Robust federated learning based on voting and scaling

Xiang-Yu Liang
School of Computer Science
Southwest Petroleum University
Chengdu 610500, China
liangxyswpu@163.com

Heng-Ru Zhang*
School of Computer Science
Southwest Petroleum University
Chengdu 610500, China
zhanghrswpu@163.com

Fan Min
School of Computer Science
Southwest Petroleum University
Chengdu 610500, China
minfan@swpu.edu.cn

Wei Tang
School of Computer Science
Southwest Petroleum University
Chengdu 610500, China
twei1123@163.com

I. Formal Security Analysis

Assumption 1: The global model objective function $F(\boldsymbol{w})$ is L-strongly convex and has an M-Lipschitz continuous gradient on $\boldsymbol{\omega}$. For any $\boldsymbol{w}, \boldsymbol{w}' \in \boldsymbol{\omega}$, we have the following:

$$F(\boldsymbol{w}) + \langle \nabla F(\boldsymbol{w}), \boldsymbol{w}' - \boldsymbol{w} \rangle + \frac{L}{2}\|\boldsymbol{w}' - \boldsymbol{w}\|^2 \leq F(\boldsymbol{w}'),$$
$$\|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}')\| \leq M\|\boldsymbol{w} - \boldsymbol{w}'\|,$$

where $\langle \cdot, \cdot \rangle$ is the inner product of two vectors, ∇ is the gradient, and $\|\cdot\|$ is the ℓ_2 norm.

Assumption 2: There exist positive constants σ_1 and γ_1 such that for every unit vector $\boldsymbol{v} \in \boldsymbol{B}$, $\langle \nabla f(D, \boldsymbol{w}^*), \boldsymbol{v} \rangle$ is sub-exponential with scaling parameters σ_1 and γ_1, i.e.,

$$\sup_{\boldsymbol{v} \in \boldsymbol{B}} \mathbb{E}[\exp(\lambda \langle \nabla f(D, \boldsymbol{w}^*), \boldsymbol{v} \rangle)] \leq e^{\sigma_1^2 \lambda^2/2}, \quad \forall |\lambda| \leq \frac{1}{\gamma_1},$$

where $\boldsymbol{B}$ denotes the unit sphere $\{\boldsymbol{v} : \|\boldsymbol{v}\| = 1\}$.

Assumption 2 is to ensure that the client uses the local dataset with high probability to find the optimal model $\boldsymbol{w}^*$. Specifically, $(1/|D_i|)\sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}^*)$ is concentrated near $\nabla F(\boldsymbol{w}^*) = 0$, where $|D_i|$ is represented as the number of elements of D_i.

Next, we define gradient difference:

$$h(D, \boldsymbol{w}) \triangleq \nabla f(D, \boldsymbol{w}) - \nabla f(D, \boldsymbol{w}^*), \tag{1}$$

which expresses the deviation of the empirical loss function from the optimal global model. Note that

$$\mathbb{E}[h(D, \boldsymbol{w})] = \nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}^*), \tag{2}$$

for each $\boldsymbol{w}$.

Assumption 3: There exist positive constant σ_2 and γ_2 such that for any $\boldsymbol{w} \in \boldsymbol{\omega}$ with $\boldsymbol{w} \neq \boldsymbol{w}^*$ and any unit vector $v \in \boldsymbol{B}$, $\langle h(D, \boldsymbol{w}) - \mathbb{E}[h(D, \boldsymbol{w})], \boldsymbol{v} \rangle / \|\boldsymbol{w} - \boldsymbol{w}^*\|$ is sub-exponential with scaling parameters σ_2 and γ_2, i.e., for all $|\lambda| < 1/\gamma_2$,

$$\sup_{\boldsymbol{w} \in \boldsymbol{\omega}, \boldsymbol{v} \in \boldsymbol{B}} \mathbb{E}[\exp(\frac{\lambda \langle h(D, \boldsymbol{w}) - \mathbb{E}[h(D, \boldsymbol{w})], \boldsymbol{v} \rangle}{\|\boldsymbol{w} - \boldsymbol{w}^*\|})] \leq e^{\sigma_2^2 \lambda^2/2},$$

where $\boldsymbol{B}$ denotes the unit sphere $\{\boldsymbol{v} : \|\boldsymbol{v}\| = 1\}$.

