

# TESSERACT: GRADIENT FLIP SCORE TO SECURE FEDERATED LEARNING AGAINST MODEL POISONING ATTACKS

**Anonymous authors**

Paper under double-blind review

## A APPENDIX

### A.1 PENALTY AND REWARD VALUE

**Penalty and reward selection.** Our design policy penalizes  $2c_{max}$  out of  $m$  clients in every iteration. Considering a completely benign scenario, we want the expected value of the reputation score of a client that has been penalized  $e$  fraction of times to be zero, where  $e = \frac{2c_{max}}{m}$ . Let a client  $i$  be penalized  $en$  number of times in  $n$  iterations. There are  $\binom{n}{en}$  ways to select the iterations where the client is penalized. After  $n$  iterations, the reputation score of client  $i$  is given by:

$$RS(i, n) = \sum_{t=0}^n \mu_d^{n-t} \mathcal{W}(i, t). \quad (1)$$

where  $\mathcal{W}(i, t)$  is a sequence of penalty and reward over time. The expected value of this reputation score over all possible sequences  $j \in \binom{n}{en}$  is

$$\begin{aligned} \mathbb{E}_j[RS(i, n)] &= \frac{1}{\binom{n}{en}} \sum_j RS(i, n) \\ &= \frac{1}{\binom{n}{en}} \sum_j \sum_t \mu_d^{n-t} \mathcal{W}(i, t) \\ &= \frac{1}{\binom{n}{en}} \sum_t \mu_d^{n-t} \sum_j \mathcal{W}(i, t) \\ &= \frac{1}{\binom{n}{en}} \sum_t \mu_d^{n-t} \left( -pen \binom{n}{en} + (r(1-e)n \binom{n}{en}) \right) \end{aligned}$$

Our setting with  $r = \frac{2c_{max}}{m}$  and  $p = 1 - r$  makes the above quantity to be zero thus ensuring that its expected reputation score increment is zero. This proof assumes that it is a random process through which (benign) clients generate their flip scores. Thus, if a subset of clients are penalized less than  $\frac{2c_{max}}{m}$  of times, they are expected to have a net neutral reputation score.

**Upper and lower bound of reputation score.** From the above expression, it is obvious that if  $\mu_d = 0$ ,  $-p \leq RS \leq r$ . When  $0 < \mu_d < 1$ , the upper and lower bounds can be computed by assuming that a client was rewarded or penalized respectively in every iteration. Assuming that the number of iterations tends towards infinity, equation (1) forms an infinite geometric sequence, that can be solved to obtain  $\frac{-p}{1-\mu_d} \leq RS \leq \frac{r}{1-\mu_d}$ . It should be noted that these reputation scores are normalized using softmax to compute the reputation weights. If the absolute value of the lower bound is not large enough (if  $\mu_d$  is set to be too small), then even after perfect detection, a malicious client can still have a significant reputation weight after softmax normalization. If  $\mu_d$  is set to a value closer to 1, then the absolute value of the lower and upper bounds increase, bringing down the contribution of malicious clients to almost zero. At the same time, redemption becomes difficult for a client in this case. This trade-off needs to be kept in mind when setting the decay parameter. We have used  $\mu_d = 0.99$  in our experiments in order to remain conservative and make recovery difficult for a client that has been penalized a lot of times. However, this is a design parameter that the user can decide.

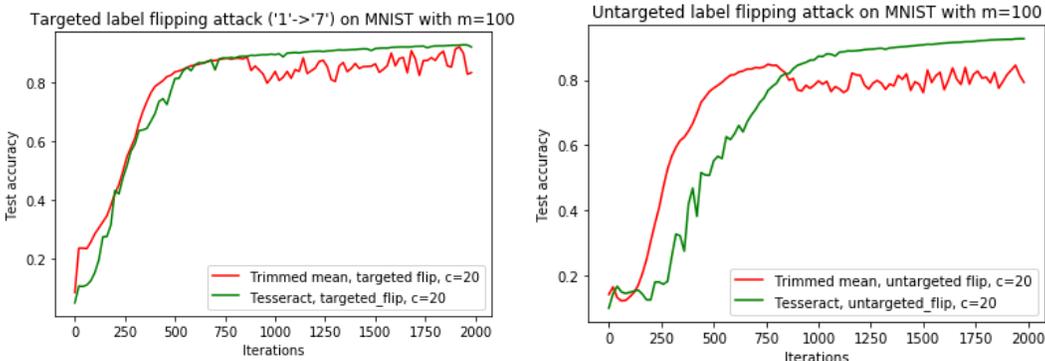


Figure 1: TESSERACT’s Performance against the targeted and untargeted label flipping attacks on the MNIST dataset. We observe that the attacks have some damage on the model, but Tesseract is able to remedy this for both attacks.

