A Theoretical Analysis

In this section, we provide the theoretical guarantee of the convergence of the detection accuracy for the proposed FedPBF.

Assumption A.1. *For any two sets $\mathcal{I}_1$ and $\mathcal{I}_2$ with Q clients, where $\mathcal{I}_1$ is composed of benign clients and $\mathcal{I}_2$ includes at least one Byzantine client, we assume $\sum_{k\in\mathcal{I}_1} F_k(\mathbf{w}_1, \mathcal{D}_k) < \sum_{k\in\mathcal{I}_2} F_k(\mathbf{w}_2, \mathcal{D}_k)$* [8]*., where $\mathbf{w}_1 = \mathbf{w} - \frac{1}{Q}\sum_{k\in\mathcal{I}_1} \mathbf{g}_k$, $\mathbf{w}_2 = \mathbf{w} - \frac{1}{Q}\sum_{k\in\mathcal{I}_2} \mathbf{g}_k$ and $\mathbf{w}$ is the global weight.*

Theorem A.1. *Let K, M, Q and T be the number of clients, the number of candidate models, the number of aggregated updates for each candidate model and the number of global iterations, respectively. If Assumption A.1 holds and $1-(1-(\frac{m}{K})^Q)^M > \frac{m}{K}$, then the detection accuracy in Definition 2 has*

$$\eta_T > \eta_{T-1} > \cdots > \eta_1 = 1-(1-(\frac{m}{K})^Q)^M > \frac{m}{K}. \tag{13}$$

Moreover, we have

$$\lim_{T\to\infty} \eta_T = 1, \tag{14}$$

and averaged detection accuracy as defined in Definition 2

$$\eta \geq 1-(1-(\frac{m}{K})^Q)^M. \tag{15}$$

Proof. Denote $\mathbf{p}_t$ to be the sample probability of all clients and $\mathbf{p}_1 = [1/K, \cdots, 1/K]$. Denote the p_t to be the probability of selecting benign clients and $p_1 = \frac{m}{K}$. Therefore, we have

$$\eta_t = 1-(1-p_t^Q)^m, \tag{16}$$

and $\eta_1 = 1-(1-(\frac{m}{K})^Q)^m$.
Notice that

$$\begin{aligned} p_t &= \eta_{t-1}\eta_{t-2}\cdots\eta_1\frac{m+tQ}{K+tQ} + \sum_{i=1}^{t-1} \Pi_{j\neq i}\eta_j(1-\eta_i)\frac{m+(t-1)Q}{K+tQ} \\ &\quad + \cdots + (1-\eta_{t-1})(1-\eta_{t-2})\cdots(1-\eta_1)\frac{m}{K+tQ} \\ &= \frac{x_t}{K+tQ}, \end{aligned} \tag{17}$$

and

$$x_t = (1-\eta_{t-1})x_{t-1} + \eta_{t-1}x_{t-1} + Q\eta_{t-1}. \tag{18}$$

We would prove the η_t and p_t are two increasing sequence by induction.
Firstly, it is easy to verify $p_2 > p_1$ if $1-(1-(\frac{m}{K})^Q)^M > \frac{m}{K}$ holds. Then since $f(x) = 1-(1-x^Q)^m$ is increasing function, thus $\eta_2 > \eta_1$.

[8] We assume that as long as there is at least one Byzantine client, the model performance must be degraded so that the loss will increase.

Secondly, if $\eta_{t-1} > \eta_{t-2}$ and $p_{t-1} > p_{t-2}$ hold, then according to Eq. (17) and (18), we have:

$$p_t = \frac{x_t}{K+tQ} = \frac{(1-\eta_{t-1})x_{t-1} + \eta_{t-1}x_{t-1} + Q\eta_{t-1}}{(k+tQ)} \tag{19}$$