Assumption 4: For any $\delta \in (0, 1)$, there exists an $M' = M'(|D_i|, \delta)$ that is non-increasing in $|D_i|, \delta$ such that

$$\mathbb{P}\{\sup_{\boldsymbol{w}, \boldsymbol{w}' \in \boldsymbol{\omega} : \boldsymbol{w} \neq \boldsymbol{w}'} \frac{\|\nabla \bar{f}_{|D_i|}(\boldsymbol{w}) - \nabla \bar{f}_{|D_i|}(\boldsymbol{w}')\|}{\|\boldsymbol{w} - \boldsymbol{w}'\|} \leq M'\} \geq 1 - \frac{\delta}{3},$$

where $\nabla \bar{f}_{|D_i|}(\boldsymbol{w}) = (\sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}))/|D_i|$.

Assumption 5: Each local training dataset D_i ($i = 1, 2, \ldots, n$) is sampled from distribution $\mathcal{X}$.

Theorem 1: Suppose Assumptions 1-5 hold, learning rate $\alpha = L/2M^2$, $\delta \in (0, 1)$, $\Delta_1 \geq \sigma_1^2/\gamma_1$, $\Delta_2 \geq \sigma_2^2/\gamma_2$, and $\boldsymbol{\omega} \subset \{\boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{w}^*\| \leq r\sqrt{d}\}$ for some positive parameter r, for any number of malicious clients, the difference between the global model aggregated by VSRFL and the optimal global model $\boldsymbol{w}^*$ without attack is bounded. For any $t \geq 1$, we have:

$$\|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq (1 - \rho)^t \|\boldsymbol{w}^0 - \boldsymbol{w}^*\| + \frac{12\alpha\Delta_1}{\rho}.$$

where $\boldsymbol{w}^t$ is the global model of the aggregation for each epoch, $\rho = 1 - (\sqrt{1 - L^2/(4M^2)} + 24\alpha\Delta_2 + 2\alpha M)$, $\Delta_1 = \sigma_1\sqrt{2/|D_i|}\sqrt{d\log 6 + \log(3/\delta)}$, $\Delta_2 = \sigma_2\sqrt{\frac{2}{|D_i|}}\sqrt{d\log\frac{18M \vee M'}{\sigma_2} + \frac{1}{2}d\log\frac{|D_i|}{d} + \log(\frac{6\sigma_2^2 r\sqrt{|D_i|}}{\gamma_2\sigma_1\delta}}$), $M \vee M' = \max(M, M')$, d is the dimension of $\boldsymbol{w}$. When $|1 - \rho| < 1$, we have $\lim_{t\to\infty} \|\boldsymbol{w}^t - \boldsymbol{w}^*\| \leq 12\alpha\Delta_1/\rho$.

II. The proving process

Recall that the optimal global model $\boldsymbol{w}^*$ is a answer to the following optimization problem: $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} F(\boldsymbol{w})$, where $F(\boldsymbol{w}) = \mathbb{E}_{D\sim\mathcal{X}}[f(D, \boldsymbol{w})]$ is the expectation of the empirical loss $f(D, \boldsymbol{w})$ on the joint training dataset D. We show that the difference between the global model trained by VSRFL and the optimal global model $\boldsymbol{w}^*$ is bounded under certain assumptions. We denote the local update set filtered by the server in epoch t by $\mathcal{S}$. We let $\hat{\boldsymbol{g}}_i = (\|\boldsymbol{g}_{median}\|/\|\boldsymbol{g}_i\|) \times \boldsymbol{g}_i$, where $i \in \mathcal{S}$ s.t. $|\mathcal{S}| > 0$. We let $|D_i|$ is size of the local dataset. We first describe our lemmas and then state our theoretical results.

Lemma 1: For any number of abnormal local updates, the gap between the global update $\boldsymbol{g}$ and the gradient $\nabla F(\boldsymbol{w})$ is bounded:

$$\|\boldsymbol{g}-\nabla F(\boldsymbol{w})\| \leq 3\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\|+2\|\nabla F(\boldsymbol{w})\|,$$

where $\boldsymbol{g}_{median}$ is the median updated of the server selection in each epoch.