## A.2 LABEL FLIPPING

The attack that we target, “the directed-deviation attack” has been shown to be the most powerful attack in federated learning Fang et al. (2020), and specifically claims to be more effective than state-of-the-art untargeted data poisoning attacks for multi-class classifiers, that is, label flipping attack, Gaussian attack, and back-gradient optimization based attacks Muñoz-González et al. (2017). They show that the existing data poisoning attacks are insufficient and cannot produce a high testing error rate, not higher than 0.2 in the presence of byzantine-robust aggregation techniques (Krum, trimmed mean, and median).

We observe that both state-of-the-art targeted and untargeted label flipping attacks are not powerful enough on the CIFAR-10 and FEMNIST datasets and have negligible damage. The attacks do have some damaging impact on the MNIST dataset, but when TESSERACT is used, the damage is completely mitigated. Thus, we verify the claims from Fang et al. (2020) and show that TESSERACT’s intuition is general enough to counteract both the more powerful “directed-deviation attacks” and the weaker state-of-the-art data poisoning attacks.

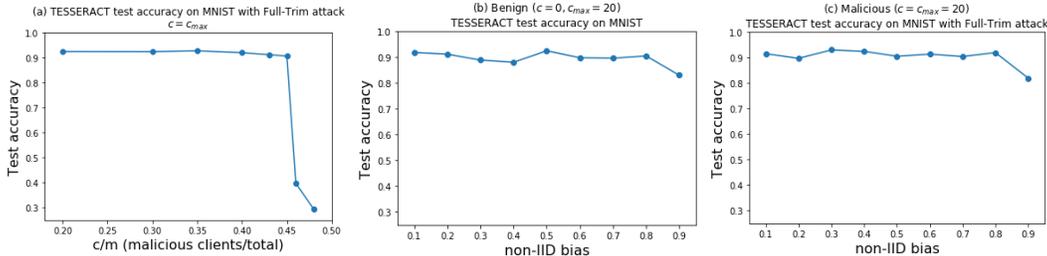
## A.3 ADDITIONAL EXPERIMENTS

Here, we provide additional evaluation of TESSERACT in two specific situations. We stress-test it first by subjecting it to a higher number of malicious clients to find the breaking point of TESSERACT, when trained on MNIST dataset in the presence of Full-trim attack. We assume that the number of compromised clients is still not greater than  $c_{max}$ , and to that end, we set  $c_{max} = c$ . Since TESSERACT requires  $c_{max} < \frac{m}{2}$ , we have swept  $c$  upto 49 where  $m$  was fixed at 100. We observe in Figure 2(a) that TESSERACT is stable upto  $c/m = 0.45$  whereas the rest of the defense techniques broke below  $c/m = 0.30$  as can be seen in Table ?? and Figure ?? with  $c_{max} = c$  set for all the defense techniques that require a knowledge of  $c_{max}$ .

Figures 2(b) and (c) show the performance of TESSERACT on MNIST dataset distributed among 100 clients with varying degrees of non-IIDness. We observe that, except for the extreme case of  $bias = 0.9$ , TESSERACT remains exceptionally stable.

Here, we describe the mathematical formulation of the adaptive-attacks. Full-Krum attack finds a vector of gradients  $u$  by solving an optimization problem described in Fang et al. (2020), and every malicious client  $i$  would send  $u$  with an additional noise to appear different. Full-Trim attack solves a different optimization problem to also come up with a vector of gradients  $u$  to which every malicious client  $i$  would add some noise to obtain  $u_i$ . The problem statement in our (two) derived versions of the above (two) attacks, namely Adaptive-Trim and Adaptive-Krum, is To find a set of vectors

$$v_i, i = 0, 1, 2, \dots, c - 1$$



**Figure 2:** Figure (a) shows the performance of TESSERACT on MNIST with increasing  $c$ . We see that TESSERACT is stable across a large range and breaks only above  $c = 0.45$  which is close to the theoretical limit of  $c = 0.5^-$ . Figure (b) and (c) show the test accuracy of TESSERACT on MNIST dataset distributed with varying non-IID bias across 100 clients in benign and malicious cases respectively. TESSERACT can be seen to be robust enough for a wide range of bias, that is 0.1 to 0.8, with a small dip in test accuracy occurring at bias = 0.9.

where  $c$  is the number of malicious clients. Here,  $v_i$  is the vector of gradients with size equal to the number of model parameters, each satisfying the constraint -

$$FS(v_i) < FS_{low}(t)$$

that is, every computed vector should have a flip-score lower than the cut-off flip-score according to the adversary’s knowledge, such that

$$\sum_i^{c-1} v_i = \sum_{i=0}^{c-1} u_i$$

where  $u_i$  were determined by the adversary originally as a valid solution to the Full-Krum and Full-Trim optimization problems. We solve this problem as described in Section 6.2. In short, we initialize  $v_i$  to some target value, and then undo the attack on “less important parameters” until the flip-score constraint is just met, and send the computed  $v_i$  for aggregation.  $v_0$  is initialized to  $u_0$ , and then updated until the flip-score constraint is satisfied. The difference  $u_0 - v_0$  is added to  $u_1$  which now becomes the initial value of  $v_1$  and so on. The results have been described and analyzed in Section 6.2.

