting: [1 1 1 1 -5]→-1 → *Voting: [3 2 1 4 -5]→5*

- If without our BEV-SGD, how attacks affect FL?
- Using our BEV-SGD, what the level of attack can FL resist?

Outline

- Introduction
- Algorithm
- **Performance Analysis**
- Simulation Results
- Conclusion

Performance Analysis --Strongest attacks

❑ Basic Assumptions

- Assumption 1 (Lipschitz continuity, smoothness):

$$\|\nabla F(\mathbf{w}_{t+1}) - \nabla F(\mathbf{w}_t)\| \leq L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|$$

- Assumption 2 (bounded gradient estimates):

$$\mathbb{E}(\mathbf{g}_{i,t}) = \mathbf{g}_t, \quad \mathbb{E}(\|\mathbf{g}_{i,t} - \mathbf{g}_t\|^2) \leq \delta^2, \quad \forall i, t,$$

❑ The Strongest Byzantine Attacks (in Theorem 1)

- Get the true gradient using training data
- Send the opposite gradient with the maximum power

$$\hat{\mathbf{g}}_{n,t} = -\mathbf{g}_{n,t}$$

The Convergence of SGD with CI Transmission (in Theorem 2)

$$\mathbb{E}[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2)] \leq \frac{1}{\sqrt{T}}\left(\frac{2L\Omega_{CI}}{\omega_{CI}^2\bar{\alpha}}(F(\mathbf{w}_0)-F(\mathbf{w}^*)) + \bar{\alpha}\left(\delta^2+\frac{1}{\Omega_{CI}}\epsilon^2 z^2\right)\right), \quad (20)$$

where

$$\omega_{CI} = Mb_0 - \sum_{n=1}^{N}\sqrt{\frac{\pi\sigma_n^2 p_n^{\max}}{2D}}, \quad (21)$$

$$\Omega_{CI} = (U+N)\left(Ub_0^2 + \sum_{n=1}^{N}\frac{2\sigma_n^2 p_n^{\max}}{D}\right), \quad (22)$$

Convergence condition

$$\boxed{\omega_{CI} > 0} \rightarrow \boxed{\frac{U}{1+\sqrt{\pi U}}}$$

- The Convergence of SGD with BEV Transmission **(in Theorem 3)**

$$\mathbb{E}[\sum_{t=1}^{T}\frac{1}{T}\|\mathbf{g}_t\|^2)] \leq \frac{1}{\sqrt{T}}\left(\frac{2L\Omega_{BEV}}{\bar{\alpha}\omega_{BEV}^2}(F(\mathbf{w}_0)-F(\mathbf{w}^*)) + \bar{\alpha}\left(\delta^2+\frac{1}{\Omega_{BEV}}\epsilon^2 z^2\right)\right), \quad (24)$$

where

$$\omega_{BEV} = \sum_{i=1}^{M}\sqrt{\frac{p_i^{\max}\pi}{2D}}\sigma_i - \sum_{n=1}^{N}\sqrt{\frac{p_n^{\max}\pi}{2D}}\sigma_n, \quad (25)$$

$$\Omega_{BEV} = (U+N)\sum_{i=1}^{U}\frac{2\sigma_i^2 p_i^{\max}}{D}, \quad (26)$$