Proof: We have the following equations:

$$\begin{aligned}
&\|\boldsymbol{g}-\nabla F(\boldsymbol{w})\| \\
&=\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{g}}_i-\nabla F(\boldsymbol{w})\| \\
&=\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{g}}_i-\boldsymbol{g}_{median}+\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\leq\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{g}}_i-\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&=\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{g}}_i+(-\boldsymbol{g}_{median})\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\leq\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{g}}_i\|+\|-\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\overset{(a)}{=}\|\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\boldsymbol{g}}_i\|+\|\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\leq\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\|\hat{\boldsymbol{g}}_i\|+\|\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\overset{(b)}{=}\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\|\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\overset{(c)}{=}2\|\boldsymbol{g}_{median}\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&=2\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})+\nabla F(\boldsymbol{w})\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&\leq 2\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\|+2\|\nabla F(\boldsymbol{w})\|+\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\| \\
&=3\|\boldsymbol{g}_{median}-\nabla F(\boldsymbol{w})\|+2\|\nabla F(\boldsymbol{w})\|, \qquad (3)
\end{aligned}$$

where (a) is because of the following equations:

$$\sqrt{g_1^2+g_2^2+\cdots+g_i^2}=\sqrt{(-g_1)^2+(-g_2)^2+\cdots+(-g_i)^2}, \\ \text{s.t. } \boldsymbol{g}=\{g_1,g_2,\cdots,g_i\}; \qquad (4)$$

(b) is because VSRFL normalizes the filtered local updates to have the same magnitude as the median, i.e., $\|\hat{\boldsymbol{g}}_i\|=\|\boldsymbol{g}_{median}\|$; and (c) is because $|\mathcal{S}|$ is represented as the number of elements of $\mathcal{S}$, e.g., $(\sum_{i\in\mathcal{S}}x)/|\mathcal{S}|=x$.

Lemma 2: Suppose Assumption 1 holds. If we choose the learning rate $\alpha=L/2M^2$, there is the following inequality:

$$\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*-\alpha\nabla F(\boldsymbol{w}^{t-1})\|\leq\sqrt{(1-\frac{L^2}{4M^2})}\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\| \\ \text{s.t. } t\geq 1.$$

Proof: By Assumption 1, we have:

$$\|\nabla F(\boldsymbol{w}^{t-1})-\nabla F(\boldsymbol{w}^*)\|\leq M\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|, \qquad (5)$$

$$F(\boldsymbol{w}^{t-1})\geq F(\boldsymbol{w}^*)+\langle\nabla F(\boldsymbol{w}^*),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle \\ +\frac{L}{2}\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2, \qquad (6)$$

$$F(\boldsymbol{w}^*)\geq F(\boldsymbol{w}^{t-1})+\langle\nabla F(\boldsymbol{w}^{t-1}),\boldsymbol{w}^*-\boldsymbol{w}^{t-1}\rangle. \qquad (7)$$

Combining equations 6 and 7, we have:

$$\begin{aligned}
&\langle\nabla F(\boldsymbol{w}^*),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle+\langle\nabla F(\boldsymbol{w}^{t-1}),\boldsymbol{w}^*-\boldsymbol{w}^{t-1}\rangle \\
&=\langle\nabla F(\boldsymbol{w}^*),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle-\langle\nabla F(\boldsymbol{w}^{t-1}),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle \\
&=\langle\nabla F(\boldsymbol{w}^*)-\nabla F(\boldsymbol{w}^{t-1}),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle \\
&=-\langle\nabla F(\boldsymbol{w}^{t-1})-\nabla F(\boldsymbol{w}^*),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle \\
&\leq-\frac{L}{2}\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2. \qquad (8)
\end{aligned}$$

Due to the that $\nabla F(\boldsymbol{w}^*)=\boldsymbol{0}$, we have the following:

$$\begin{aligned}
&\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*-\alpha\nabla F(\boldsymbol{w}^{t-1})\|^2 \\
&=\|-\alpha(\nabla F(\boldsymbol{w}^{t-1})-\nabla F(\boldsymbol{w}^*))+\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2 \\
&=\alpha^2\|(\nabla F(\boldsymbol{w}^{t-1})-\nabla F(\boldsymbol{w}^*))\|^2+\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2 \\
&-2\alpha\langle\nabla F(\boldsymbol{w}^{t-1})-\nabla F(\boldsymbol{w}^*),\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\rangle \\
&\leq\alpha^2M^2\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2+\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2 \\
&-\alpha L\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2 \\
&=(1+\alpha^2M^2-\alpha L)\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2. \qquad (9)
\end{aligned}$$

We let $\alpha=L/2M^2$, therefore:

$$\begin{aligned}
&\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*-\alpha\nabla F(\boldsymbol{w}^{t-1})\|^2 \\
&\leq(1+\alpha^2M^2-\alpha L)\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2 \\
&=(1-\frac{L^2}{4M^2})\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|^2, \qquad (10)
\end{aligned}$$

which concludes the proof.