We formulate an even stronger attack where the adversary also has the knowledge of its own reputation score ( $W_R$ ) in order to come up with attacked gradients with better chances of success. We call this a “Weighted-Adaptive-Trim” attack. The modified constraint is

$$\sum_i^{c-1} W_{R,i} v_i = \sum_{i=0}^{c-1} W_{R,i} u_i$$

TESSERACT successfully defends against this attack when evaluated on MNIST, as can be seen from the following results in Figure 3.

#### A.4 CONVERGENCE ANALYSIS

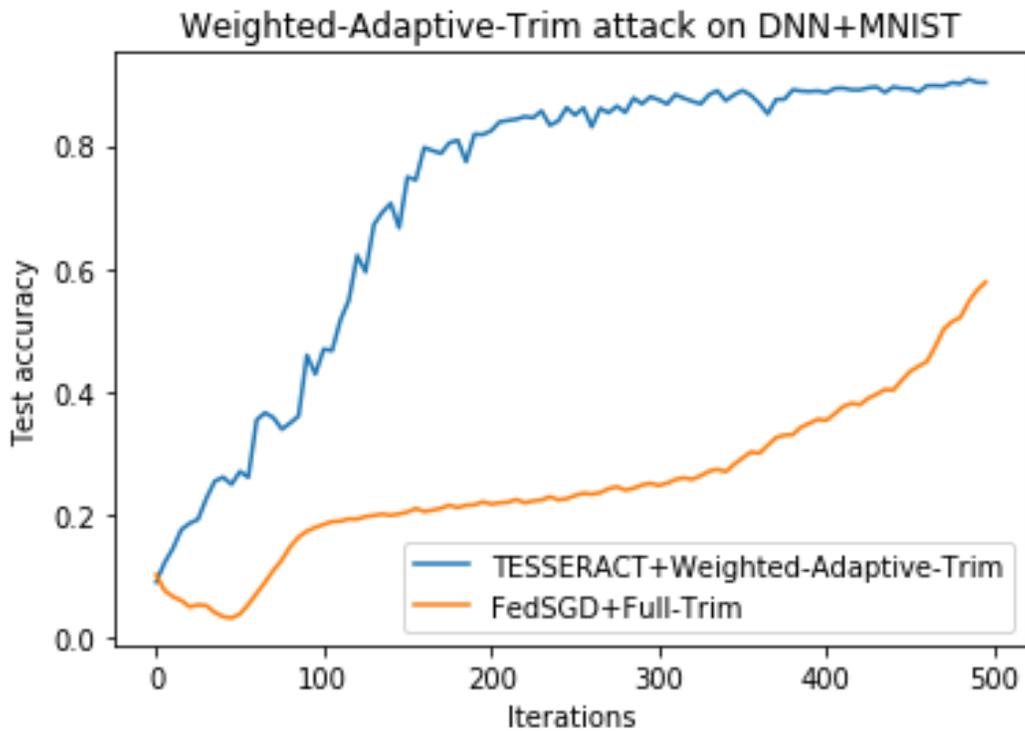
Let the  $k$ -th client hold  $n_k$  training data batches:  $x_{k,1}, \dots, x_{k,n_k}$ . The local objective function  $LM_k(\cdot)$  is given by