Convergence condition

$\boxed{\omega_{BEV} > 0} \rightarrow \boxed{N \leq \frac{U}{2}}$

$\boxed{\frac{U}{2} \geq \frac{U}{1+\sqrt{\pi U}}}$

Performance Analysis --Comparison

- For large learning rate

$$O(\frac{1}{\Omega\sqrt{T}}) \qquad \Omega_{BEV} > \Omega_{CI}$$

BEV is better than CI

- For small learning rate

$$O(\frac{\Omega}{\omega^2\sqrt{T}})$$

Depends on the specific parameters

- No attackers for small learning rate
  - CI has $\omega_{CI}^2 = \Omega_{CI}$

$$O(\frac{1}{\sqrt{T}})$$

Error-free case

  - BEV has $\omega_{BEV}^2 \leq \Omega_{BEV}$

$$O(\frac{\Omega_{BEV}}{\omega_{BEV}^2\sqrt{T}})$$

CI is better than BEV

Outline

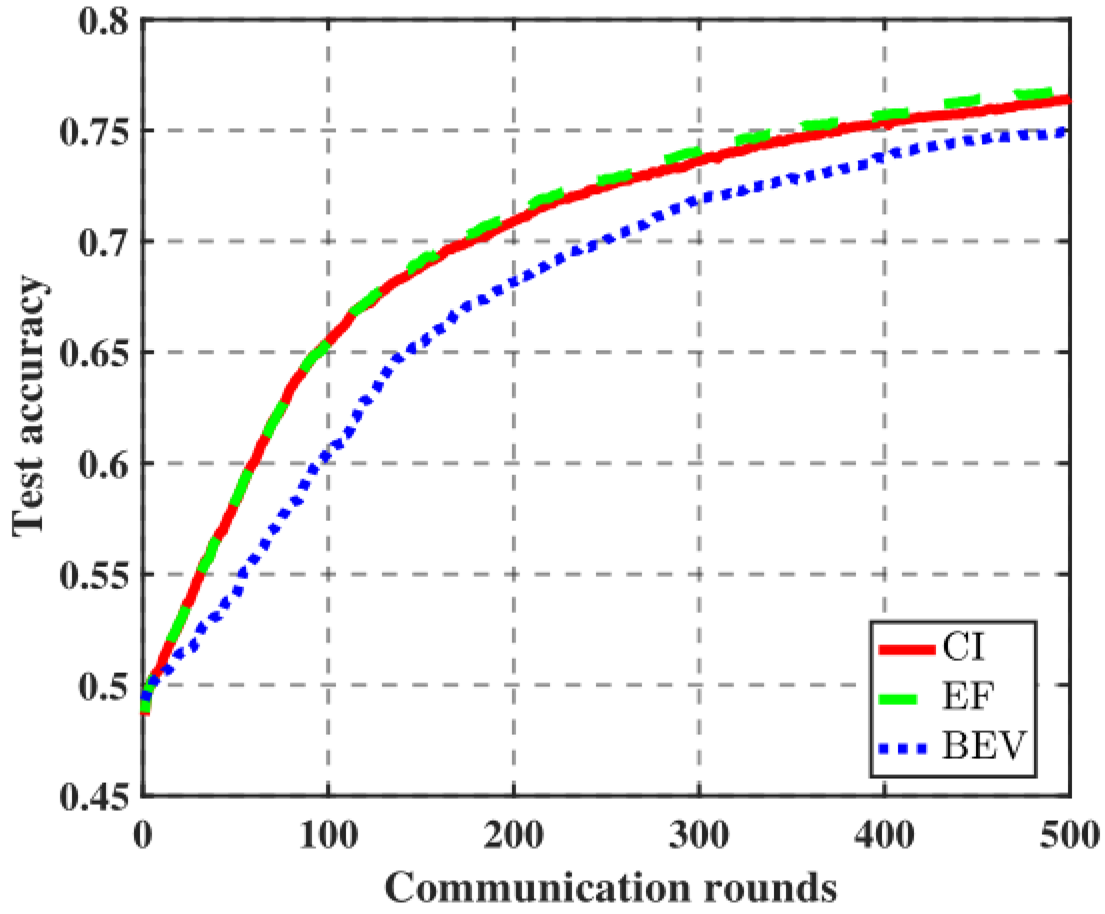
- Introduction
- Algorithm
- Performance Analysis
- **Simulation Results**
- Conclusion

Simulation Results

- Wireless network setting
  - 10 workers
  - Rayleigh fading model.
- Task
  - Image classification with the MNIST dataset
- Scenarios
  - Case 1: Without any attacks
  - Case 2: Only one attacker who is far from the server
  - Case 3: Only one attacker who is close to the server
  - Case 4: Randomly selected several attackers

