TPDS 2022

A Decentralized Federated Learning Framework via Committee Mechanism with Convergence Guarantee

Chunjiang Che[1], Xiaoli Li[1], Chuan Chen[1]
Xiaoyu He[1], and Zibin Zheng[2]

Speaker: Chunjiang Che

1. School of Computer Science, Sun Yat-sen University
2. School of Software Engineering, Sun Yat-sen University

中山大學
SUN YAT-SEN UNIVERSITY

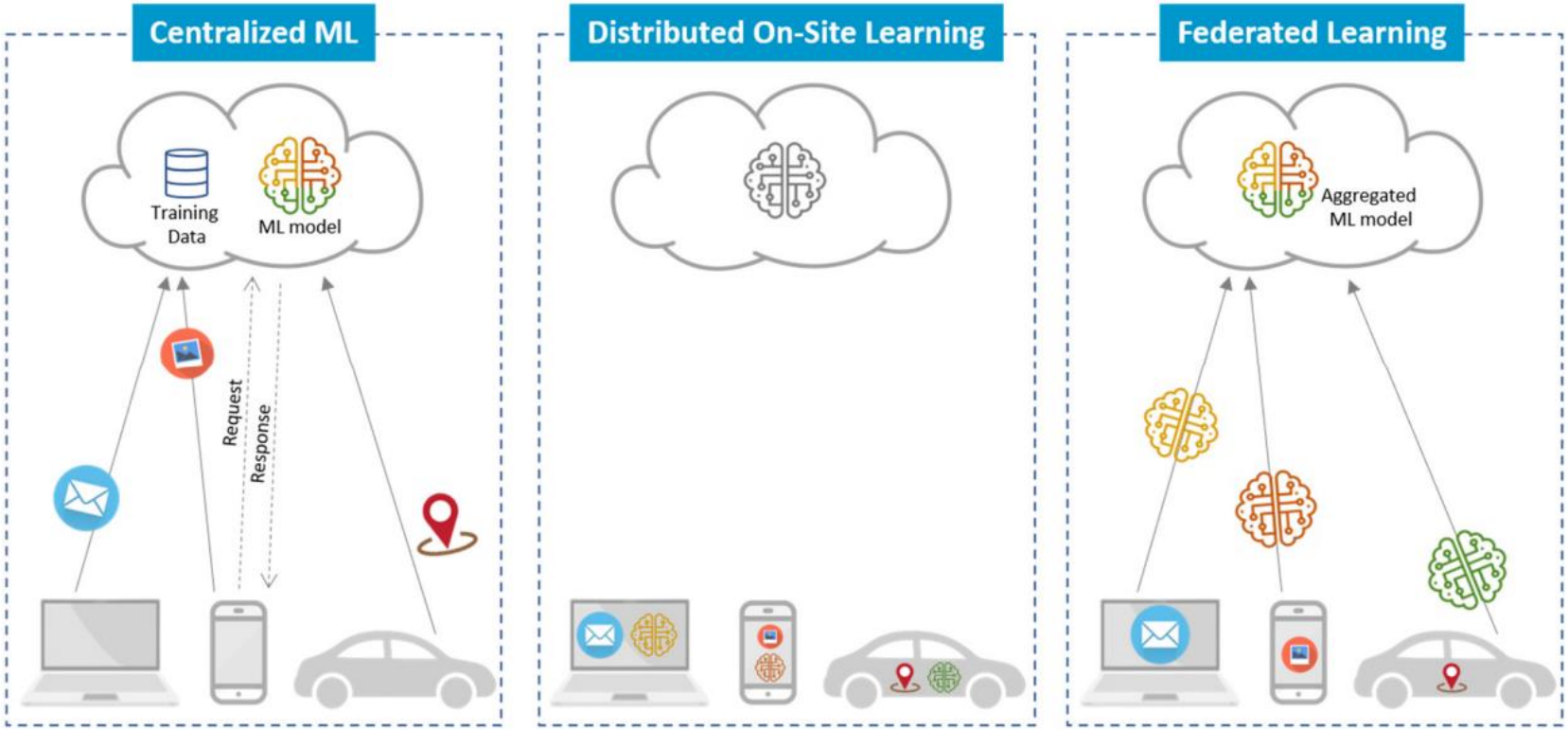
Outline



- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

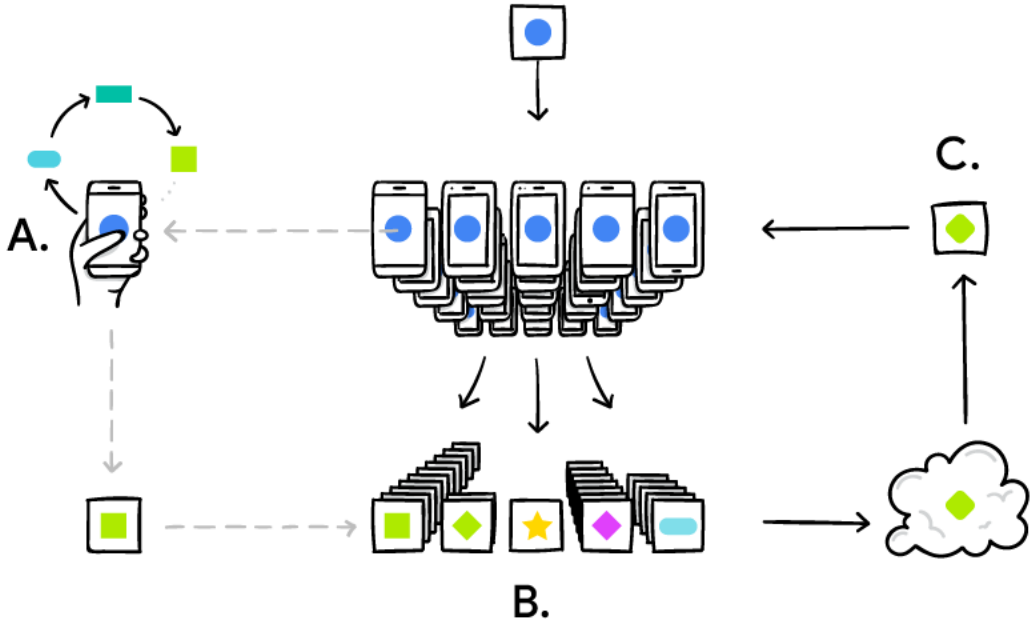
Distributed Machine Learning

- Centralized ML method violates user privacy
  - Federated Learning protects user privacy



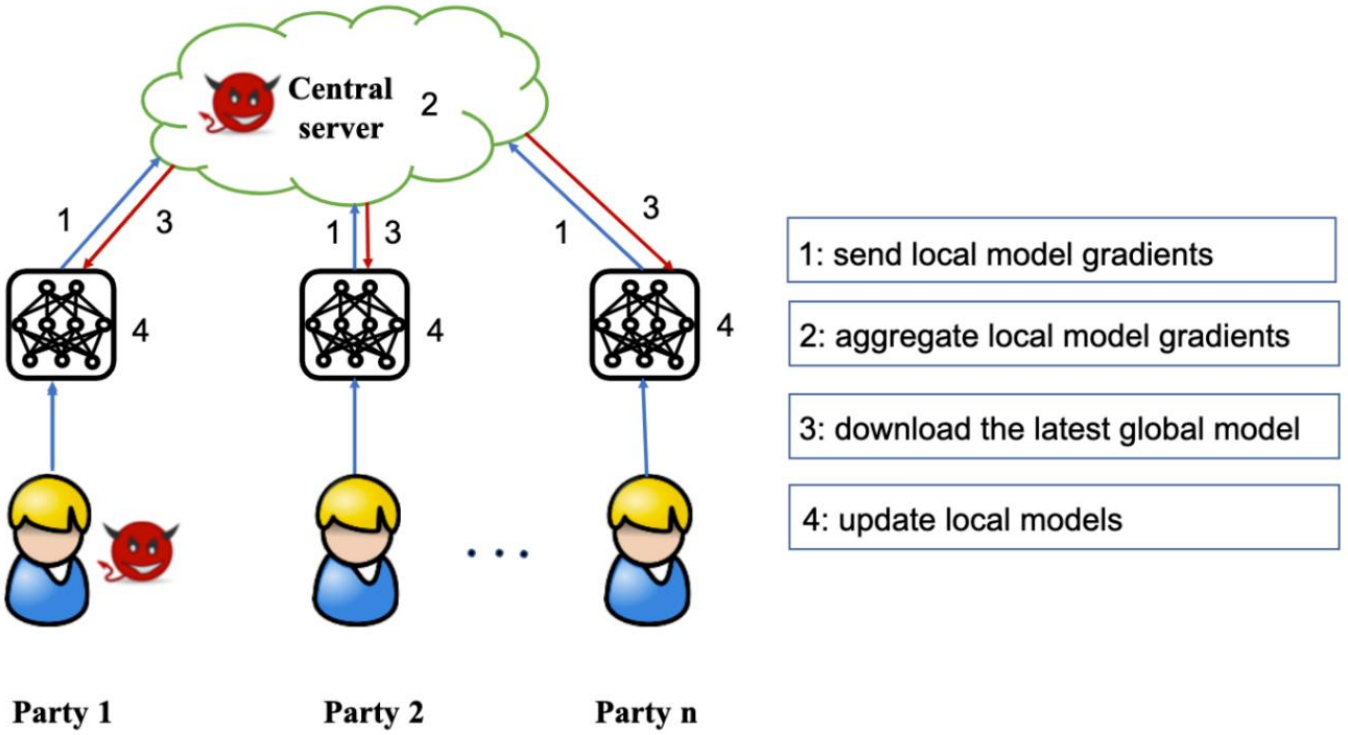
Federated Learning

- Algorithm
  - Local training, gradient uploading, aggregation, model downloading



