# Supplementary Material for "One-Bit Aggregation for Over-the-Air Federated Learning Against Byzantine Attacks"

Yifan Miao, Wanli Ni, and Hui Tian, *Senior Member, IEEE*

## Appendix A
## Proof of Theorem 1

To begin with, we extend (6) in Assumption 2 by substituting $\mathbf{w}^{t+1}=\mathbf{w}^t-\eta\mathbf{v}^t$ to expose the bit error rate due to various factors, which is presented as Lemma 1.

***Lemma 1:*** Assuming that the model parameter vector is $\mathbf{w}^t$ at the $t$-th communication round, then we have

$$\mathbb{E}\left[F^{t+1}-F^t|\mathbf{w}^t\right]\leq-\eta\left\|\mathbf{g}^t\right\|_1+\frac{\eta^2}{2}\|\mathbf{L}\|_1 + 2\eta\sum\nolimits_{i=1}^{Q}\left|g_i^t\right|\mathbb{P}\left[\operatorname{sign}(\bar{g}_i^t)\neq\operatorname{sign}(g_i^t)\right], \quad (12)$$

where $\mathbb{P}\left[\operatorname{sign}(\bar{g}_i^t)\neq\operatorname{sign}(g_i^t)\right]$ represents the error probability of each element of the gradient vector after demodulation compared with the real situation, which can reflect the strength of Byzantine attacks and the stochasticity of wireless channels.

*Proof:* See [1] for details due to space limitation. ■

Set $P_i^{\mathrm{err}}=\mathbb{P}\left[\operatorname{sign}(\bar{g}_i^t)\neq\operatorname{sign}(g_i^t)\right]$ and then we focus on deriving the theoretical upper bound of $P_i^{\mathrm{err}}$. If there is no Byzantine attacker among all devices, let $X_i$ represent the number of EDs whose sign of the $i$-th element of the local gradient vector is received incorrectly. It can be seen as the sum of $K$ independent Bernoulli trails and follows a binomial distribution (BD). The failure and success probabilities of each trial are given by $p_i=\mathbb{P}[\operatorname{sign}(\bar{g}_{k,i}^t)\neq\operatorname{sign}(g_i^t)]$ and $q_i=\mathbb{P}[\operatorname{sign}(\bar{g}_{k,i}^t)=\operatorname{sign}(g_i^t)]$, respectively. Then we have $\mathbb{E}[X_i]=Kp_i=K(\frac{1}{2}-\varepsilon_i)$ and $\operatorname{Var}(X_i)=Kp_iq_i=K(\frac{1}{4}-\varepsilon_i^2)$ with $\varepsilon_i=\frac{1}{2}-p_i=q_i-\frac{1}{2}$.

Byzantine attackers actively send opposite gradients to interfere with global gradient aggregation, thus the probability of the BS receiving correct and incorrect signs from them is opposite to normal devices. Let $\tilde{X}_i$ be the number of EDs with the wrong sign for the $i$-th gradient element under Byzantine attacks. As the sum of two BDs, $\tilde{X}_i$ also follows a BD with

$$\mathbb{E}[\tilde{X}_i]=Vp_i+Uq_i=\frac{1}{2}(U+V)+\varepsilon_i(U-V), \quad (13)$$

$$\operatorname{Var}(\tilde{X}_i)=Vp_iq_i+Uq_ip_i=(U+V)(\frac{1}{4}-\varepsilon_i^2). \quad (14)$$