Simulation Results --Case 1

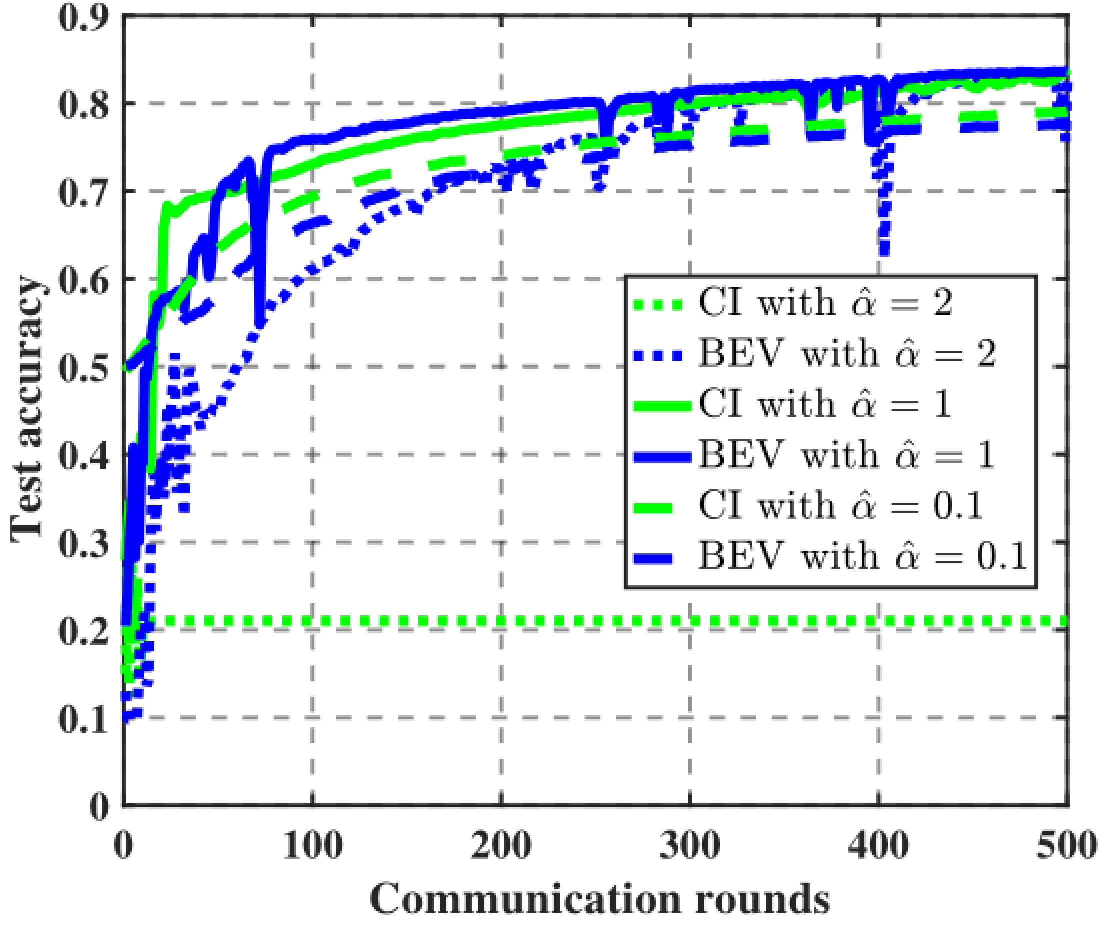
❑ Performance without Attacks



CI is almost the same as the error-free case, which is better than BEV

Simulation Results --Case 2

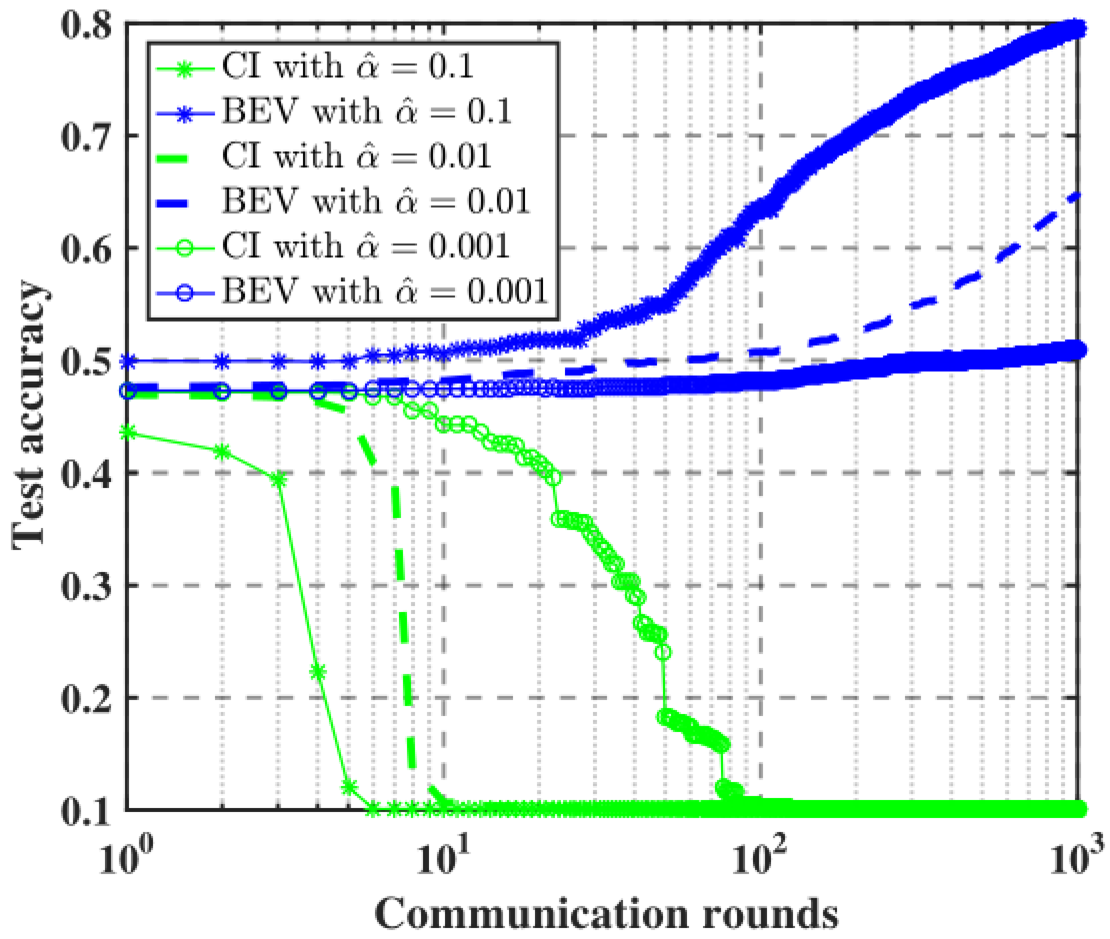
❑ Performance under a Single Attacker with Weak Channel Gain