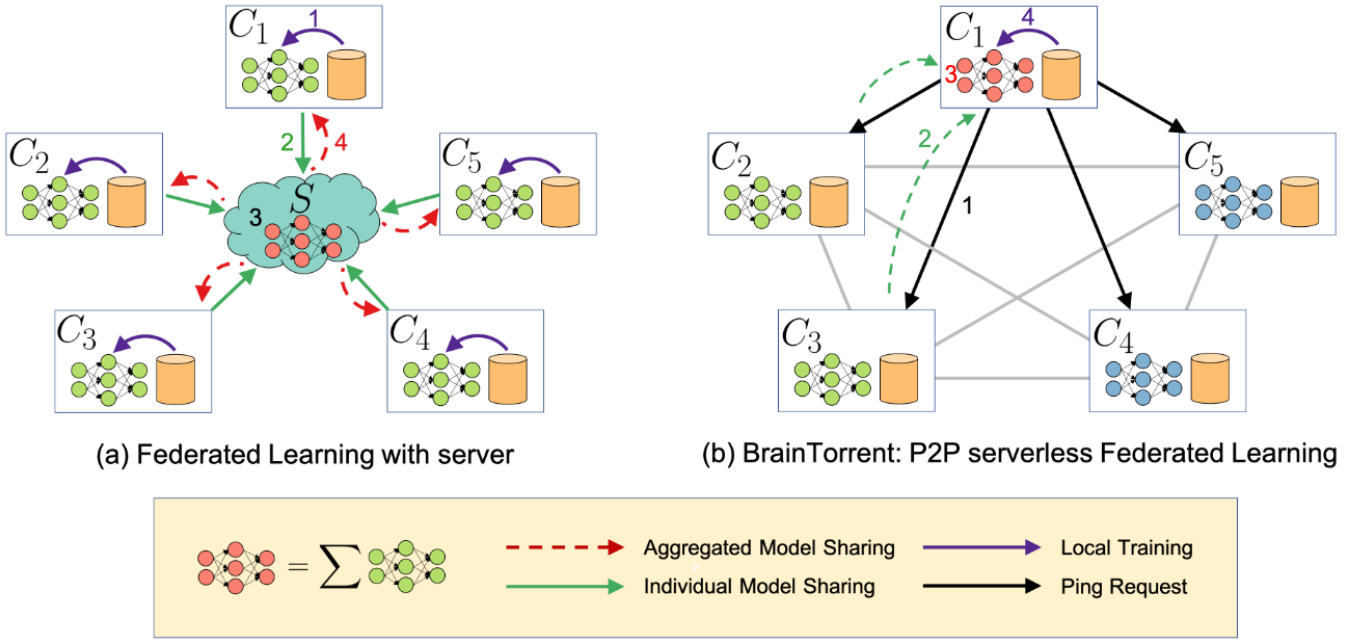
Malicious Participant

- Byzantine attacks reduce model performance
  - Current Byzantine-tolerance algorithms: Median, Trimmed Mean, Krum and Multi-Krum



Decentralized Federated Learning

- BrainTorrent
  - Nodes complete aggregation through P2P protocol

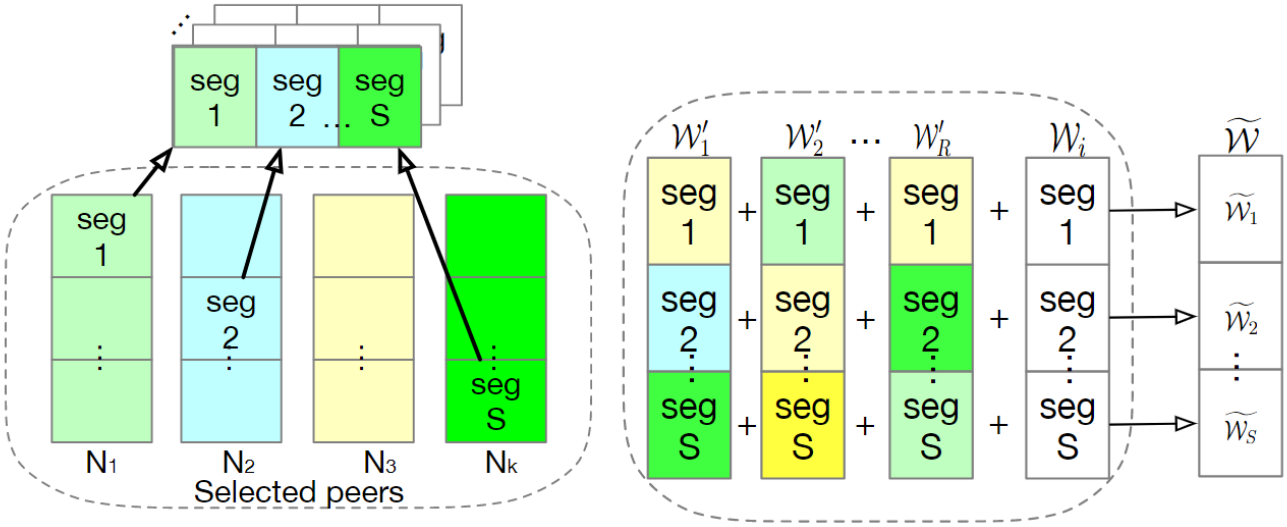


(a) Federated Learning with server

(b) BrainTorrent: P2P serverless Federated Learning

Decentralized Federated Learning

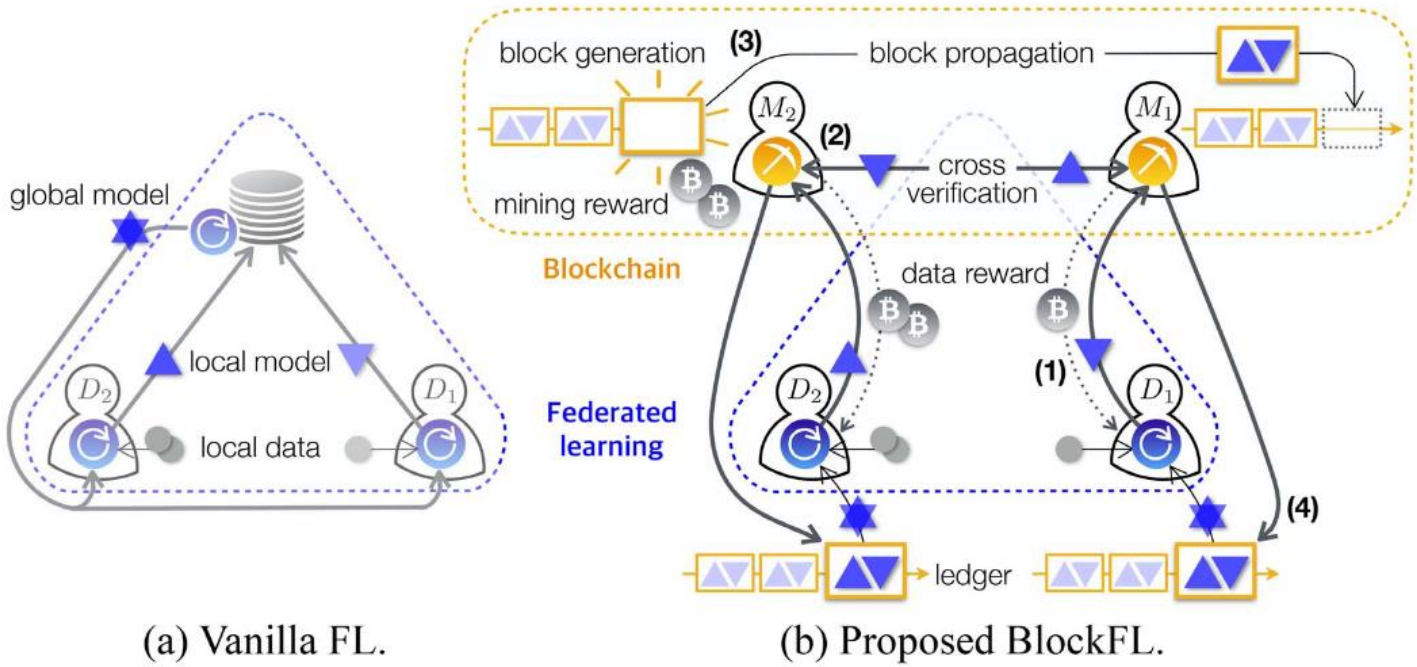
- Gossip
  - Nodes complete aggregation through Gossip protocol



(a) Segmented Pulling (b) Segmented Aggregation

Decentralized Federated Learning

- Blockchain
  - Nodes complete aggregation through Blockchain consensus



(a) Vanilla FL. (b) Proposed BlockFL.

Limitation

- Current Byzantine-tolerance algorithms
  - Cannot achieve satisfactory performance
  - Cannot provide effective defense in the presence of malicious servers
- Current decentralized FL frameworks
  - Not resistant to Byzantine attacks
  - High communication overhead

CMFL: A Robust Decentralized Federated Learning Framework

Outline



- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

Problem Formulation

- Local objective function

$$F_k(\mathbf{w}) = \frac{1}{n_k} \sum_{\xi \in \mathcal{D}_k} f(\mathbf{w}, \xi).$$

- Global objective function

$$F(\mathbf{w}) = \sum_{k=1}^{K} p_k F_k(\mathbf{w})$$

- Local SGD training

$$g_k(\mathbf{w}_{k,i}^t, \mathcal{B}_{k,i}^t) = \frac{1}{|\mathcal{B}_{k,i}^t|} \sum_{\xi \in \mathcal{B}_{k,i}^t} \nabla f(\mathbf{w}_{k,i}^t, \xi).$$

Problem Formulation

- Local update

$$\mathbf{w}_{k,i+1}^t = \mathbf{w}_{k,i}^t - \eta_i^t g_k(\mathbf{w}_{k,i}^t, \mathcal{B}_{k,i}^t)$$

- Aggregation

$$\overline{g}^t = \sum_{k \in S^t} p_{k,S^t} g_k(\mathbf{w}_{k,\tau}^t, \mathcal{B}_{k,\tau}^t)$$

- Global update

$$\overline{\mathbf{w}}^{t+1} = \overline{\mathbf{w}}^t - \eta_t \overline{g}^t$$