Lemma 3: Suppose Assumption 2 holds. For any $\delta\in(0,1)$ and any positive integer $|D_i|$, we let

$$\Delta_1(|D_i|,d,\delta,\sigma_1)=\sqrt{2}\sigma_1\sqrt{\frac{d\log 6+\log(3/\delta)}{|D_i|}}.$$

We let $\Delta_1=\Delta_1(|D_i|,d,\delta,\sigma_1)$. If $\Delta_1\leq\sigma_1^2/\gamma_1$, we have

$$\mathbb{P}\{\|\frac{1}{|D_i|}\sum_{X_j\in D_i}\nabla f(X_j,\boldsymbol{w}^*)-\nabla F(\boldsymbol{w}^*)\|\geq 2\Delta_1\}\leq\frac{\delta}{3}.$$

For fixed δ and σ_1, if $d=o(|D_i|)$,

$$\Delta_1=\sqrt{2}\sigma_1\sqrt{\frac{d\log 6+\log(3/\delta)}{|D_i|}}\to 0\ as\ |D_i|\to\infty.$$

So, if γ_1 is fixed, $\Delta_1\leq\sigma_1^2/\gamma_1$ holds when l is large enough.

Proof: We let $\mathcal{V}=\{v_1,\ldots,v_{N_{1/2}}\}$ denote an $\frac{1}{2}$-cover of unit sphere $\boldsymbol{B}$. It is show in [1], [2] that $\log N_{1/2}\leq d\log 6$, and

$$\begin{aligned}
&\|\frac{1}{|D_i|}\sum_{X_j\in D_i}\nabla f(X_j,\boldsymbol{w}^*)-\nabla F(\boldsymbol{w}^*)\| \\
&\leq 2\sup_{v\in\mathcal{V}}\{\langle\frac{1}{|D_i|}\sum_{X_j\in D_i}\nabla f(X_j,\boldsymbol{w}^*)-\nabla F(\boldsymbol{w}^*),v\rangle\}. \qquad (11)
\end{aligned}$$

By Assumption 2, the condition $\Delta_1 \le \sigma_1^2/\gamma_1$, and the concentration inequalities for sub-exponential random variables, for $v \in \mathcal{V}$ we have:

$$\mathbb{P}\{\langle \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}^*) - \nabla F(\boldsymbol{w}^*), v\rangle \ge \Delta_1\} \le \exp(-\frac{|D_i|\Delta_1^2}{2\sigma_1^2}). \quad (12)$$

Recall that in $\mathcal{V}$ contains at most 6^d vetors. In view of the union bound, if further yields that

$$\begin{aligned}\mathbb{P}\{2\sup_{v\in\mathcal{V}}\{\langle \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}^*) - \nabla F(\boldsymbol{w}^*), v\rangle\} \ge 2\Delta_1\} &\le 6^d \exp(-\frac{|D_i|\Delta_1^2}{2\sigma_1^2}) \\ &= \exp(-\frac{|D_i|\Delta_1^2}{2\sigma_1^2} + d\log 6). \quad (13)\end{aligned}$$

Therefore,

$$\mathbb{P}\{\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}^*) - \nabla F(\boldsymbol{w}^*)\| \ge 2\Delta_1\} \le \exp(-\frac{|D_i|\Delta_1^2}{2\sigma_1^2} + d\log 6). \quad (14)$$

We conclude the proof by equation $\Delta_1 = \sqrt{2}\sigma_1\sqrt{(d\log 6 + \log(3/\delta))/|D_i|}$.

***Lemma** 4:* Suppose Assumption 3 holds and fix any $\boldsymbol{w} \in \boldsymbol{\omega}$. We let

$$\Delta_1'(|D_i|, d, \delta, \sigma_2) = \sqrt{2}\sigma_2\sqrt{\frac{d\log 6 + \log(3/\delta)}{|D_i|}}.$$

We let $\Delta_1' = \Delta_1'(|D_i|, d, \delta, \sigma_2)$. If $\Delta_1' \le \sigma_2^2/\gamma_2$, then

$$\mathbb{P}\{\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla h(X_j, \boldsymbol{w}) - \mathbb{E}[h(X, \boldsymbol{w})]\| \ge 2\Delta_1'(\boldsymbol{w} - \boldsymbol{w}^*)\} \le \frac{\delta}{3}.$$

Similar to Δ_1, if δ, σ_1 and σ_2 are fixed, and $d = o(|D_i|)$, then for all sufficiently large l, it holds that $\Delta_1'(l, d, \delta, \sigma_2) \le \sigma_2^2/\gamma_2$.