$$LM_k(\mathbf{w}) = \frac{1}{n_k} \sum_{j=1}^{n_k} l(\mathbf{w}; x_{k,j}),$$

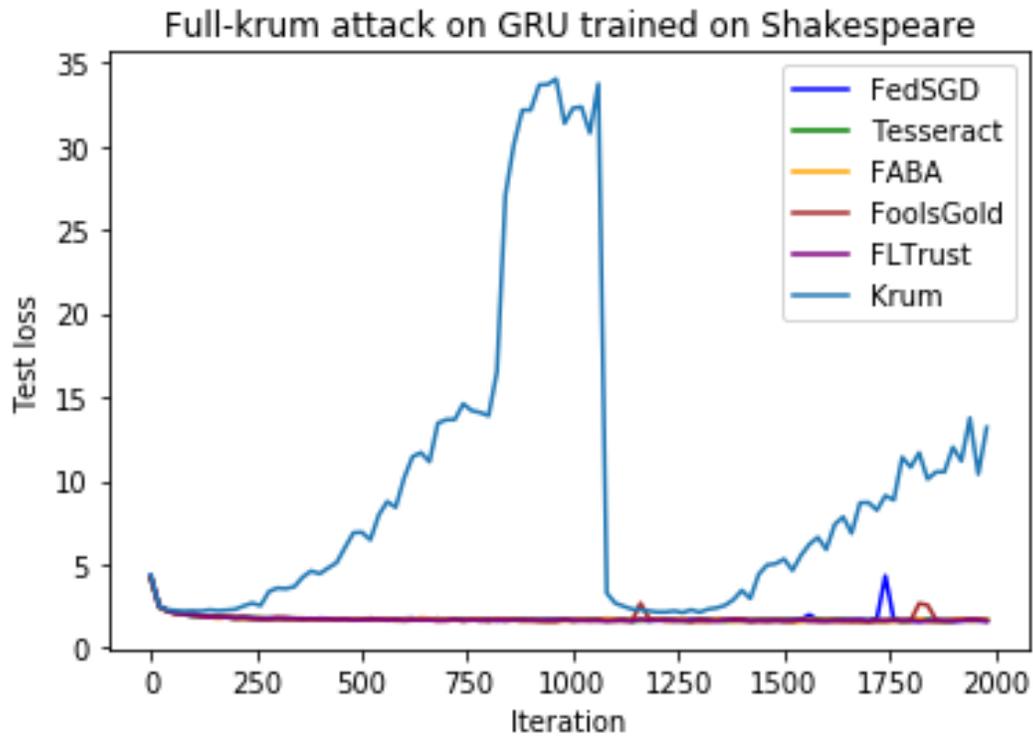
where  $l(\cdot; \cdot)$  is the specified loss function for each client.

The global objective function is defined as

$$GM_{k,t}(\mathbf{w}) = \sum_{k=1}^m p_{k,t} LM_k(\mathbf{w}).$$



*Figure 3: The figure shows the test accuracy of TESSERACT when evaluated on MNIST dataset under the default conditions with  $m = 100$ ,  $c = 20$  where the adversary launches the Weighted-Adaptive-Trim attack on the system, compared with the baseline performance of FedSGD against the Full-Trim attack. TESSERACT successfully defends against this attack to achieve a 90% accuracy.*



*Figure 4: The figure shows the impact of the Full-Krum attack on Krum aggregation as compared to the other aggregation techniques. It has a devastating impact of Krum for which the attack was specifically tailored for.*

The global model is updated as

$$\mathbf{w}_{t+1}^k = \mathbf{w}_t^k - \eta_t \sum_{k=1}^m p_{k,t} \nabla LM_k(\mathbf{w}_t^k),$$

where  $p_{k,t} = \text{softmax}(RS_{k,t})$  is the softmax of reputation score of client  $k$  at time  $t$ .

We update the weights by averaging the weights from selected clients  $\bar{\mathbf{w}}_t = \sum_{k=1}^m p_{k,t} \mathbf{w}_t^k$ . For convenience, we also define  $g_t = \sum_{k=1}^m p_{k,t} \nabla LM_k(\mathbf{w}_t^k, \xi_t^k)$ , where  $\xi_t^k$  is the selected local data.

#### A.4.1 ANALYSIS ON CONSECUTIVE STEPS

To bound the expectation of the global objective function at time  $T$  from its optimal value, we first consider to analyze the global weight from the optimal weights by calculating single step SGD:

$$\begin{aligned} \|\bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \|\bar{\mathbf{w}}_t - \eta_t g_t - \mathbf{w}^* - \eta_t \bar{g}_t + \eta_t \bar{g}_t\|^2 \\ &= \|\bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t\|^2 + 2\eta_t \langle \bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t, \bar{g}_t - g_t \rangle + \eta_t^2 \|\bar{g}_t - g_t\|^2. \end{aligned} \quad (2)$$

The first term of Equation. 2 can be expressed as

$$\|\bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t\|^2 = \|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 - 2\eta_t \langle \bar{\mathbf{w}}_t - \mathbf{w}^*, \bar{g}_t \rangle + \eta_t^2 \|\bar{g}_t\|^2. \quad (3)$$

The second term of Equation. 3 can be expressed as

$$\begin{aligned} -2\eta_t \langle \bar{\mathbf{w}}_t - \mathbf{w}^*, \bar{g}_t \rangle &= -2\eta_t \sum_{k=1}^m p_{k,t} \langle \bar{\mathbf{w}}_t - \mathbf{w}^*, \nabla LM_k(\mathbf{w}_t^k) \rangle \\ &= -2\eta_t \sum_{k=1}^m p_{k,t} \langle \bar{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla LM_k(\mathbf{w}_t^k) \rangle \\ &\quad - 2\eta_t \sum_{k=1}^m p_{k,t} \langle \mathbf{w}_t^k - \mathbf{w}^*, \nabla LM_k(\mathbf{w}_t^k) \rangle. \end{aligned} \quad (4)$$