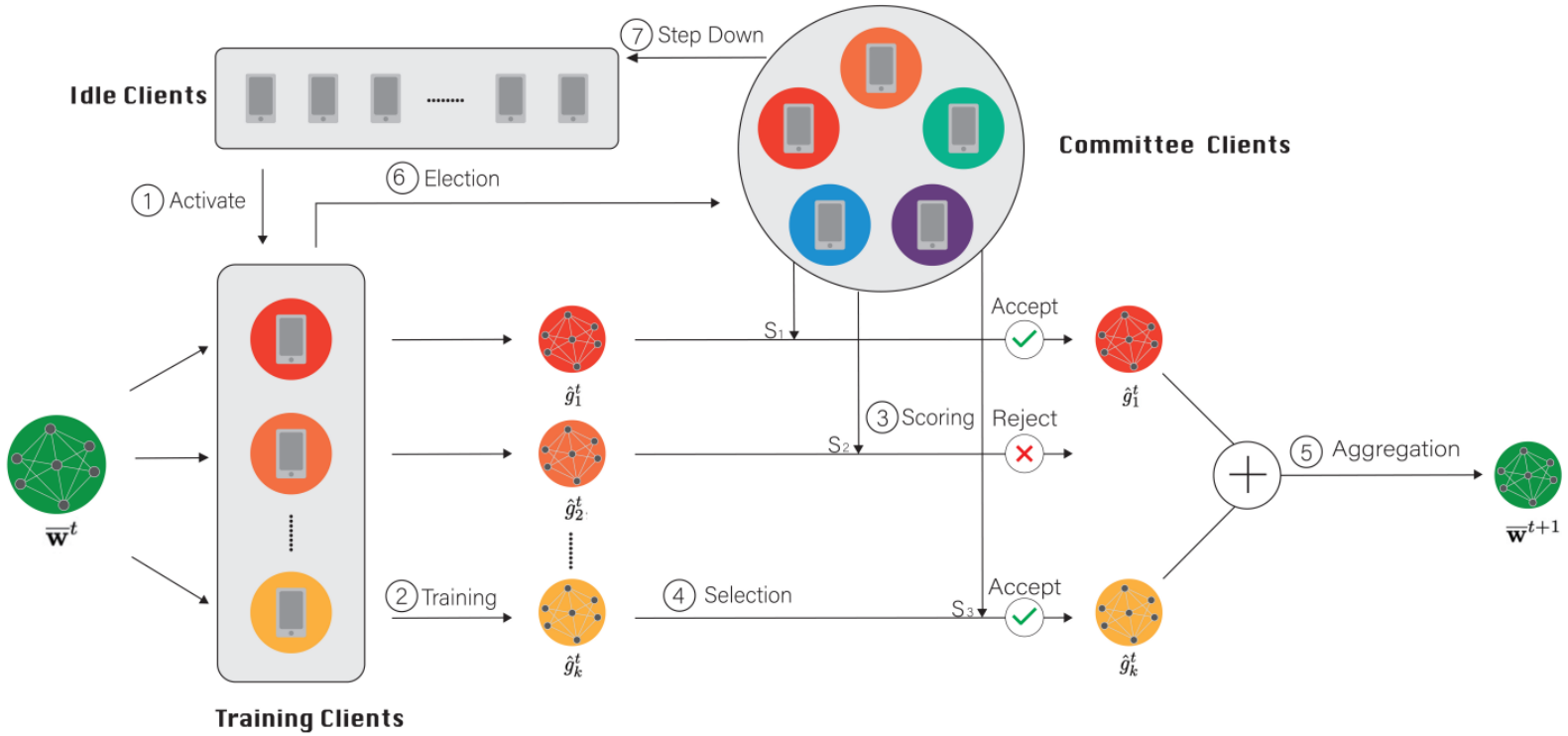
Outline



- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

Committee Mechanism based FL

- CMFL utilize the committee mechanism to implement a decentralized FL framework



Scoring System

- The scoring system based on Euclidean distance to distinguish malicious and honest gradients
  - Honest gradients get a high score
  - Malicious gradients get a low score

$$\mathcal{P}_k^c = \frac{1}{||\hat{g}_k^t - \hat{g}_c^t||_2^2}.$$

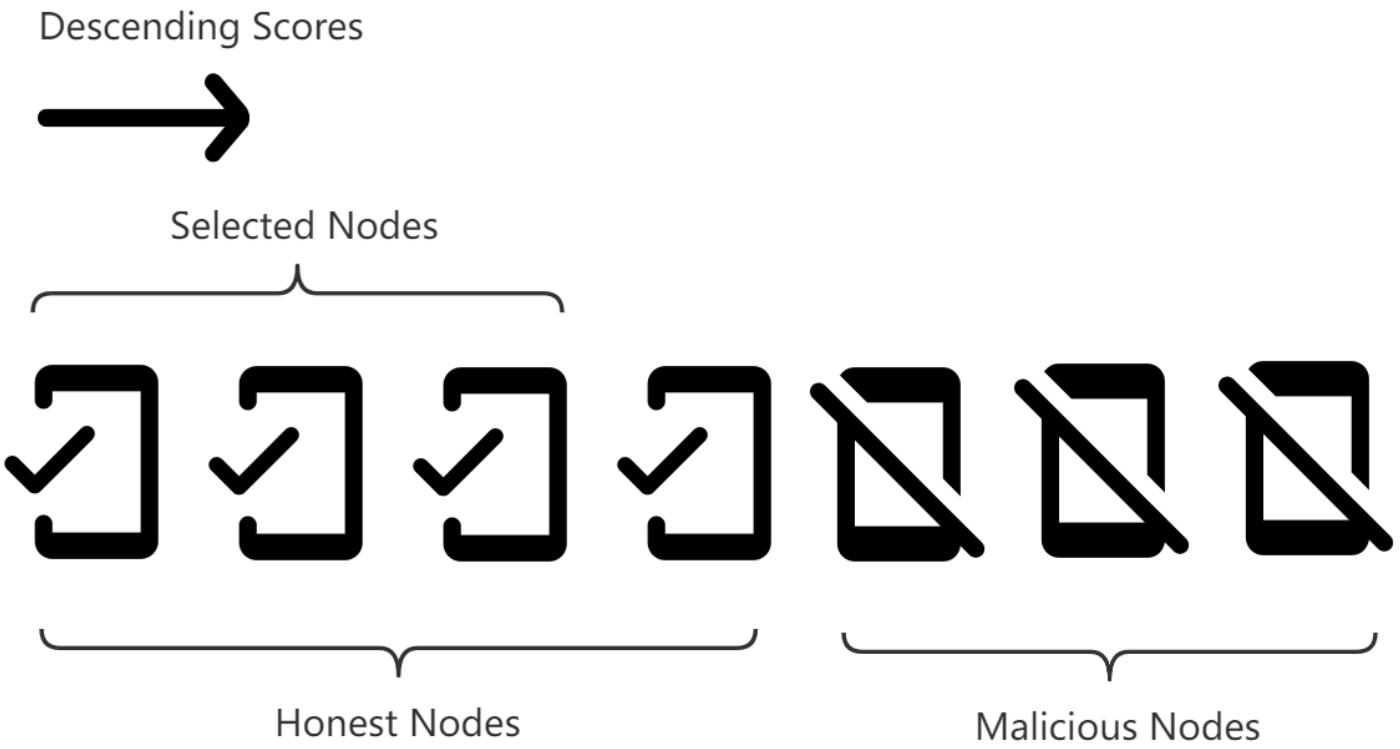
The c-th committee client's score

$$\mathcal{P}_k = \frac{1}{\frac{1}{C}\sum_c^C ||\hat{g}_k^t - \hat{g}_c^t||_2^2} = \frac{C}{\sum_c^C \frac{1}{\mathcal{P}_k^c}}$$

Total score

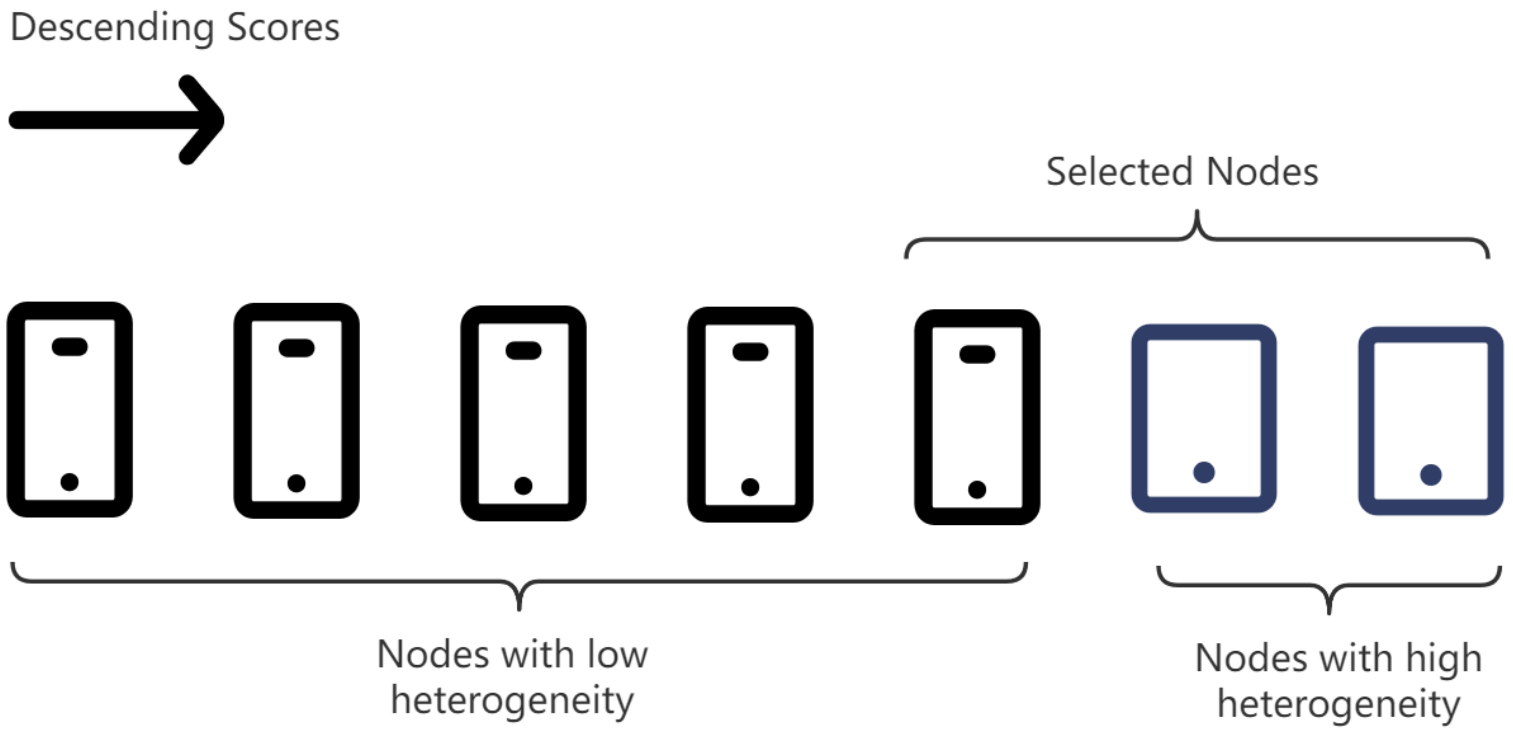
Selection Strategy

- Selection strategy Ⅰ
  - Consideration: Robustness
  - Applicable scenario: Byzantine attack scenario