Therefore, we only need to prove

$$\begin{aligned} & p_t > p_{t-1} \\ \iff & \frac{(1-\eta_{t-1})x_{t-1} + \eta_{t-1}x_{t-1} + Q\eta_{t-1}}{(k+tQ)} > \frac{x_{t-1}}{K+(t-1)Q} \\ \iff & \eta_{t-1}(K+(t-1)Q) > x_{t-1} \\ \iff & \eta_{t-1} > p_{t-2} \end{aligned} \tag{20}$$

It is noted that $f(x) = 1-(1-x^Q)^m < x$ at the range $(0, b)$ and $f(x) = 1-(1-x^Q)^m > x$ at the range $(b, 1)$. Also, $f(m/K) > m/K$ and $p_{t-1} > p_{t-2} > \cdots > m/K$, thus we obtain

$$\eta_{t-1} > p_{t-1} > p_{t-2}. \tag{21}$$

Therefore, we derive $p_t > p_{t-1}$, since $f(x)$ is increasing, $\eta_t > \eta_{t-1}$. Consequently, the conclusion of induction is also true with t. This completes the proof of the η_t and p_t are two increasing sequences.
Since $\eta_t < 1$ and $p_t < 1$, according to the monotone bounded convergence theorem, the sequence $\{\eta_t\}$ and $\{p_t\}$ have the limit η and p respectively. Suppose $\eta < 1$, then we take limits w.r.t t of Eq. (17) for both sides and obtain

$$p = 0, \tag{22}$$

which is a contradictory of $p > p_1 = m/K$. Noted that $\eta = f(p)$, thus $p = 1$.
Finally, we derive

$$\eta_t > \eta_{t-1} > \cdots > \eta_1 = 1-(1-(\frac{m}{K})^Q)^M > \frac{m}{K} \tag{23}$$

Moreover, we have

$$\lim_{t\to\infty} \eta_t = 1, \eta \geq 1-(1-(\frac{m}{K})^Q)^M. \tag{24}$$

Theorem A.1 demonstrates 1) the detection accuracy increases with the training iteration and converges to one, which reflects that our algorithm selects the benign updates more accurately during training and achieves the 100% accuracy finally, so the Byzantine Fault-tolerance in Definition 1 is guaranteed ; 2) the detection accuracy has the lower bound $\bar{l} = 1-(1-(\frac{m}{K})^Q)^M$, which is influenced by M, Q and $\frac{m}{K}$ (experimental results in Appendix B.2 also elucidate this phenomenon). Specifically, we have

- the lower bound $\bar{l}$ is increasing w.r.t $\frac{m}{K}$ and $\lim_{\frac{m}{K}\to 1} \bar{l} = 1$, which is explained $\bar{l}$ improves as there are more benign clients.

– the lower bound $\bar{l}$ is increasing w.r.t M and $\lim_{M\to\infty} \bar{l} = 1$. It is reasonable that the $\bar{l}$ becomes larger when there are more candidate models to evaluate for the server.
– the lower bound $\bar{l}$ is decreasing w.r.t Q and $\lim_{Q\to\infty} \bar{l} = 0$. When the number of aggregated updates for each candidate model increases, the probability of the presence of Byzantine attackers for the candidate model is increasing, making it hard to choose the sets without any malicious updates. Therefore, the convergence of our proposed method is influenced.

B Experiments

B.1 Robustness under Misreporting by Byzantine Clients

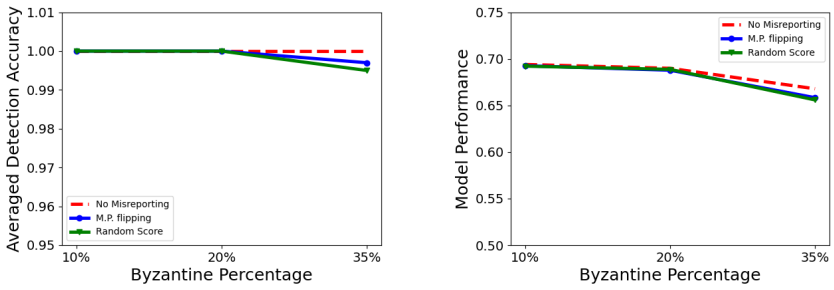


Fig. 3: Averaged Detection accuracy (left) and model performance (right) of FedPBF with different Byzantine client percentages (10%, 20%, 35%) with misreporting under sign flipping attack (with IID setting for classification of CIFAR10).