By Cauchy-Schwarz inequality and AM-GM inequality, we have

$$-2\eta_t \sum_{k=1}^m p_{k,t} \langle \bar{\mathbf{w}}_t - \mathbf{w}_t^k, \nabla LM_k(\mathbf{w}_t^k) \rangle \leq \frac{1}{\eta_t} \|\bar{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \eta_t \|\nabla LM_k(\mathbf{w}_t^k)\|^2. \quad (5)$$

By the  $\mu$ -strong convexity of  $LM_k(\cdot)$ , we have

$$-2\eta_t \sum_{k=1}^m p_{k,t} \langle \mathbf{w}_t^k - \mathbf{w}^*, \nabla LM_k(\mathbf{w}_t^k) \rangle \leq -(LM_k(\mathbf{w}_t^k) - LM_k(\mathbf{w}^*)) - \frac{\mu}{2} \|\mathbf{w}_t^k - \mathbf{w}^*\|^2. \quad (6)$$

By the convexity of  $\|\cdot\|$  and the L-smoothness of  $LM_k(\cdot)$ , we can express third term of Equation. 3 as

$$\begin{aligned} \eta_t^2 \|\bar{g}_t\|^2 &\leq \eta_t^2 \sum_{k=1}^m p_{k,t} \|\nabla LM_k(\mathbf{w}_t^k)\|^2 \\ &\leq 2L\eta_t^2 \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k^*). \end{aligned} \quad (7)$$

Combining Equations. 3 – 7, we have

$$\begin{aligned}
\|\bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t\|^2 &\leq \|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 2L\eta_t^2 \sum_{k=1}^m p_{k,t}(LM_k(\mathbf{w}_t^k) - LM_k^*) \\
&\quad + \eta_t \sum_{k=1}^m p_{k,t} \left( \frac{1}{\eta_t} \|\bar{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 + \eta_t \|\nabla LM_k(\mathbf{w}_t^k)\|^2 \right) \\
&\quad - 2\eta_t \sum_{k=1}^m p_{k,t} ((LM_k(\mathbf{w}_t^k) - LM_k(\mathbf{w}^*)) + \frac{\mu}{2} \|\mathbf{w}_t^k - \mathbf{w}^*\|^2) \\
&= (1 - \mu\eta_t) \|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \sum_{k=1}^m p_{k,t} \|\bar{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \\
&\quad + 2L\eta_t^2 \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k^*) + \eta_t^2 \sum_{k=1}^m p_{k,t} \|\nabla LM_k(\mathbf{w}_t^k)\|^2 \quad (8) \\
&\quad - 2\eta_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k(\mathbf{w}^*)) \\
&\leq (1 - \mu\eta_t) \|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \sum_{k=1}^m p_{k,t} \|\bar{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 \\
&\quad + 4L\eta_t^2 \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k^*) \\
&\quad - 2\eta_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k(\mathbf{w}^*)),
\end{aligned}$$

where we use the L-smoothness of  $LM_k(\cdot)$  in the last inequality.

We use  $\gamma_t = 2\eta_t(1 - 2L\eta_t)$ , and the last two terms of Equation. 8 are

$$\begin{aligned}
&4L\eta_t^2 \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k^*) - 2\eta_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - LM_k(\mathbf{w}^*)) \\
&= -\gamma_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - GM^*) - \gamma_t \sum_{k=1}^m p_{k,t} (GM^* - LM_k^*) \\
&\quad + 2\eta_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}^*) - LM_k^*) \\
&= -\gamma_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - GM^*) - \gamma_t \sum_{k=1}^m p_{k,t} (GM^* - LM_k^*) \quad (9) \\
&\quad + 2\eta_t \sum_{k=1}^m p_{k,t} (GM^* - LM_k^*) \\
&= -\gamma_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - GM^*) + (2\eta_t - \gamma_t) \sum_{k=1}^m p_{k,t} (GM^* - LM_k^*) \\
&= -\gamma_t \sum_{k=1}^m p_{k,t} (LM_k(\mathbf{w}_t^k) - GM^*) + 4L\eta_t^2 \Gamma,
\end{aligned}$$

where  $\Gamma = \sum_{k=1}^m p_{k,t} (GM^* - LM_k^*) = GM^* - \sum_{k=1}^m p_{k,t} LM_k^*$ .