Due to the majority vote adopted, $\tilde{X}_i$ must be greater than $\frac{K}{2}$ for $\operatorname{sign}(\bar{g}_i^t)\neq\operatorname{sign}(g_i^t)$, hence we have

$$P_i^{\mathrm{err}}=\mathbb{P}(\tilde{X}_i\geq\frac{K}{2})=\mathbb{P}[\tilde{X}_i-\mathbb{E}[\tilde{X}_i]\geq\frac{K}{2}-\mathbb{E}[\tilde{X}_i]] \leq\frac{1}{1+\frac{\left(\frac{K}{2}-\mathbb{E}[\tilde{X}_i]\right)^2}{\operatorname{Var}(\tilde{X}_i)}}\leq\frac{1}{2}\sqrt{\frac{V+U}{(V-U)^2}\left(\frac{1}{4\varepsilon_i^2}-1\right)}, \quad (15)$$

where the two scalings utilize the Cantellis' inequality and the AM-GM inequality, respectively. Given the variance bound and the unimodal symmetric gradient noise postulated in Assumptions 3 and 4, the Gauss' inequality is adopted to obtain $\frac{1}{4\varepsilon_i^2}-1\leq\frac{4}{S_i^2}$ by monotonicity discussion, where $S_i=\sqrt{n_b}|g_i^t|/\sigma_i$. Then, (15) can be written as $P_i^{\mathrm{err}}\leq\frac{\sqrt{V+U}}{S_i(V-U)}$. Substituting $\eta=\frac{1}{\sqrt{\|\mathbf{L}\|_1 n_b}}$ and $n_b=\frac{1}{\gamma}T$ into (12), we have

$$\mathbb{E}\left[F^{t+1}-F^t|\mathbf{w}^t\right]\leq-\eta\left\|\mathbf{g}^t\right\|_1+\frac{\eta^2}{2}\|\mathbf{L}\|_1+\frac{2\eta\sqrt{V+U}}{\sqrt{n_b}(V-U)}\|\boldsymbol{\sigma}\|_1 =-\frac{\sqrt{\gamma}}{\sqrt{\|\mathbf{L}\|_1 T}}\left\|\mathbf{g}^t\right\|_1+\frac{\gamma}{2T}+\frac{2\gamma\sqrt{V+U}}{\sqrt{\|\mathbf{L}\|_1}T(V-U)}\|\boldsymbol{\sigma}\|_1. \quad (16)$$

Now we extend the expectation of randomness in the optimization trajectory and perform the telescoping sum over the $T$ iterations:

$$F^0-F^*\geq F^0-\mathbb{E}\left[F^t\right]=\mathbb{E}[\sum\nolimits_{t=0}^{T-1}(F^t-F^{t+1})] =\sqrt{\frac{\gamma T}{\|\mathbf{L}\|_1}}\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\|\mathbf{g}^t\right\|_1\right]-\frac{2\gamma\sqrt{V+U}}{\sqrt{\|\mathbf{L}\|_1}(V-U)}\|\boldsymbol{\sigma}\|_1-\frac{\gamma}{2}. \quad (17)$$

Finally, we shift and reorganize (17) to reach the conclusion in Theorem 1, which completes the proof.

## Appendix B
## Proof of Theorem 2

Similar to noise-free channels, the primary objective remains to find the upper bound of $P_i^{\mathrm{err}}$, in which the analysis for $P_{k,i}^{\mathrm{err}}=\mathbb{P}[\operatorname{sign}(\bar{g}_{k,i}^t)\neq\operatorname{sign}(g_i^t)]$ is a vital step. Under Rayleigh fading channels, $P_{k,i}^{\mathrm{err}}$ can be divided into $\mathbb{P}[\operatorname{sign}(\tilde{g}_{k,i}^t)\neq\operatorname{sign}(g_i^t)]$ and $\mathbb{P}[\operatorname{sign}(\bar{g}_{k,i}^t)\neq\operatorname{sign}(\tilde{g}_{k,i}^t)]$. $p_i=\mathbb{P}[\operatorname{sign}(\tilde{g}_{k,i}^t)\neq\operatorname{sign}(g_i^t)]$ represents the probability that the sign of stochastic gradient $\tilde{g}_{k,i}^t$ does not match the true gradient $g_i^t$ under the error-free transmission, which has been detailed discussed in Appendix A. $P_k^{\mathrm{com}}=\mathbb{P}[\operatorname{sign}(\bar{g}_{k,i}^t)\neq\operatorname{sign}(\tilde{g}_{k,i}^t)]$ is the probability that the sign of demodulated gradient $\bar{g}_{k,i}^t$ is inconsistent with the transmitted one $\tilde{g}_{k,i}^t$ due to imperfect