Under large learning rate, our BEV is better, but for small learning rate, CI is better

Simulation Results --Case 3

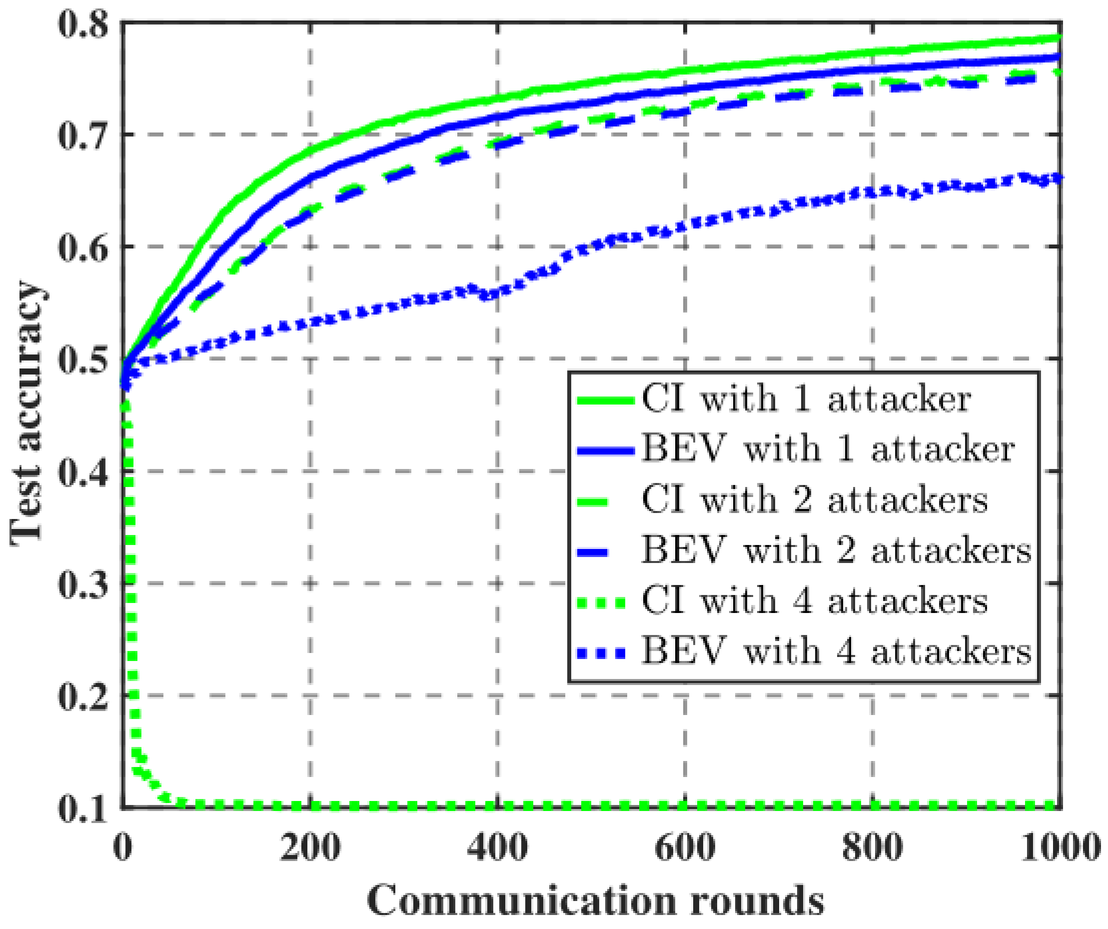
- Performance under a Single Attacker with Large Channel Gain



Our BEV is better than CI, regardless learning rate

Simulation Results --Case 4

❑ Performance with Multiple Randomly Selected Attackers



Our BEV can defend more attackers

Outline

- Introduction
- Algorithm
- Performance Analysis
- Simulation Results
- **Conclusion**

Conclusion

- Without attacks, CI is better than BEV
- Under weak attacks for small learning rate, CI is better than BEV
- Under weak attacks for large learning rate, BEV is better than CI
- Under strong attacks, BEV is better than CI.
- **In practice, BEV is a better option for potential attacks**

THANK YOU

Questions?

Xin FAN

Email: fanxin@bjtu.edu.cn