The first term of Equation. 9

$$\begin{aligned}
& \sum_{k=1}^m p_{k,t}(LM_k(\mathbf{w}_t^k) - GM^*) \\
&= \sum_{k=1}^m p_{k,t}(LM_k(\mathbf{w}_t^k) - LM_k(\bar{\mathbf{w}}_t)) + \sum_{k=1}^m p_{k,t}(LM_k(\bar{\mathbf{w}}_t) - GM^*) \\
&\geq \sum_{k=1}^m p_{k,t} \langle \nabla LM_k(\bar{\mathbf{w}}_t), \mathbf{w}_t^k - \bar{\mathbf{w}}_t \rangle + \sum_{k=1}^m p_{k,t}(LM_k(\bar{\mathbf{w}}_t) - GM^*) \\
&= \sum_{k=1}^m p_{k,t} \langle \nabla LM_k(\bar{\mathbf{w}}_t), \mathbf{w}_t^k - \bar{\mathbf{w}}_t \rangle + GM(\bar{\mathbf{w}}_t) - GM^* \\
&\geq -\frac{1}{2} \sum_{k=1}^m p_{k,t} (\eta_t \|LM_k(\bar{\mathbf{w}}_t)\|^2 + \frac{1}{\eta_t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2) + GM(\bar{\mathbf{w}}_t) - GM^* \\
&\geq -\sum_{k=1}^m p_{k,t} (\eta_t L(LM_k(\bar{\mathbf{w}}_t) - LM_k^*) + \frac{1}{2\eta_t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2) + GM(\bar{\mathbf{w}}_t) - GM^*,
\end{aligned} \tag{10}$$

where the first inequality results from the convexity of  $LM_k(\cdot)$ , the second inequality from AM-GM inequality and the third inequality from L-smoothness of  $LM_k(\cdot)$ .

Therefore, Equation. 9 becomes

$$\begin{aligned}
& -\gamma_t \sum_{k=1}^m p_{k,t}(LM_k(\mathbf{w}_t^k) - GM^*) + 4L\eta_t^2\Gamma \\
&\leq \gamma_t \left( \sum_{k=1}^m p_{k,t} (\eta_t L(LM_k(\bar{\mathbf{w}}_t) - LM_k^*) + \frac{1}{2\eta_t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2) \right) \\
&\quad - \gamma_t (GM(\bar{\mathbf{w}}_t) - GM^*) + 4L\eta_t^2\Gamma \\
&= \gamma_t \left( \sum_{k=1}^m p_{k,t} (\eta_t L(LM_k(\bar{\mathbf{w}}_t) - GM^*) + \frac{1}{2\eta_t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2) \right) \\
&\quad + \gamma_t \eta_t L\Gamma - \gamma_t (GM(\bar{\mathbf{w}}_t) - GM^*) + 4L\eta_t^2\Gamma \\
&= \gamma_t (\eta_t L - 1) \sum_{k=1}^m p_{k,t} (LM_k(\bar{\mathbf{w}}_t) - GM^*) \\
&\quad + \frac{\gamma_t}{2\eta_t} \sum_{k=1}^m p_{k,t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2 + (4L\eta_t^2 + \gamma_t \eta_t L)\Gamma,
\end{aligned} \tag{11}$$

With  $GM(\bar{\mathbf{w}}_t) - GM^* > 0$  and  $\eta_t L - 1 < 0$ , we have

$$\gamma_t (\eta_t L - 1) \sum_{k=1}^m p_{k,t} (LM_k(\bar{\mathbf{w}}_t) - GM^*) \leq 0, \tag{12}$$

and recall  $\gamma_t = 2\eta_t(1 - 2L\eta_t)$ , so  $\frac{\gamma_t}{2\eta_t} \leq 1$  and  $4L\eta_t^2 + \gamma_t \eta_t L \leq 6L\eta_t^2$ .

Therefore,

$$-\gamma_t \sum_{k=1}^m p_{k,t}(LM_k(\mathbf{w}_t^k) - GM^*) + 4L\eta_t^2\Gamma \leq \sum_{k=1}^m p_{k,t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2 + 6L\eta_t^2\Gamma. \tag{13}$$

Thus, Equation. 8 becomes

$$\|\bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t\|^2 \leq (1 - \mu\eta_t) \|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 2 \sum_{k=1}^m p_{k,t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2 + 6L\eta_t^2\Gamma. \tag{14}$$