Y. Miao, W. Ni and H. Tian are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: miaoyifan@bupt.edu.cn; charleswall@bupt.edu.cn; tianhui@bupt.edu.cn;).

wireless channels, which has been given by (5). Accordingly, for Byzantine attackers and normal devices, we have

$$P_{u,i}^{\mathrm{err}} = P_u^{\mathrm{com}} p_i + (1 - P_u^{\mathrm{com}})(1 - p_i),\ \forall u \in \mathcal{U}, \tag{18}$$

$$P_{v,i}^{\mathrm{err}} = P_v^{\mathrm{com}} (1 - p_i) + (1 - P_v^{\mathrm{com}}) p_i,\ \forall v \in \mathcal{V}. \tag{19}$$

Let $Y_i$ represent the number of EDs whose sign of the $i$-th local gradient element is incorrectly received. Then, it can be viewed as the sum of $K$ Bernoulli trails with varying success probablity and obeys the Poisson binomial distribution (PBD). BD is a special case in PBD, which can provide some interesting inequalities for subsequent derivation.

***Lemma** 2:* If the random variable $Z$ follows a PBD with $Z \sim \mathrm{PB}(P_1, P_2, ..., P_N)$ and $\hat{Z}$ follows a BD with $\hat{Z} \sim \mathrm{B}(N, \hat{P})$, we have $\mathbb{P}[Z \leq \lambda] \geq \mathbb{P}[\hat{Z} \leq \lambda]$ for $N\hat{P} \leq \lambda \leq N$.

*Proof:* The approach of Schur convexity and optimization is applied to complete the proof, as detailed in [2]. ■

According to Lemma 2, we can obtain $\mathbb{P}[Y_i \geq \frac{U+V}{2}] \leq \mathbb{P}[\hat{Y}_i \geq \frac{U+V}{2}]$, where $\hat{Y}_i$ follows a BD w.r.t. $Y_i$ and $\hat{P} = \frac{1}{U+V}(\sum_{u\in\mathcal{U}} P_{u,i}^{\mathrm{err}} + \sum_{v\in\mathcal{V}} P_{v,i}^{\mathrm{err}})$. Therefore, substituting $\varepsilon_i = \frac{1}{2} - \hat{P}$ and (18) (19), we can obtain

$$\varepsilon_i = \frac{\frac{1}{2} - p_i}{U+V}\Big(\sum\nolimits_{v\in\mathcal{V}}(1-2P_v^{\mathrm{com}}) - \sum\nolimits_{u\in\mathcal{U}}(1-2P_u^{\mathrm{com}})\Big). \tag{20}$$

From $\frac{1}{4(\frac{1}{2}-p_i)^2} - 1 \leq \frac{4}{S_i^2}$ in Appendix A, we have $(\frac{1}{2} - p_i)^2 \geq \frac{S_i^2}{16+4S_i^2}$. Then, with reference to (15), we have

$$P_i^{\mathrm{err}} \leq \frac{(V+U)^{\frac{3}{2}}}{S_i(V-U)\Delta} + \frac{V+U}{2(V-U)}\sqrt{\frac{U+V}{\Delta^2} - \frac{1}{U+V}}, \tag{21}$$

where $\Delta = \sum_{v\in\mathcal{V}}(1 - 2P_v^{\mathrm{com}}) - \sum_{u\in\mathcal{U}}(1 - 2P_u^{\mathrm{com}})$ and the contraction is based on the fact that $\sqrt{a} + \sqrt{b} \geq \sqrt{a+b}$. Substituting the above result into (12) and performing the same procedure as (16) and (17), the conclusion in Theorem 2 can easily be drawn, thus completing the proof.

## REFERENCES


[1] H. Li, R. Wang, W. Zhang, and J. Wu, "One bit aggregation for federated edge learning with reconfigurable intelligent surface: Analysis and optimization," *IEEE Trans. Wireless Commun.*, Feb. 2023.

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2022.