Proof: It is similar to the proof of Lemma 3. Let $\mathcal{V} = \{v_1, \dots, v_{N_{1/2}}\}$ denote an $\frac{1}{2}$-cover of unit sphere $\boldsymbol{B}$. This exists $\log N_{1/2} \le d\log 6$, and

$$\begin{aligned}&\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla h(X_j, \boldsymbol{w}) - \mathbb{E}[h(X, \boldsymbol{w})]\| \\ &\le 2\sup_{v\in\mathcal{V}}\{\langle \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla h(X_j, \boldsymbol{w}) - \mathbb{E}[h(X, \boldsymbol{w})], v\rangle\}. \quad (15)\end{aligned}$$

By Assumption 3, the condition $\Delta_1' \le \sigma_2^2/\gamma_2$, and the concentration inequalities for sub-exponential random variables, for $v \in \mathcal{V}$ we have:

$$\mathbb{P}\{\langle \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla h(X_j, \boldsymbol{w}) - \mathbb{E}[h(X, \boldsymbol{w})], v\rangle \ge \Delta_1'(\boldsymbol{w} - \boldsymbol{w}^*)\} \le \exp(-\frac{|D_i|(\Delta_1')^2}{2\sigma_2^2}). \quad (16)$$

Recall that in $\mathcal{V}$ contains at most 6^d vetors. In view of the union bound, if further yields that

$$\begin{aligned}&\mathbb{P}\{2\sup_{v\in\mathcal{V}}\{\langle \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla h(X_j, \boldsymbol{w}) - \mathbb{E}[h(X, \boldsymbol{w})], v\rangle\} \\ &\ge 2\Delta_1'(\boldsymbol{w} - \boldsymbol{w}^*)\} \le 6^d \exp(-\frac{|D_i|(\Delta_1')^2}{2\sigma_2^2}) \\ &= \exp(-\frac{|D_i|(\Delta_1')^2}{2\sigma_2^2} + d\log 6). \quad (17)\end{aligned}$$

Therefore,

$$\mathbb{P}\{\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla h(X_j, \boldsymbol{w}) - \mathbb{E}[h(X, \boldsymbol{w})]\| \ge 2\Delta_1'(\boldsymbol{w} - \boldsymbol{w}^*)\} \le \exp(-\frac{|D_i|(\Delta_1')^2}{2\sigma_2^2} + d\log 6). \quad (18)$$

We conclude the proof by equation $\Delta_1' = \sqrt{2}\sigma_2\sqrt{(d\log 6 + \log(3/\delta))/|D_i|}$.

***Lemma** 5:* Given a real number $r > 0$, we let

$$\Delta_2(|D_i|) = \sigma_2\sqrt{\frac{2}{|D_i|}}\sqrt{K_1 + K_2 + K_3},$$

where $K_1 = d\log\frac{18M\vee M'}{\sigma_2}$, $K_2 = \frac{1}{2}d\log\frac{|D_i|}{d}$, $K_3 = \log(\frac{6\sigma_2^2 r\sqrt{|D_i|}}{\gamma_2\sigma_1\delta})$, and $|D_i|$ is size of the local dataset.

Suppose Assumption 2 - Assumption 5 hold, and $\boldsymbol{\omega} \subset \{\boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{w}^*\| \le r\sqrt{d}\}$ for some positive parameter r. For any $\delta \in (0,1)$ and any integer $|D_i|$, if $\Delta_1 \le \sigma_1^2/\gamma_1$ and $\Delta_2 \le \sigma_2^2/\gamma_2$, we have:

$$\begin{aligned}&\mathbb{P}\{\forall \boldsymbol{w} \in \boldsymbol{\omega} : \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \\ &\le 8\Delta_2\|\boldsymbol{w} - \boldsymbol{w}^*\| + 4\Delta_1\} \ge 1 - \delta.\end{aligned}$$

Proof: Our proof is mainly based on the ε-net argument [1], [3]. We let $\tau = \frac{\gamma_2\sigma_1}{2\sigma_2^2}\sqrt{\frac{d}{|D_i|}}$ and $\ell^* = \lceil r\sqrt{d}/\tau\rceil$. For any integer $1 \le \ell \le \ell^*$, we let $\boldsymbol{\omega}_l \triangleq \{\boldsymbol{w} : \|\boldsymbol{w} - \boldsymbol{w}^*\| \le r\sqrt{d}\}$. Given an integer ℓ, we let $\boldsymbol{w}_1, \dots, \boldsymbol{w}_{N_{\varepsilon_\ell}}$ be an ε-cover of $\boldsymbol{\omega}_\ell$, where $\varepsilon_\ell = (\sigma_2\tau\ell\sqrt{d/|D_i|})/(M \vee M')$, and $M \vee M' = \max\{M, M'\}$. We know $\log N_{\varepsilon_\ell} \le d\log\left(\frac{3\tau\ell}{\varepsilon_\ell}\right)$ from [2]. For any $\boldsymbol{w} \in \boldsymbol{\omega}$, there exists a k_ℓ ($1 \le k_\ell \le N_{\varepsilon_\ell}$) such that $\|\boldsymbol{w} - \boldsymbol{w}_{k_\ell}\| \le \varepsilon_\ell$. By triangle's inequality, we have:

$$\begin{aligned}\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}) - \nabla F(\boldsymbol{w})\| &\le \|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}_{k_\ell})\| \\ &+ \| \frac{1}{|D_i|} \sum_{X_j \in D_i} (\nabla f(X_j, \boldsymbol{w}) - \nabla f(X_j, \boldsymbol{w}_{k_\ell}))\| \\ &+ \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}_{k_\ell}) - \nabla F(\boldsymbol{w}_{k_\ell})\|. \quad (19)\end{aligned}$$

In view of Assumption 1, we have:

$$\|\nabla F(\boldsymbol{w}) - \nabla F(\boldsymbol{w}_{k_\ell})\| \le M\|\boldsymbol{w} - \boldsymbol{w}_{k_\ell}\| \le M\varepsilon_\ell. \quad (20)$$

We define event

$$\mathcal{E}_1 = \{ \sup_{\boldsymbol{w},\boldsymbol{w}' \in \boldsymbol{\omega}: \boldsymbol{w} \neq \boldsymbol{w}'} \frac{\|\nabla \bar{f}_{|D_i|}(\boldsymbol{w}) - \nabla \bar{f}_{|D_i|}(\boldsymbol{w}')\|}{\|\boldsymbol{w} - \boldsymbol{w}'\|} \leq M' \}, \tag{21}$$

where $\nabla \bar{f}_{|D_i|}(\boldsymbol{w}) = (\sum_{X_j \in D_i} \nabla f(X_i, \boldsymbol{w}))/|D_i|$.

By Assumption 4, we have $\mathbb{P}\{\mathcal{E}_1\} \geq 1 - \delta/3$. On event $\mathcal{E}_1$, we have the following:

$$\sup_{\boldsymbol{w},\boldsymbol{w}' \in \boldsymbol{\omega}: \boldsymbol{w} \neq \boldsymbol{w}'} \| \frac{1}{|D_i|} \sum_{X_i \in D_i} (f(X_i, \boldsymbol{w}) - f(X_i, \boldsymbol{w}_{k_\ell}))\| \leq M' \varepsilon_\ell. \tag{22}$$

By triangle's inequality, we have:

$$\begin{aligned}
&\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}_{k_\ell}) - \nabla F(\boldsymbol{w}_{k_\ell})\| \\
&\leq \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}^*) - \nabla F(\boldsymbol{w}^*)\| \\
&+ \| \frac{1}{|D_i|} \sum_{X_j \in D_i} (\nabla f(X_j, \boldsymbol{w}_{k_\ell}) - \nabla f(X_j, \boldsymbol{w}^*)) \\
&- (\nabla F(\boldsymbol{w}_{k_\ell}) - \nabla F(\boldsymbol{w}^*))\| \\
&\overset{(a)}{\leq} \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}^*) - \nabla F(\boldsymbol{w}^*)\| \\
&+ \| \frac{1}{|D_i|} \sum_{X_j \in D_i} h(X_j, \boldsymbol{w}_{k_\ell}) - \mathbb{E}[h(X, \boldsymbol{w}_{k_\ell})]\|,
\end{aligned} \tag{23}$$

where (a) is because Equations 1 and 2.

We define events as:

$$\mathcal{E}_2 = \{ \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_i, \boldsymbol{w}^*) - \nabla F(\boldsymbol{w}^*)\| \leq 2\Delta_1 \}, \tag{24}$$

$$\begin{aligned}
\mathcal{F}_\ell = \{ \sup_{1 \leq k \leq N_\epsilon} \| \frac{1}{|D_i|} \sum_{X_j \in D_i} h(X_j, \boldsymbol{w}_k) - \mathbb{E}[h(X, \boldsymbol{w}_k)]\| \\
\leq 2\tau\ell\Delta_2 \}.
\end{aligned} \tag{25}$$