### A.5 BOUND FOR VARIANCE OF GRADIENTS

Next, to bound the gradient, using assumption 3, we have

$$\begin{aligned}
\mathbb{E}\|g_t - \bar{g}_t\|^2 &= \mathbb{E}\left\|\sum_{k=1}^m p_{k,t}(\nabla LM_k(\mathbf{w}_t^k, \xi_t^k) - \nabla LM_k(\mathbf{w}_t^k))\right\|^2 \\
&= \sum_{k=1}^m p_{k,t}^2 \mathbb{E}\|\nabla LM_k(\mathbf{w}_t^k, \xi_t^k) - \nabla LM_k(\mathbf{w}_t^k)\|^2 \\
&\leq \sum_{k=1}^m p_{k,t}^2 \sigma_k^2.
\end{aligned} \tag{15}$$

#### A.5.1 BOUND FOR DIVERGENCE OF WEIGHTS

Based on Assumption 5, for malicious clients  $k = 1, 2, \dots, c$ , we have

$$\begin{aligned}
p_{k,t} &= \text{softmax}(RS_{km}^t) \\
&= \frac{e^{RS_{km}^t}}{\sum_{i=1}^m RS_i^t} \\
&= \frac{e^{RS_{km}^{t-M} - \delta_m}}{\sum_{i=1}^c e^{RS_i^{t-M} - \delta_m} + \sum_{i=c+1}^m e^{RS_i^{t-M} + \delta_b}} \\
&= \frac{e^{RS_{km}^{t-M}}}{\sum_{i=1}^c e^{RS_i^{t-M}} + \sum_{i=c+1}^m e^{RS_i^{t-M} + \delta_b + \delta_m}} \\
&\leq \frac{e^{RS_{km}^{t-M}}}{\sum_{i=1}^c e^{RS_i^{t-M}} + \sum_{i=c+1}^m e^{RS_i^{t-M}}} \\
&= p_{k,t-M}.
\end{aligned} \tag{16}$$

To bound the weights, we assume within  $E$  communication steps, there exists  $t_0 < t$ , such that  $t - t_0 \leq E - 1$  and  $\mathbf{w}_{t_0}^k = \bar{\mathbf{w}}_{t_0}$  for all  $k = 1, 2, \dots, m$ . And we know  $\eta_t$  is non-increasing and  $\eta_{t_0} \leq 2\eta_t$ . With the fact  $\mathbb{E}\|X - \mathbb{E}X\|^2 \leq \mathbb{E}\|X\|^2$  and Jensen inequality, we have

$$\begin{aligned}
\mathbb{E}\sum_{k=1}^m p_{k,t}\|\bar{\mathbf{w}}_t - \mathbf{w}_t^k\|^2 &\leq \mathbb{E}\sum_{k=1}^m p_{k,t}\|\bar{\mathbf{w}}_{t_0} - \mathbf{w}_{t_0}^k\|^2 \\
&\leq \sum_{k=1}^m p_{k,t}\mathbb{E}\sum_{t_0}^{t-1} (E-1)\eta_t^2\|LM_k(\mathbf{w}_t^k, \xi_t^k)\|^2 \\
&\leq \sum_{k=1}^m p_{k,t}\mathbb{E}\sum_{t_0}^{t-1} (E-1)\eta_{t_0}^2 G^2 \\
&\leq \sum_{k=1}^m p_{k,t}\mathbb{E}(E-1)^2\eta_{t_0}^2 G^2 \\
&= \sum_{k=1}^c p_{k,t}\mathbb{E}(E-1)^2\eta_{t_0}^2 G^2 + \sum_{k=c+1}^m p_{k,t}\mathbb{E}(E-1)^2\eta_{t_0}^2 G^2 \\
&\leq \sum_{k=1}^c p_{k,t}\mathbb{E}(E-1)^2\eta_{t_0}^2 G^2 + 4\eta_t^2(E-1)^2 G^2 \\
&\leq 4\sum_{k=1}^c p_{k,t}\eta_t^2(E-1)^2 G^2 + 4\eta_t^2(E-1)^2 G^2 \\
&\leq 4\sum_{k=1}^c p_{k,0}\eta_t^2(E-1)^2 G^2 + 4\eta_t^2(E-1)^2 G^2,
\end{aligned} \tag{17}$$

where  $p_{k,0}$  is the initial probability of  $k$ th malicious client.