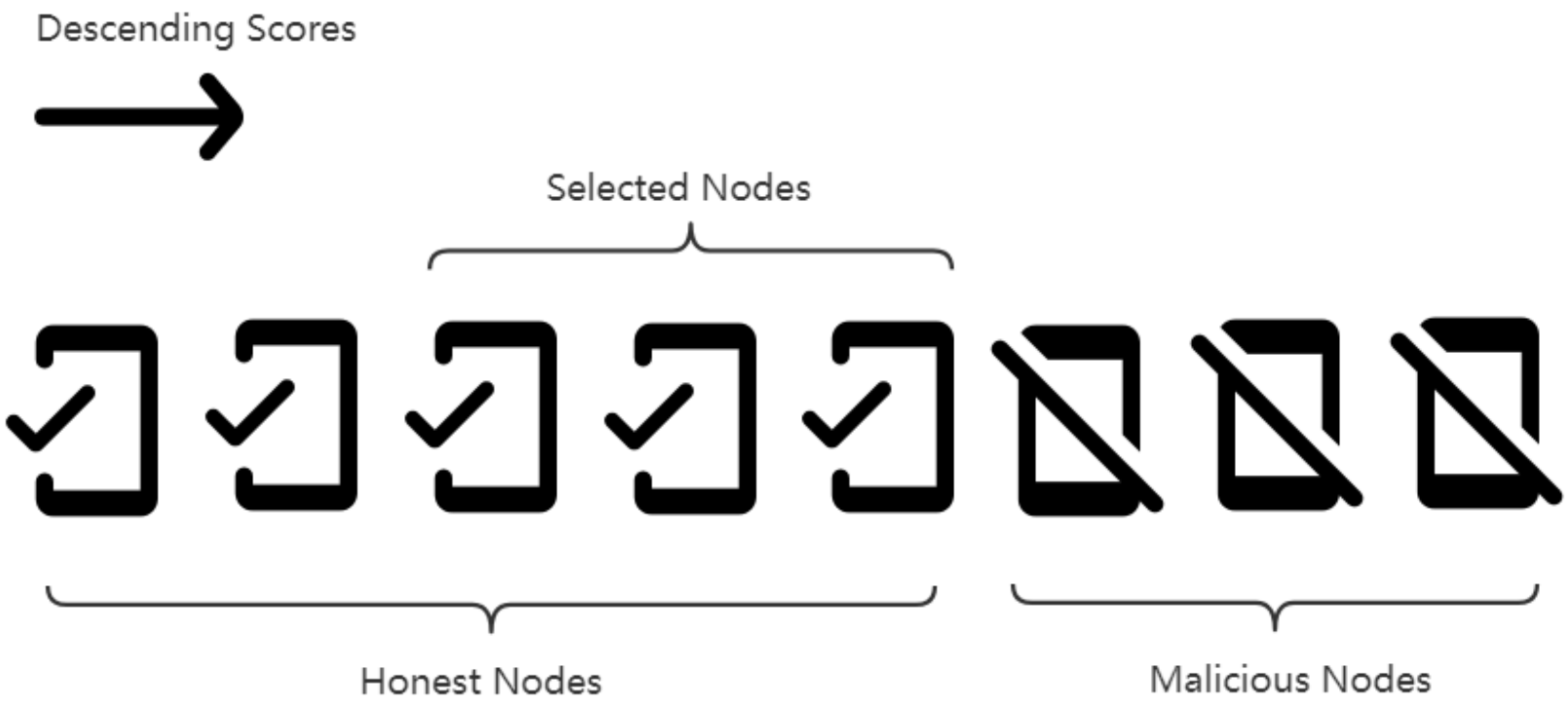
Selection Strategy

- Selection strategy Ⅱ
  - Consideration: Heterogeneity
  - Applicable scenario: Non-attack scenario



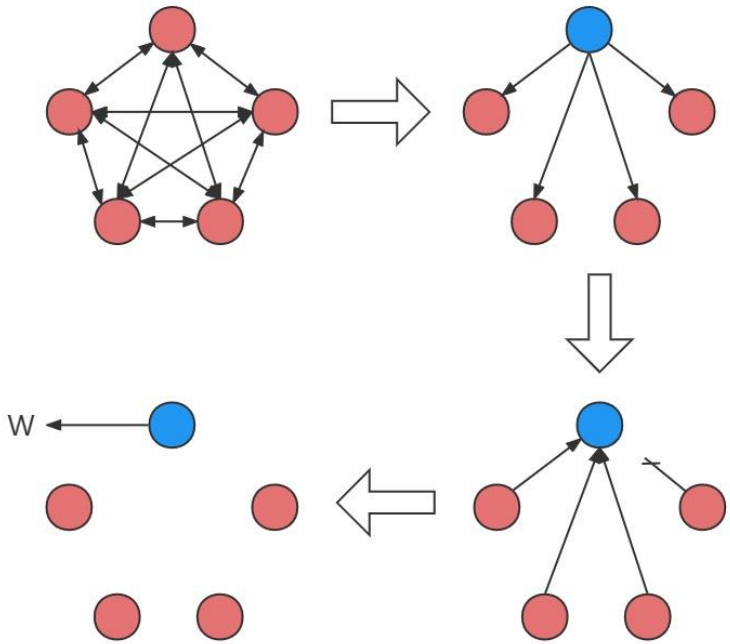
Election Strategy

- Considerations
  - Robustness
  - System Stability



Committee Consensus Protocol

- The committee members reach a consensus by voting.



Algorithm 1 Committee Consensus Protocol

Input: t, S_c^t, S_b^t
Output: global model $\mathbf{w}^t$, new committee client set S_c^{t+1}

1: **for** $c \in S_c^t$ **do**
2: **for** $k \in S_b^t$ **do**
3: The c-th committee client broadcasts score $\mathcal{P}_k^c$ to other committee clients.
4: **end for**
5: **end for**
6: **for** $c \in S_c^t$ **do**
7: **for** $k \in S_b^t$ **do**
8: The c-th committee client calculates the total score $\mathcal{P}_k$ of the k-th training client.
9: **end for**
10: **end for**
11: The committee performs voting(S_a^t, aggregtion) as Algorithm2.
12: The committee performs voting(S_c^{t+1}, None) as Algorithm2.

Algorithm 2 The voting algorithm

Input: The client set S and the opertation o

1: **while** No consensus on S **do**
2: The primary committee client p is selected from committee clients by random.
3: **for** $k \in S_c^t$ **do**
4: The k-th committee client decides its client set S^k according to the corresponding strategy (selection strategy/election strategy).
5: **end for**
6: p performs the operation o.
7: p creates a request $\langle$ *Request*, S^p, *operation*, *timestamp* $\rangle$ and broadcasts it to all the replicate committee clients.
8: **for** $k \in S_c^t \setminus \{p\}$ **do**
9: The k-th replicate committee client receives the request from p.
10: **if** $S^k == S^p$ **then**
11: The k-th committee client performs the operation o.
12: Return $\langle$ *Reply*, *timestamp*, S^p, *response* $\rangle$ to p.
13: **end if**
14: **if** p receives at least $\lfloor C/2 \rfloor + 1$ response **then**
15: Consensus is reached on S.
16: **end if**
17: **end for**
18: **end while**

Training Algorithm

- Random Sampling
- Local Training
- Gradient Verification
- Global Model Updating
- Election of New Committee Members

Algorithm 3 The training algorithm of CMFL

Input: $\tau, T, K, m, C, \eta_t, \eta_i^t$
Output: target global model $\mathbf{w}^T$

Initialize $\mathbf{w}^1$, S_c^1 and S_b^1 randomly.
for $t = 1$ to T **do**
  for $k \in S_b^t \cup S_c^t$ **do**
    for $i = 1$ to τ **do**
      The k-th client runs the SGD over the local dataset by $\mathbf{w}_{k,i}^t \Leftarrow \mathbf{w}_{k,i-1}^t - \eta_i^t g_k(\mathbf{w}_{k,i-1}^t, \mathcal{B}_{k,i-1}^t)$.
    end for
  end for
  for $k \in S_b^t$ **do**
    for $c \in S_c^t$ **do**
      The k-th training client send its local gradient $\hat{g}_k^t$ to the c-th committee client.
      The c-th committee client assign a score $\mathcal{P}_k^c$ on the k-th training client/local gradient according to the scoring system.
    end for
  end for
  The committee clients execute CCP to complete the selection, aggregation, and election process.
  The S_b^t be reinitialized to form the S_b^{t+1}.
  for $k \in S_b^{t+1} \cup S_c^{t+1}$ **do**
    The k-th client downloads the global model from the primary committee client.
  end for
end for

Outline



- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

Assumption

- Assumption 1 and 2 are standard conditions in FL setting

Assumption 1. (Lipschitz Gradient). *$F_1, ..., F_K$ are all $\mathcal{L}$-smooh: for all $\mathbf{v}, \mathbf{w} \in R^n$, $k = 1, ..., K$, $F_k(\mathbf{v}) \leq F_k(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{L}{2}||\mathbf{v} - \mathbf{w}||_2^2$.*

Assumption 2. (μ-strongly Convex Gradient). *$F_1, ..., F_K$ are all μ-strongly convex: for all $\mathbf{v}, \mathbf{w} \in R^n$, $k = 1, ..., K$, $F_k(\mathbf{v}) \geq F_k(\mathbf{w}) + (\mathbf{v} - \mathbf{w})^T \nabla F_k(\mathbf{w}) + \frac{\mu}{2}||\mathbf{v} - \mathbf{w}||_2^2$*

Assumption

- Assumption 3 is a form of bounded variance between the local objective functions over full dataset and the batch dataset

Assumption 3. (Bounded Variance). *For the mini-batch* $B_{k,i}^t$ *uniformly sampled randomly from k-th client's dataset* $\mathcal{D}_k$*, the resulting stochastic gradient is unbiased:* $E[g_k(\mathbf{w}_{k,i}^t, \mathcal{B}_{k,i}^t)] = \nabla F_k(\mathbf{w}_{k,i}^t)$ *for all* $k = 1, ..., K, t = 1, ..., T, i = 1, ..., \tau$. *And the variance of stochastic grandient in each client is bounded:* $E||g_k(\mathbf{w}_{k,i}^t, \mathcal{B}_{k,i}^t) - \nabla F_k(\mathbf{w}_{k,i}^t)||^2 \leq \sigma^2$.