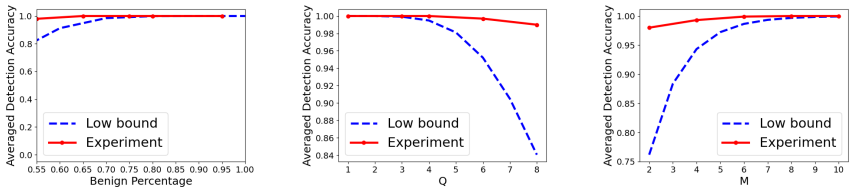
To confuse the judgment of the server, Byzantine clients may intentionally misreport model performance of candidate models for them to evaluate. Therefore, we adopt the median to filter out the misreported model performance by Byzantine clients as Eq. (10). We further evaluate the model performance on the CIFAR10 dataset with different Byzantine client percentages (10%, 20% and 35%) under three misreporting conditions: 1) **Mode Performance flipping (M.P. flipping):** Byzantine clients flip the model performances of candidate models, i.e., exchanging the highest j_{th} of the candidate model with the lowest j_{th} score of the candidate model for $j = 1, \cdots, \lfloor \frac{M}{2} \rfloor$. 2) **Random Score:** Byzantine clients randomly select a model performance value from 0 to 1. 3) **No Misreporting:** the Byzantine clients report the model performance honesty. Fig. 3 shows the model performance and averaged detection accuracy of FedPBF with Byzantine clients with different kinds of misreporting. It shows that the FedPBF is robust

against misreporting, i.e. there is no performance degradation when Byzantine client percentage equals 10% and 20%. Even if the Byzantine client percentage reaches 35%, the averaged detection accuracy and model performance only drop less than 1%.

Hyper-parameter	Logistic Regression	LeNet	AlexNet
Number of Clients	100	100	20
Optimization method	SGD	Adam	SGD
Learning rate	0.125	0.0125	0.01
Weight Decay	0.0	0.002	0.001
Batch size	32	32	32
Data Distribution	Non-IID ($\beta = 0.5$)	Non-IID ($\beta = 0.5$)	IID and Non-IID ($\beta = 0.5$,1)
Local rounds	1 iteration	3 epochs	3 epochs
Communication rounds	3000	3000	600
M and Q	$M = 40, Q = 10$	$M = 40, Q = 10$	$M = 10, Q = 3$

Table 3: Hyper-parameters for training.

B.2 Ablation Study



(a) M=10 Q=3 (b) M=10 $\frac{m}{K} = 80\%$ (c) Q=3 $\frac{m}{K} = 80\%$

Fig. 4: Low bound (blue dotted line) and experiment (red solid line) averaged detection accuracy of FedPBF with different hyperparameters under sign flipping attack (with IID setting for classification of CIFAR10)

.

To answer **Question 3**, we use CIFAR10 on AlexNet under the IID setting to test the reliability of the theoretical averaged detection accuracy lower bound in Sect. A. Fig. 4 (a) shows how the averaged detection accuracy is affected by $\frac{m}{K}$ for a fixed $Q = 3$ and $M = 10$. Fig. 4 (b) shows the changing trend of averaged detection accuracy when the Q value is changed at a fixed $M = 10$ and $\frac{m}{K} = 80\%$. Fig. 4 (c) shows the changing trend of averaged detection accuracy when the

value of M is changed at a fixed $Q = 3$ and $\frac{m}{K} = 80\%$. The averaged detection accuracy of the experiment is consistently larger than the theoretical lower bound (blue dotted line), which means one can select appropriate hyperparameters M, Q to ensure required model performances. Therefore, we set $Q = 10$, $M = 40$ in experiments in Tab. 2 unless stated otherwise.