### A.5.2 CONVERGENCE BOUND

Combining Equation.(2)(14)(15)(17), we have

$$\begin{aligned} \|\bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &= \|\bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t\|^2 + 2\eta_t \langle \bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t, \bar{g}_t - g_t \rangle + \eta_t^2 \|\bar{g}_t - g_t\|^2 \\ &\leq (1 - \mu\eta_t) \|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 2 \sum_{k=1}^m p_{k,t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2 + 6L\eta_t^2 \Gamma \\ &\quad + 2\eta_t \langle \bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t, \bar{g}_t - g_t \rangle + \eta_t^2 \|\bar{g}_t - g_t\|^2. \end{aligned} \quad (18)$$

Since  $\mathbb{E}[g_t] = \bar{g}_t$ , Therefore,

$$\begin{aligned} \mathbb{E}\|\bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &\leq (1 - \mu\eta_t) \mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 2\mathbb{E} \sum_{k=1}^m p_{k,t} \|\mathbf{w}_t^k - \bar{\mathbf{w}}_t\|^2 + 6L\eta_t^2 \Gamma \\ &\quad + 2\eta_t \mathbb{E} \langle \bar{\mathbf{w}}_t - \mathbf{w}^* - \eta_t \bar{g}_t, \bar{g}_t - g_t \rangle + \mathbb{E}\eta_t^2 \|\bar{g}_t - g_t\|^2 \\ &\leq (1 - \mu\eta_t) \mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + 8 \sum_{k=1}^c p_{k,0} \eta_t^2 (E-1)^2 G^2 + 8\eta_t^2 (E-1)^2 G^2 \\ &\quad + 6L\eta_t^2 \Gamma + \eta_t^2 \sum_{k=1}^m p_{k,t}^2 \sigma_k^2 \\ &= (1 - \mu\eta_t) \mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 \\ &\quad + \eta_t^2 \left[ 8 \sum_{k=1}^c p_{k,0} (E-1)^2 G^2 + 8(E-1)^2 G^2 + 6L\Gamma + \sum_{k=1}^m p_{k,t}^2 \sigma_k^2 \right] \end{aligned} \quad (19)$$

We set  $\eta_t = \frac{\beta}{t+\gamma}$  for some  $\beta > \frac{1}{\mu}$  and  $\gamma > 0$ , such that  $\eta_1 \leq \min\{\frac{1}{\mu}, \frac{1}{4L}\} = \frac{1}{4L}$  and  $\eta_t \leq 2\eta_{t+E}$ .

We want to prove  $\mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 \leq \frac{v}{\gamma+t}$ , where  $v = \max\{\frac{\beta^2 B}{\beta\mu-1}, (\gamma+1)\mathbb{E}\|\bar{\mathbf{w}}_1 - \mathbf{w}^*\|^2\}$  and  $B = 8 \sum_{k=1}^c p_{k,0} (E-1)^2 G^2 + 8(E-1)^2 G^2 + 6L\Gamma + \sum_{k=1}^m p_{k,t}^2 \sigma_k^2$ .

Firstly, the definition of  $v$  ensures that  $\mathbb{E}\|\bar{\mathbf{w}}_1 - \mathbf{w}^*\|^2 \leq \frac{v}{\gamma+1}$ . Assume the conclusion holds for some  $t$ , we have

$$\begin{aligned} \mathbb{E}\|\bar{\mathbf{w}}_{t+1} - \mathbf{w}^*\|^2 &\leq (1 - \mu\eta_t) \mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 + \eta_t^2 B \\ &\leq \left(1 - \frac{\beta\mu}{t+\gamma}\right) \frac{v}{t+\gamma} + \frac{\beta^2 B}{(t+\gamma)^2} \\ &= \frac{t+\gamma-1}{(t+\gamma)^2} v + \left[\frac{\beta^2 B}{(t+\gamma)^2} - \frac{\beta\mu-1}{(t+\gamma)^2} v\right] \\ &\leq \frac{v}{t+\gamma+1}. \end{aligned} \quad (20)$$

By the  $L$ -smoothness of  $GM(\cdot)$ ,  $\mathbb{E}[GM(\bar{\mathbf{w}}_t)] - GM^* \leq \frac{L}{2} \mathbb{E}\|\bar{\mathbf{w}}_t - \mathbf{w}^*\|^2 \leq \frac{L}{2} \frac{v}{\gamma+t}$ .

Thus we have

$$\mathbb{E}[GM(\mathbf{w}_T)] - GM^* \leq \frac{2}{\mu^2} \cdot \frac{L}{\gamma+T} \left( \sum_{k=1}^m p_k^2 \sigma_k^2 + 6L\Gamma + 8G^2 + 8G^2 \sum_{k=1}^c p_{k,0} + \frac{\mu^2}{4} \|w_0 - w^*\|^2 \right).$$

## REFERENCES

Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Local model poisoning attacks to byzantine-robust federated learning. In *29th USENIX Security Symposium (USENIX Security 20)*, Boston, MA, August 2020. URL <https://www.usenix.org/conference/usenixsecurity20/presentation/fang>.

Luis Muñoz-González, Battista Biggio, Ambra Demontis, Andrea Paudice, Vasin Wongrassamee, Emil C. Lupu, and Fabio Roli. Towards poisoning of deep learning algorithms with back-gradient optimization. *arXiv preprint arXiv:1708.08689*, 2017.