Assumption

- Assumption 4 is fairly standard in nonconvex optimization literature

Assumption 4. (Bounded Gradient). *The local gradient's expected squared norm is uniformly bounded:* $\mathbb{E}||g_k(\mathbf{w}_{k,i}^t, \mathcal{B}_{k,i}^t)||^2 \leq G^2$ *for all* $k = 1, ..., K, t = 1, ..., T, i = 1, ..., \tau$.

Assumption

- Assumption 5 is used to constrain the optimal client set and the committee client set caused by the committee mechanism

Assumption 5. (Bounded Objective Function). *For any aggregation client set* $S_a \notin \varnothing$ *and the optimal committee client set* $S_c^* \notin \varnothing$, *the difference of local optimal objective function is bounded:* $\mathbb{E}[||\sum_{k \in S_a} p_{k,S_a} F_k^* - \sum_{k' \in S_c^*} p_{k',S_c^*} F_{k'}^*||] \leq \kappa^2$, *where* S_c^* *satisfies that* $S_c^* = \arg\min_{S_c} \sum_{k \in S_c} p_{k,S_c} F_k^*$ *and* $|S_c^*| = C$.

Definition

- The definition of data heterogeneity

Definition 1. ***(Degree of Heterogeneity).*** *We use*

$$\Gamma = F^* - \sum_{k=1}^{K} p_k F_k^* = \sum_{k=1}^{K} p_k (F_k(\mathbf{w}^*) - F_k(\mathbf{w}_k^*))$$

to quantify the degree of heterogeneity among the clients.

- In the Non-IID FL scenario, the Γ remains nonzero and its value reflects the heterogeneity of the data distribution. In the IID FL scenario, with the growth of K it gradually goes to zero

Definition

- The definition of selection gap between aggregation set and committee set caused by committee mechanism

Definition 2. ***(Aggregation-Committee Gap).*** *For any aggregation client set S_a^t, we define*

$$\varphi(S_a^t, \mathbf{w}) = \frac{\mathbb{E}[\sum_{k \in S_a^t} p_{k,S_a^t} F_k(\mathbf{w}) - \sum_{k' \in S_c^*} p_{k',S_c^*} F_{k'}^*]}{F(\mathbf{w}) - \sum_{k=1}^{K} p_k F_k^*} \geq 0, \tag{10}$$

where $\mathbb{E}$ denotes the expectation over all randomness in the previous iterations, and S_c^* denotes the optimal committee client set.

$$\varphi_{min} = \min_{S_a^t, \mathbf{w}} \varphi(S_a^t, \mathbf{w}), \varphi_{max} = \max_{S_a^t} \varphi(S_a^t, \mathbf{w}^*)$$

Convergence Result

- Analysis
  - Selection strategy Ⅰ results in low $\varphi_{\min}$ and slow convergence rate
  - Selection strategy Ⅱ results in high $\varphi_{\min}$ and fast convergence rate

Theorem 1. (Convergence of Committee Mechanism based Federated Learning). *Under Assumption 1 to 5, Definition 1 to 2 and the learning rate η_t, where $\eta_t = \frac{1}{\mu(t+\gamma)}$ and $\gamma = \frac{4L}{\mu}$, the error after T rounds of CMFL satisfies*

$$\begin{aligned}&\mathbb{E}[F(\overline{\mathbf{w}}^T)] - F^* \\ \leq &\frac{1}{T+\gamma}\left[\frac{4L(32\tau^2G^2 + \sum_{k=1}^{K} p_k\sigma_k^2) + 24L^2\kappa^2}{3\mu^2\varphi_{min}}\right. \\ &\left. + \frac{8L^2\Gamma}{\mu^2} + \frac{L\gamma||\overline{\mathbf{w}}^1 - \mathbf{w}^*||^2}{2}\right] + \frac{8L\Gamma}{3\mu}\left(\frac{\varphi_{max}}{\varphi_{min}} - 1\right),\end{aligned} \tag{12}$$

where L, μ and σ represent the constant light indicated in the Assumption 1, 2 and 3. G and κ represent the upper bound values defined in Assumption 4 and 5 .Γ denotes the heterogeneity among the clients according to the Definition

Computation Complexity

- Local Training
  - Computation overhead: $O((N_t + N_c) \mid D \mid \bullet \mid \mathbf{w} \mid)$
  - Computation time: $O(\mid D \mid \bullet \mid \mathbf{w} \mid)$
- Scoring
  - Computation overhead: $O(N_t \bullet N_c \bullet \mid \mathbf{w} \mid^2)$
  - Computation time: $O(N_t \bullet \mid \mathbf{w} \mid^2)$
- Aggregation
  - Computation overhead: $O(m \bullet N_c \bullet \mid \mathbf{w} \mid)$
  - Computation time: $O(m \bullet \mid \mathbf{w} \mid)$
- Other computation ovehead can be ignored

Communication Complexity



- Gradient Uploading
  - Communication overhead: $O(N_t \bullet N_c \bullet |\mathbf{w}|)$
  - Communication time: $O(N_c \bullet |\mathbf{w}| / r)$
- Global Model Downloading
  - Communication overhead: $O(N_t \bullet |\mathbf{w}|)$
  - Communication time: $O(N_t \bullet |\mathbf{w}| / r)$
- Other communication overhead can be ignored

Outline



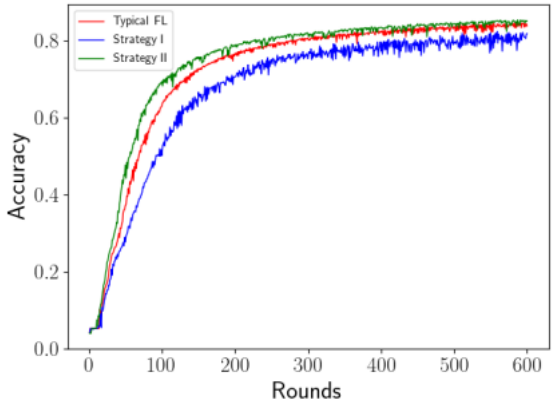
- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

Experimental Settings

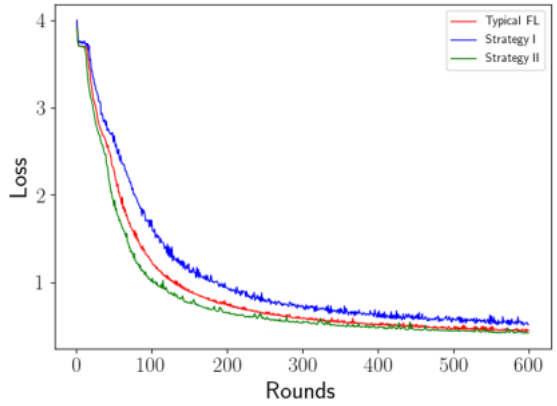
- Dataset & Model
  - FEMNIST-AlexNet
  - Sentiment140-LSTM
- Baseline methods
  - Typical Federated Learning
  - Median
  - Trimmed Mean
  - Krum & Multi-Krum
  - BrainTorrent
  - Gossip Federated Learning

Normal Training Experiment

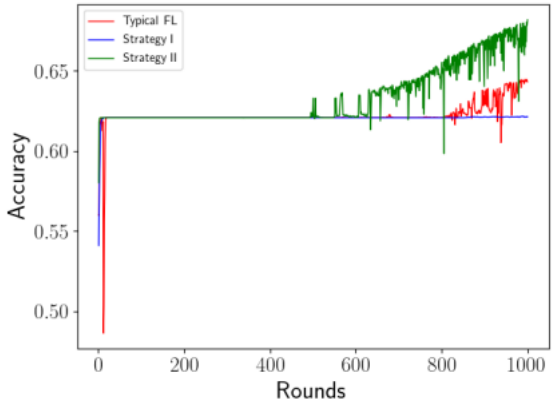
- CMFL under selection strategy Ⅱ achieves the best performance in non-attack scenario



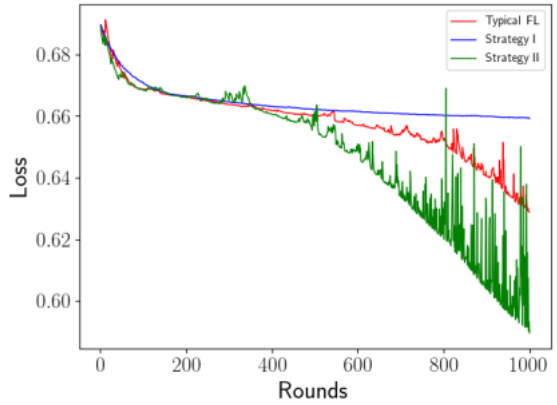
(a) FEMNIST-Accuracy



(b) FEMNIST-Loss



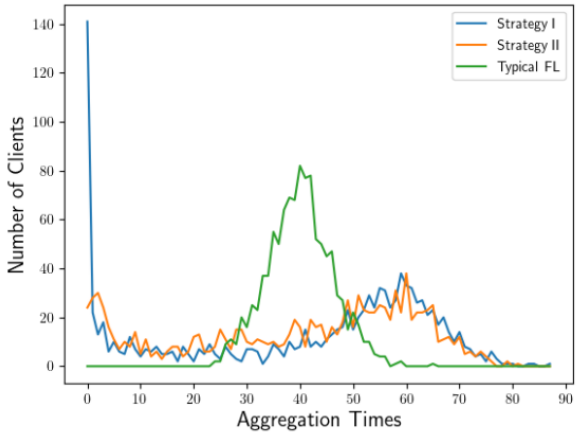
(c) Sentiment140-Accuracy



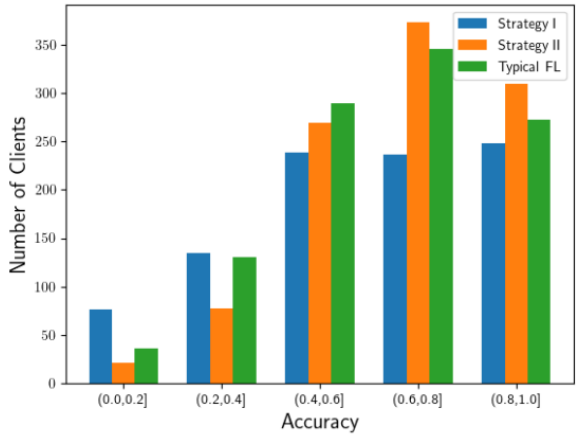
(d) Sentiment140-Loss

Normal Training Experiment

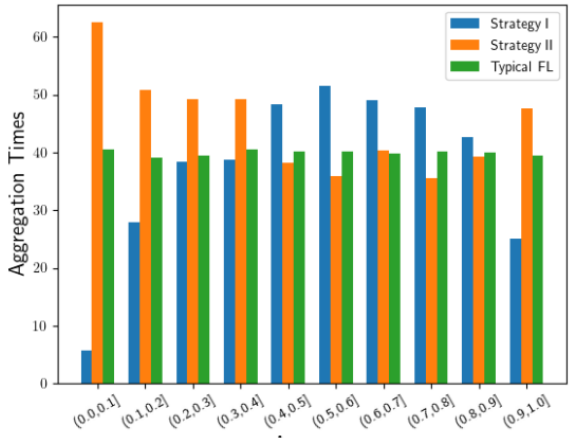
- CMFL under selection strategy Ⅱ allows the global model to learn a more comprehensive knowledge