Since $\Delta_1 \leq \sigma_1^2/\gamma_1$, it follows from Lemma 3 that $\mathbb{P}\{\mathcal{E}_2\} \geq 1 - \delta/3$. For $\Delta_2 \leq \sigma_2^2/\gamma_2$ from Lemma 4, we have:

$$\begin{aligned}
&\mathbb{P}\{\mathcal{F}_\ell^c\} = \mathbb{P}\{ \sup_{1 \leq k \leq N_{\epsilon_\ell}} \| \frac{1}{|D_i|} \sum_{X_j \in D_i} h(X_j, \boldsymbol{w}_k) \\
&- \mathbb{E}[h(X, \boldsymbol{w}_k)]\| > 2\tau\ell\Delta_2 \} \\
&\leq \sum_{k=1}^{N_{\epsilon_\ell}} \mathbb{P}\{ \| \frac{1}{|D_i|} \sum_{X_j \in D_i} h(X_j, \boldsymbol{w}_k) - \mathbb{E}[h(X, \boldsymbol{w}_k)]\| > 2\tau\ell\Delta_2 \} \\
&\leq \frac{\delta}{3\ell^*} \frac{1}{(\frac{3\tau\ell}{\epsilon_\ell})^d} (\frac{3\tau\ell}{\epsilon_\ell})^d = \frac{\delta}{3\ell^*}.
\end{aligned} \tag{26}$$

In conclusion, by combining Equations 19, 20, 22, and 23, on event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{F}_\ell$, we have:

$$\begin{aligned}
&\sup_{\boldsymbol{w} \in \boldsymbol{\omega}_\ell} \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \\
&\leq (M + M')\epsilon_\ell + 2\Delta_1 + 2\Delta_2\tau\ell \\
&\overset{(a)}{\leq} 4\Delta_2\tau\ell + 2\Delta_1,
\end{aligned} \tag{27}$$

where (a) is due to $(M \vee M')\epsilon_\ell \leq \Delta_2\tau\ell$. We let event $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap (\cap_{\ell=1}^{\ell^*} \mathcal{F}_\ell)$. By the union bound, we have $\mathbb{P}\{\mathcal{E}\} \geq 1 - \delta$. Moreover, suppose event $\mathcal{E}$ holds. For any $\boldsymbol{w} \in \boldsymbol{\omega}_\ell$, there exists an $1 \leq \ell \leq \ell^*$ such that $(\ell - 1)\tau < \|\boldsymbol{w} - \boldsymbol{w}^*\| \leq \ell\tau$. If $2 \leq \ell \leq 2(\ell - 1)$, we have:

$$\begin{aligned}
\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}) - \nabla F(\boldsymbol{w})\| &\leq 4\Delta_2\tau\ell + 2\Delta_1 \\
&\leq 8\Delta_2 \|\boldsymbol{w} - \boldsymbol{w}^*\| + 2\Delta_1.
\end{aligned} \tag{28}$$

If $\ell = 1$, we have:

$$\| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \leq 4\Delta_2\tau\ell + 2\Delta_1 \overset{(a)}{\leq} 4\Delta_1, \tag{29}$$

where (a) is due to that $\Delta_2 \leq \sigma_2^2/\gamma_2$ and $\Delta_1 \geq \sigma_1 \sqrt{d/|D_i|}$. Combining inequalities 5 and 29, we have:

$$\begin{aligned}
&\sup_{\boldsymbol{w} \in \boldsymbol{\omega}_{\ell^*}} \| \frac{1}{|D_i|} \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \\
&\leq 8\Delta_2 \|\boldsymbol{w} - \boldsymbol{w}^*\| + 4\Delta_1.
\end{aligned} \tag{30}$$

The proposition follows by the Assumption that $\boldsymbol{\omega} \subset \boldsymbol{\omega}_{\ell^*}$. In addition, we let $\boldsymbol{g}_{median} = 1/(|D_i|) \sum_{X_j \in D_i} \nabla f(X_j, \boldsymbol{w})$, that is because the median update selected by the server at each epoch is a local update uploaded by a certain client.