(a)



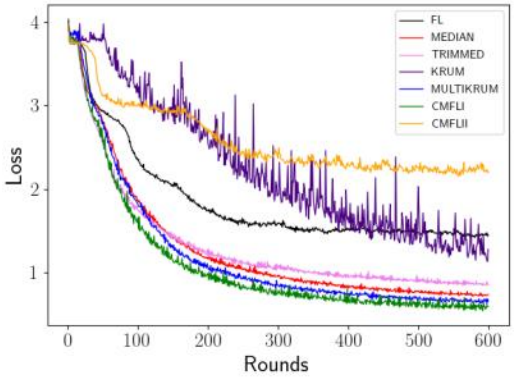
(b)



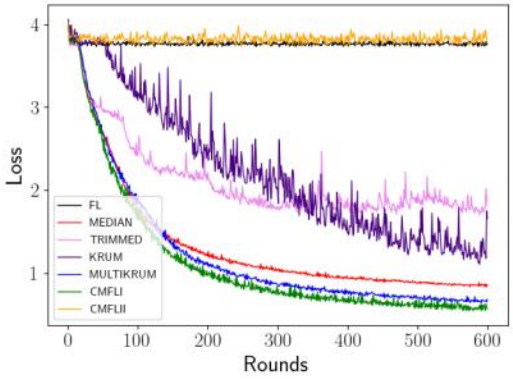
(c)

Robustness Comparative Experiment

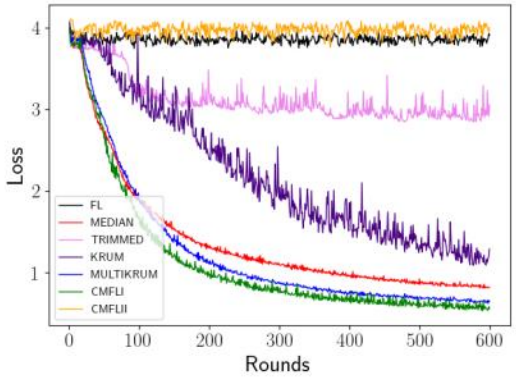
- CMFL under selection strategy Ⅰ achieves the best performance in Byzantine-attack scenario



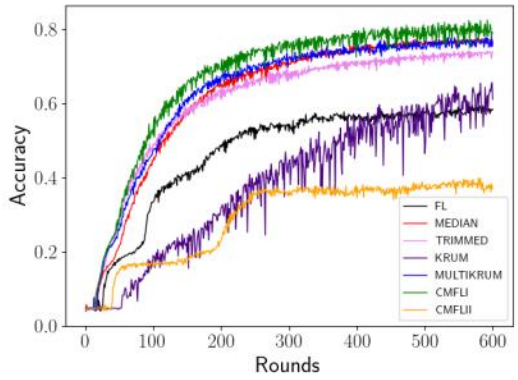
(a) Gradient Scaling Attack Loss



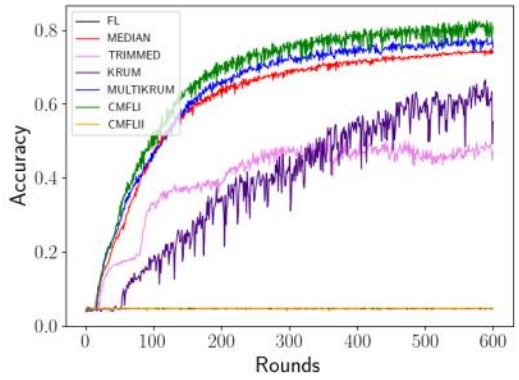
(b) Same-value Attack Loss



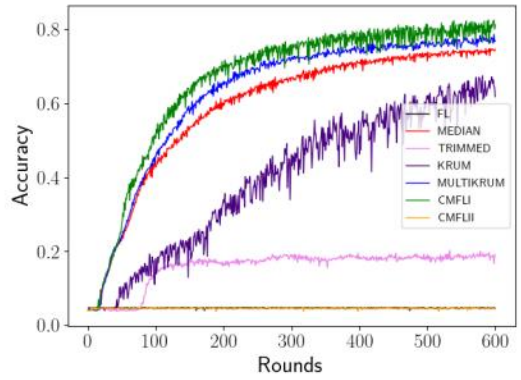
(c) Back-gradient Attack Loss



(d) Gradient Scaling Attack Accuracy



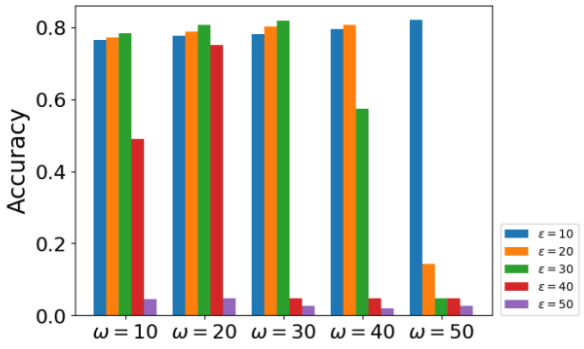
(e) Same-value Attack Accuracy



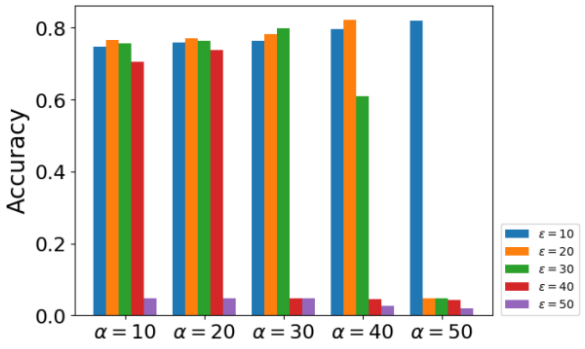
(f) Back-gradient Attack Accuracy

Hyper-parameter Analysis Experiment

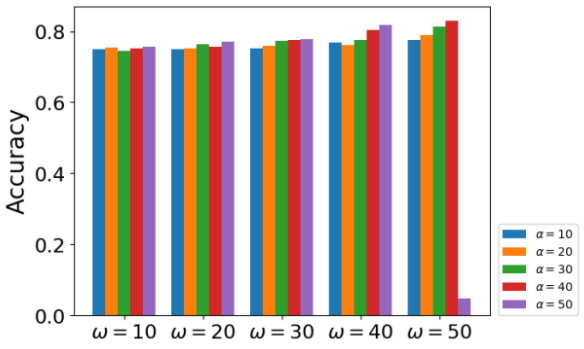
- Extreme parameter settings cause poor performance



(a) $\alpha = 40$



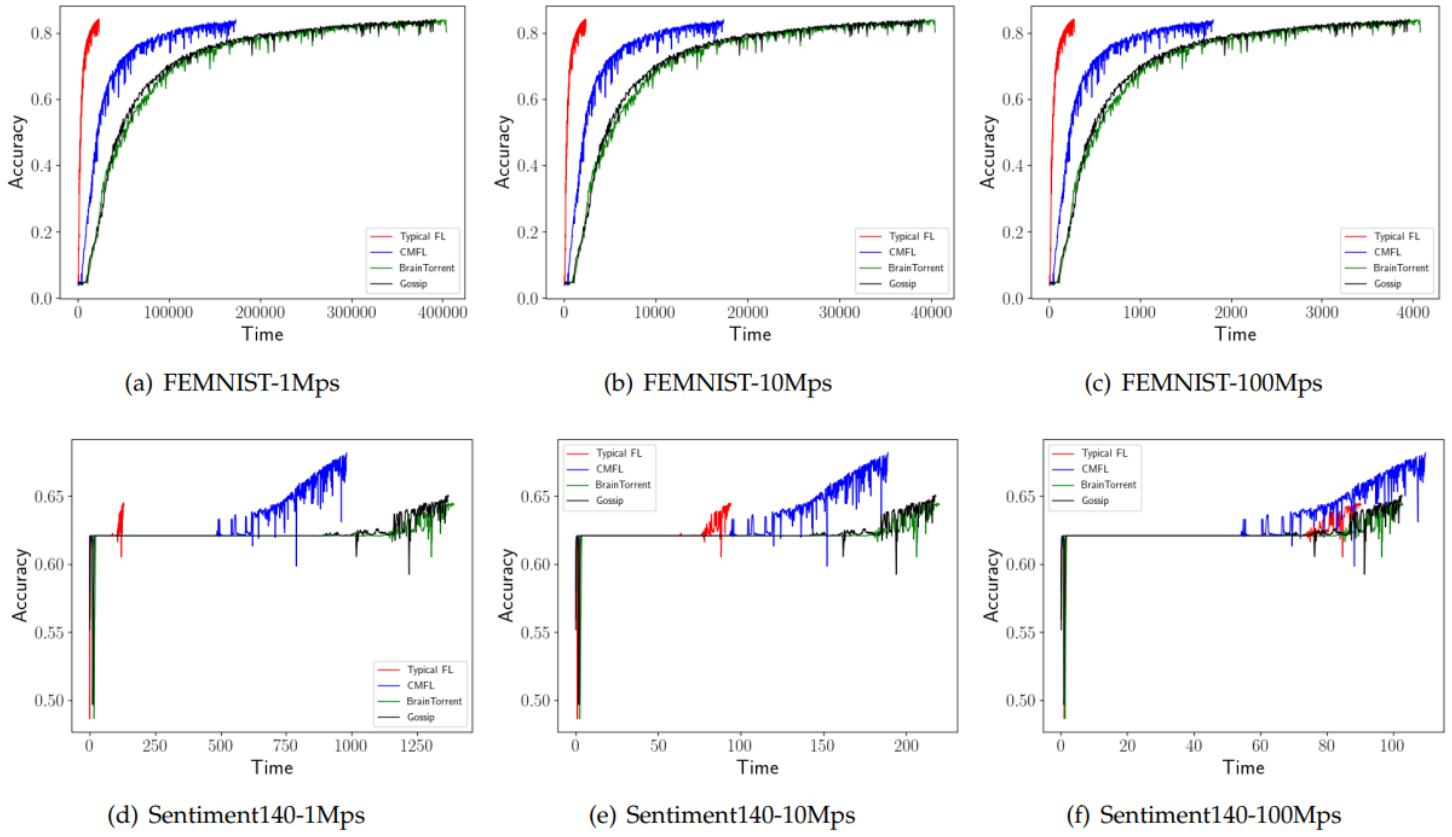
(b) $\omega = 40$



(c) $\epsilon = 10$

Efficiency Experiment

- CMFL achieves a better performance than other two decentralized FL framework (BrainTorrent and Gossip)

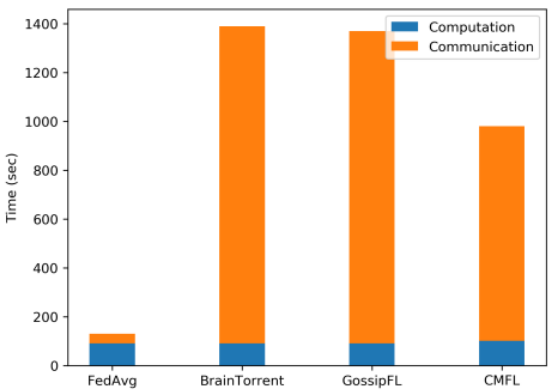


(a) FEMNIST-1Mps (b) FEMNIST-10Mps (c) FEMNIST-100Mps

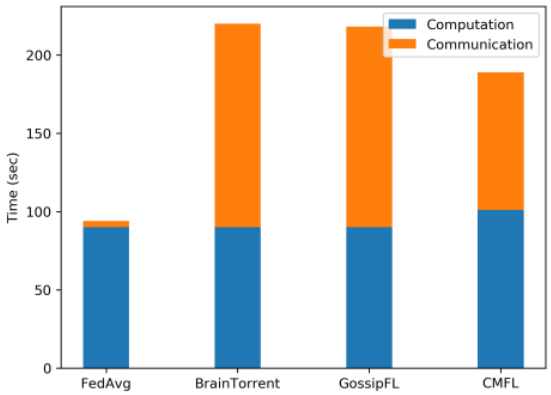
(d) Sentiment140-1Mps (e) Sentiment140-10Mps (f) Sentiment140-100Mps

Efficiency Experiment

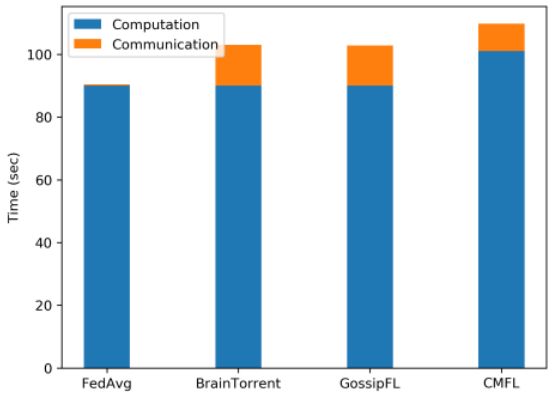
- The communication overhead is lower than other two decentralized FL framework



(a) Sentiment140-1Mps



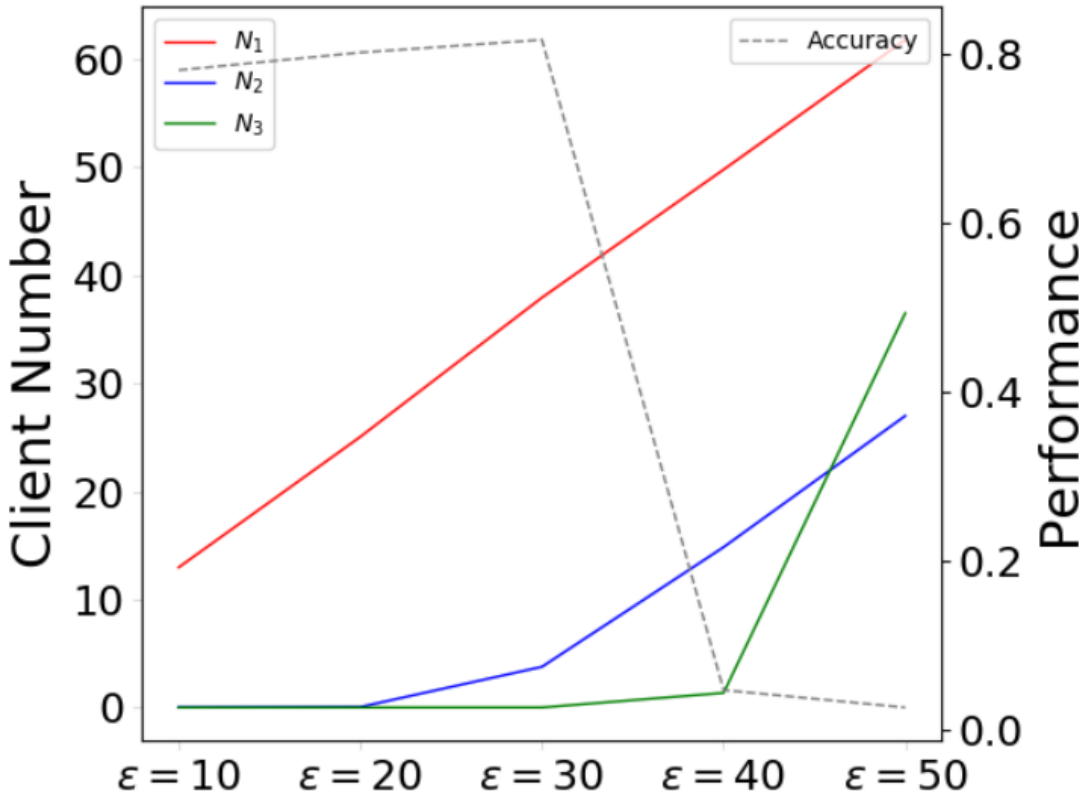
(b) Sentiment140-10Mps



(c) Sentiment140-100Mps

Committee Member Analysis Experiment

- The committee system is robust enough



Outline



- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

Definition

- (t) is defined as the index of the total local SGD iterations
- Local SGD training

$$\mathbf{w}_k^{(t+1)} = \mathbf{w}_k^{(t)} - \eta_t g_k(\mathbf{w}_k^{(t)}, \mathcal{B}_k^{(t)}),$$

- Aggregation

$$\overline{\mathbf{w}}^{(t+1)} = \overline{\mathbf{w}}^{(t)} - \eta_t \overline{g}^{(t)} = \overline{\mathbf{w}}^{(t)} - \eta_t \Big(\sum_{k \in S_a^{(t)}} p_{k, S_a^{(t)}} g_k(\mathbf{w}_k^{(t)}, \mathcal{B}_k^{(t)}) \Big),$$

- Local Updating

$$\mathbf{w}_k^{(t+1)} = \begin{cases} \mathbf{w}_k^{(t)} - \eta_t g_k(\mathbf{w}_k^{(t)}, \mathcal{B}_k^{(t)}), (t+1)\%\tau \neq 0 \\ \overline{\mathbf{w}}^{(t+1)}, (t+1)\%\tau = 0. \end{cases}$$

Lemma

Lemma 1. ***(Global-Local Gradient Product).*** *According to the definitions of* $\nabla F_k(\mathbf{w}_k^{(t)})$*, we have that*

$$
\begin{aligned}
&-\langle \overline{\mathbf{w}}^{(t)} - \mathbf{w}^*, \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) \rangle \\
\leq &\frac{1}{2\eta_t} \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2 + \frac{\eta_t}{2} \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\nabla F_k(\mathbf{w}_k^{(t)})||^2 - \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}_k^{(t)}) - F_k(\mathbf{w}^*)) \\
&- \frac{\mu}{2} \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\mathbf{w}_k^{(t)} - \mathbf{w}^*||^2.
\end{aligned}
$$

Lemma 2. ***(Local Parameter-Optimal Gap).*** *Let* $v_t = 2\eta_t - 4L\eta_t^2 = 2\eta_t(1 - 2L\eta_t)$*, we have that*

$$
\begin{aligned}
&- \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}_k^{(t)}) - F_k^*) \\
\leq &-(1 - \eta_t L) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\overline{\mathbf{w}}^{(t)}) - F^*) + \frac{1}{v_t} \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\mathbf{w}_k^{(t)} - \overline{\mathbf{w}}^{(t)}||^2.
\end{aligned}
$$

Lemma 3. ***(Client Heterogeneous Bound).*** *According to the Assumption 4, we have that*

$$
\mathbb{E}[\sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2] \leq 16\eta_t^2 \tau^2 G^2.
$$

Proof of Theorem 1

- According to the global updating:

$$\mathcal{H}(\mathbf{w}, t+1) = ||\overline{\mathbf{w}}^{(t+1)} - \mathbf{w}^*||^2 = ||\overline{\mathbf{w}}^{(t)} - \eta_t \overline{g}^{(t)} - \mathbf{w}^*||^2.$$

- Above formula can be expanded as:

$$\begin{aligned}
&\mathcal{H}(\mathbf{w}, t+1) \\
=&||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^* - \eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)})||^2 + ||\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) - \eta_t \overline{g}^{(t)}||^2 \\
&\underbrace{+ 2\eta_t \langle \overline{\mathbf{w}}^{(t)} - \mathbf{w}^* - \eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}), \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) - \overline{g}^{(t)} \rangle}_{A_1}.
\end{aligned}$$