Proof of Theorem 1: With the help of the above lemma, we can prove the theory that the difference between the aggregated result of the global model at epoch t and the optimal solution is bounded. We have:

$$
\begin{aligned}
&\|\boldsymbol{w}^t-\boldsymbol{w}^*\| \\
&=\|\boldsymbol{w}^{t-1}-\alpha\boldsymbol{g}^{t-1}-\boldsymbol{w}^*\| \\
&=\|\boldsymbol{w}^{t-1}-\alpha\nabla F(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*+\alpha\nabla F(\boldsymbol{w}^{t-1})-\alpha\boldsymbol{g}^{t-1}\| \\
&\leq\|\boldsymbol{w}^{t-1}-\alpha\nabla F(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\|+\alpha\|\nabla F(\boldsymbol{w}^{t-1})-\boldsymbol{g}^{t-1}\| \\
&\overset{(a)}{=}\|\boldsymbol{w}^{t-1}-\alpha\nabla F(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\|+\alpha\|\boldsymbol{g}^{t-1}-\nabla F(\boldsymbol{w}^{t-1})\| \\
&\overset{(b)}{\leq}\|\boldsymbol{w}^{t-1}-\alpha\nabla F(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\|+2\alpha\|\nabla F(\boldsymbol{w}^{t-1})\| \\
&+3\alpha\|\boldsymbol{g}^{t-1}_{median}-\nabla F(\boldsymbol{w}^{t-1})\| \\
&\overset{(c)}{=}\underbrace{\|\boldsymbol{w}^{t-1}-\alpha\nabla F(\boldsymbol{w}^{t-1})-\boldsymbol{w}^*\|}_{A_1}+2\alpha\underbrace{\|\nabla F(\boldsymbol{w}^{t-1})-\nabla F(\boldsymbol{w}^*)\|}_{A_2} \\
&+3\alpha\underbrace{\|\boldsymbol{g}^{t-1}_{median}-\nabla F(\boldsymbol{w}^{t-1})\|}_{A_3} \\
&\overset{(d)}{\leq}\sqrt{1-L^2/(4M^2)}\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|+2\alpha M\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\| \\
&+3\alpha(8\Delta_2\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|+4\Delta_1) \\
&=(\sqrt{1-L^2/(4M^2)}+24\alpha\Delta_2+2\alpha M)\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|+12\alpha\Delta_1,
\end{aligned}
\tag{31}
$$

where (a) has the same reason as Lemma $1(a)$; (b) is obtained according to Lemma 1, (c) is due to $\nabla F(\boldsymbol{w}^*)=0$; and A_1, A_2, and A_3 in (c) are Lemma 2, Assumption 1, and Lemma 5, respectively.

By letting $\rho=1-(\sqrt{1-L^2/(4M^2)}+24\alpha\Delta_2+2\alpha M)$, we have:

$$
\begin{aligned}
&\|\boldsymbol{w}^t-\boldsymbol{w}^*\| \\
&\leq(1-\rho)\|\boldsymbol{w}^{t-1}-\boldsymbol{w}^*\|+12\alpha\Delta_1 \\
&\leq(1-\rho)[(1-\rho)\|\boldsymbol{w}^{t-2}-\boldsymbol{w}^*\|+12\alpha\Delta_1]+12\alpha\Delta_1 \\
&=(1-\rho)^2\|\boldsymbol{w}^{t-2}-\boldsymbol{w}^*\|+12[1+(1-\rho)]\alpha\Delta_1 \\
&\leq(1-\rho)^2[(1-\rho)\|\boldsymbol{w}^{t-3}-\boldsymbol{w}^*\|+12\alpha\Delta_1]+12[1+(1-\rho)]\alpha\Delta_1 \\
&=(1-\rho)^3\|\boldsymbol{w}^{t-3}-\boldsymbol{w}^*\|+12[1+(1-\rho)+(1-\rho)^2]\alpha\Delta_1 \\
&\ldots \\
&\leq 12[1+(1-\rho)+(1-\rho)^2+\cdots+(1-\rho)^{t-1}]\alpha\Delta_1 \\
&+(1-\rho)^t\|\boldsymbol{w}^0-\boldsymbol{w}^*\| \\
&=(1-\rho)^t\|\boldsymbol{w}^0-\boldsymbol{w}^*\|+12\frac{1-(1-\rho)^t}{1-(1-\rho)}\alpha\Delta_1 \\
&=(1-\rho)^t\|\boldsymbol{w}^0-\boldsymbol{w}^*\|+\frac{12\alpha\Delta_1}{\rho}-\frac{12(1-\rho)^t\alpha\Delta_1}{\rho} \\
&\leq(1-\rho)^t\|\boldsymbol{w}^0-\boldsymbol{w}^*\|+\frac{12\alpha\Delta_1}{\rho}.
\end{aligned}
\tag{32}
$$

Thus, we conclude the proof.

REFERENCES

[1] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proc. ACM Meas. Anal. Comput. Syst.*, 2017.

[2] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *arXiv preprint arXiv:1011.3027*, 2010.

[3] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," in *ISOC Network and Distributed System Security Symposium*, 2021.