Proof of Theorem 1

- Due to Assumption 3 and Lemma 1:

$$\begin{aligned}
&\mathcal{H}(\mathbf{w}, t+1) \\
\leq &||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^*||^2 + \underbrace{4L\eta_t^2 \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}_k^{(t)}) - F_k^*)}_{A_2} \underbrace{- 2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}_k^{(t)}) - F_k(\mathbf{w}^*))}_{A_3} \\
&+ \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2 - \eta_t \mu \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\mathbf{w}_k^{(t)} - \mathbf{w}^*||^2 + ||\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) - \eta_t \overline{g}^{(t)}||^2 + A_1.
\end{aligned}$$

- The difference of A2 and A3 can be written as:

$$\begin{aligned}
&A_2 - A_3 \\
=&(4L\eta_t^2 - 2\eta_t) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k(\mathbf{w}_k^{(t)}) - 4L\eta_t^2 \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k^* + 2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k(\mathbf{w}^*) \\
=&(4L\eta_t^2 - 2\eta_t) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k(\mathbf{w}_k^{(t)}) - (4L\eta_t^2 - 2\eta_t) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k^* + 2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k(\mathbf{w}^*) - 2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} F_k^* \\
=&(4L\eta_t^2 - 2\eta_t) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}_k^{(t)}) - F_k^*) + 2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}^*) - F_k^*).
\end{aligned}$$

Proof of Theorem 1

- Let $v_t = 2\eta_t - 4L\eta_t^2 = 2\eta_t(1-2L\eta_t)$:

$$\begin{aligned}
&\mathcal{H}(\mathbf{w}, t+1) \\
\leq &||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^*||^2 - v_t \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}}(F_k(\mathbf{w}_k^{(t)}) - F_k^*) + 2\eta_t \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}}(F_k(\mathbf{w}^*) - F_k^*) + \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2 \\
&- \eta_t\mu \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} ||\mathbf{w}_k^{(t)} - \mathbf{w}^*||^2 + ||\eta_t \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) - \eta_t \overline{g}^{(t)}||^2 + A_1.
\end{aligned}$$

- Due to Lemma 2:

$$\begin{aligned}
&\mathcal{H}(\mathbf{w}, t+1) \\
\leq &||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^*||^2 + 2\eta_t \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}}(F_k(\mathbf{w}^*) - F_k^*) + 2 \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2 \\
&- v_t(1-\eta_t L) \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}}(F_k(\overline{\mathbf{w}}^{(t)}) - F_k^*) - \eta_t\mu \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} ||\mathbf{w}_k^{(t)} - \mathbf{w}^*||^2 \\
&+ ||\eta_t \sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) - \eta_t \overline{g}^{(t)}||^2 + A_1.
\end{aligned}$$

Proof of Theorem 1

- Solve the expectation:

$$\begin{aligned}
&\mathbb{E}[\mathcal{H}(\mathbf{w}, t+1)] = \mathbb{E}[||\overline{\mathbf{w}}^{(t+1)} - \mathbf{w}^*||^2] \\
\leq &\mathbb{E}[||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^*||^2] + \mathbb{E}[2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}^*) - F_k^*)] + \mathbb{E}[2 \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2] \\
&- \mathbb{E}[v_t(1 - \eta_t L) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\overline{\mathbf{w}}^{(t)}) - F_k^*)] - \mathbb{E}[\eta_t \mu \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\mathbf{w}_k^{(t)} - \mathbf{w}^*||^2] \\
&+ \mathbb{E}[||\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} \nabla F_k(\mathbf{w}_k^{(t)}) - \eta_t \overline{g}^{(t)}||^2] + \mathbb{E}[A_1].
\end{aligned}$$

- Due to Assumption 3 and the expectation of A1 is zero:

$$\begin{aligned}
&\mathbb{E}[||\overline{\mathbf{w}}^{(t+1)} - \mathbf{w}^*||^2] \\
=&(1 - \eta_t \mu) \mathbb{E}[||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^*||^2] + \mathbb{E}[\mathcal{Q}(\mathbf{w}, k, t)] + \eta_t^2 \sum_{k=1}^{K} p_k^2 \sigma_k^2,
\end{aligned}$$

where

$$\begin{aligned}
&\mathcal{Q}(\mathbf{w}, k, t) \\
=&2\eta_t \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\mathbf{w}^*) - F_k^*) + 2 \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} ||\overline{\mathbf{w}}^{(t)} - \mathbf{w}_k^{(t)}||^2 - v_t(1 - \eta_t L) \sum_{k \in S_a^{(t)}} p_{k,S_a^{(t)}} (F_k(\overline{\mathbf{w}}^{(t)}) - F_k^*).
\end{aligned}$$

Proof of Theorem 1

- Due to Lemma 3:

$$\mathbb{E}[\mathcal{Q}(\mathbf{w},k,t)]$$

$$\begin{aligned}&= -v_t(1-\eta_t L)\mathbb{E}[\sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} F_k(\overline{\mathbf{w}}^{(t)}) - \sum_{k'\in S_c^*} p_{k',S_c^*} F_{k'}^*] + 2\eta_t(\mathbb{E}[\sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} F_k(\mathbf{w}^*) - \sum_{k'\in S_c^*} p_{k',S_c^*} F_{k'}^*] \\ &\quad - (2\eta_t - v_t(1-\eta_t L))\mathbb{E}[\sum_{k\in S_a^{(t)}} p_{k,S_a^{(t)}} F_k^* - \sum_{k'\in S_c^*} p_{k',S_c^*} F_{k'}^*] + 32\eta_t^2\tau^2 G^2.\end{aligned}$$

- According to the Assumption 5 and Definition 1 and 2:

$$\mathbb{E}[\mathcal{Q}(\mathbf{w},k,t)]$$

$$\leq \underbrace{-v_t(1-\eta_t L)\varphi_{min}\mathbb{E}[(F(\overline{\mathbf{w}}) - \sum_{k=1}^{K} p_k F_k^*)]}_{A_4} + 2\eta_t\varphi_{max}\Gamma + 6L\eta_t^2\kappa^2 + 32\eta_t^2\tau^2 G^2.$$

- A4 can be expanded as:

$$A_4 \leq -\frac{3\eta_t\mu\varphi_{min}}{8}\mathbb{E}[||\overline{\mathbf{w}}^{(t)} - \mathbf{w}^*||^2] - 2\eta_t\varphi_{min}\Gamma + 6\eta_t^2\varphi_{min}L\Gamma.$$

Proof of Theorem 1

So we have

$$\begin{aligned}&\mathbb{E}[\mathcal{Q}(\mathbf{w},k,t)]\\&=-\frac{3\eta_t\mu\varphi_{min}}{8}\mathbb{E}[||\overline{\mathbf{w}}^{(t)}-\mathbf{w}^*||^2]+2\eta_t\Gamma(\varphi_{max}-\varphi_{min})+\eta_t^2(6\varphi_{min}L\Gamma+32\tau^2G^2+6L\kappa^2).\end{aligned}\tag{34}$$

As a result, we have

$$\begin{aligned}&\mathbb{E}[||\overline{\mathbf{w}}^{(t+1)}-\mathbf{w}^*||]\\&\leq[1-\eta_t\mu(1+\frac{3\varphi_{min}}{8})]\mathbb{E}[||\overline{\mathbf{w}}^{(t)}-\mathbf{w}^*||^2]+2\eta_t\Gamma(\varphi_{max}-\varphi_{min})+\eta_t^2(6\varphi_{min}L\Gamma+32\tau^2G^2+6L\kappa^2+\sum_{k=1}^{K}p_k\sigma_k^2).\end{aligned}\tag{35}$$

By defining $\Delta_{t+1}=\mathbb{E}[||\overline{\mathbf{w}}^{(t+1)}-\mathbf{w}^*||]$, $B=1+\frac{3\varphi_{min}}{8}$, $C=6\varphi_{min}L\Gamma+32\tau^2G^2+6L\kappa^2+\sum_{k=1}^{K}p_k\sigma_k^2$, $D=\Gamma(\varphi_{max}-\varphi_{min})$, we have

$$\Delta_{t+1}\leq(1-\eta_t\mu B)\Delta_t+\eta_t^2C+\eta_tD.\tag{36}$$

If we set $\Delta_t\leq\frac{\psi}{t+\gamma}$, $\eta_t=\frac{\beta}{t+\gamma}$ and $\beta>\frac{1}{\mu B}$, $\gamma>0$ by induction, we have

$$\psi=\max\left\{\gamma||\overline{\mathbf{w}}^1-\mathbf{w}^*||^2,\frac{1}{\beta\mu B-1}(\beta^2C+D\beta(t+\gamma))\right\}.\tag{37}$$

Then by the L-smoothness of $F(\cdot)$,

$$\mathbb{E}[F(\overline{\mathbf{w}}^{(t)})]-F^*\leq\frac{L}{2}\Delta_t\leq\frac{L}{2}\frac{\psi}{\gamma+t}.\tag{38}$$

Finally, we complete the proof of Theorem 1:

$$\begin{aligned}&\mathbb{E}[F(\overline{\mathbf{w}}^T)]-F^*\\&\leq\frac{1}{T+\gamma}\left[\frac{4L(32\tau^2G^2+\sum_{k=1}^{K}p_k\sigma_k^2)+24L^2\kappa^2}{3\mu^2\varphi_{min}}+\frac{8L^2\Gamma}{\mu^2}+\frac{L\gamma||\overline{\mathbf{w}}^1-\mathbf{w}^*||^2}{2}\right]+\frac{8L\Gamma}{3\mu}\left(\frac{\varphi_{max}}{\varphi_{min}}-1\right),\end{aligned}\tag{39}$$

where the T means the maximal communication rounds, which satisfies $T=i\tau$ for $i=1,2,...$ in realistic scenarios. □

Outline



- Background
- Problem Formulation
- Framework
- Theoretical Analysis
- Experimental Results
- Proof of Theorem
- Conclusion

Conclusion

- Advantages
  - Strong Robustness
  - Decentralized Setting
  - High Communication Efficiency
  - Theoretical Guarantee
  - Scalable Framework
  - Better Performance on Non-IID dataset
- Future Works
  - Backdoor attack
  - System Heterogeneity

Thank You!
Q & A