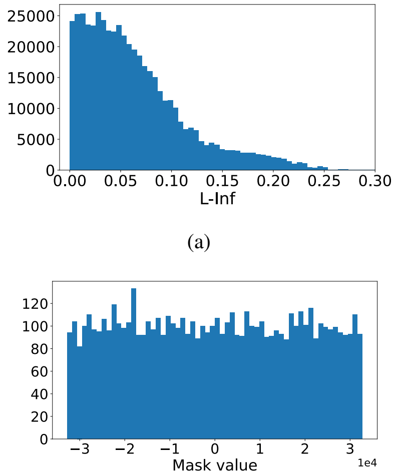
A. Malicious Seed Modification

Fig. 7(a) demonstrates the histogram of L_∞ norm of *benign* gradients, observed throughout training across 100 users for CIFAR-10 dataset. As seen, majority of the gradient norms are bounded in $[0, 0.25]$. Masks are generated using seeds through a pseudorandom generator (PRG), and malicious users cannot control the resulting error when changing the seed. Fig. 7(b) shows a histogram of mask values generated from random seeds over 10000 runs. As shown, changing the seed may cause unpredictable and drastic changes in the mask. By changing the random seed, the generated masks can vary anywhere between -3×10^4 to 3×10^4, which is much larger than the normal observed range for model updates. As such, when a malicious user changes the random seed from which the masks are generated, it can lead to easily recognizable errors in the gradient that raises alarms for the server. Thus, in our threat model malicious users are incentivized to use the correct seed when computing masks.



(a)

(b)

Figure 7: Histogram of (a) ResNet-20 gradient norms observed during training on CIFAR-10, and (b) mask values when changing the random seed.

B. Effect of Aggregation Method on Accuracy

zPROBE leverages the median of averaged model updates across user clusters to check whether the incoming updates are benign or Byzantine. An alternative aggregation strategy is to directly use the median of cluster means, rather than performing the subsequent per-user checks. Fig. 8 shows the test accuracy of zPROBE as training progresses, when compared to the above baseline aggregation method that applies the coordinate-wise median of cluster means. As seen, compared to zPROBE, this baseline suffers from a large accuracy degradation, since all information in benign user updates is lost by replacing the aggregation with the median, which can potentially contain Byzantine workers.

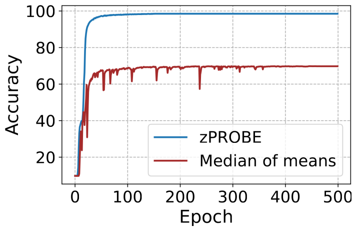
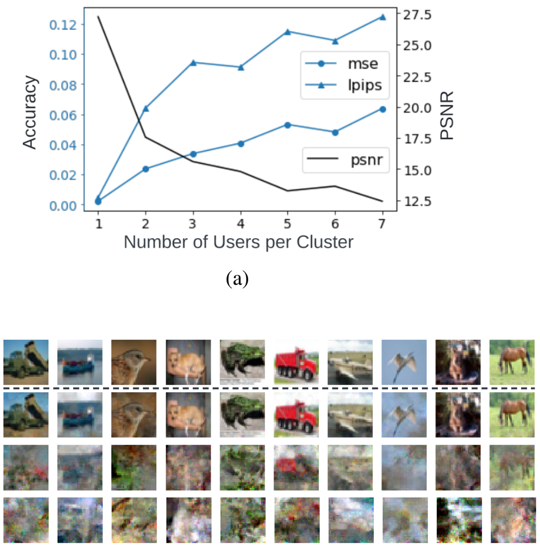


Figure 8: Test accuracy of zPROBE compared with an aggregation methodology that uses the median of cluster means.

C. Effect of Cluster Size on Inversion Attack

Fig. 9 shows the effect of cluster size on gradient inversion attacks. In Fig 9(a) we show the effectiveness of the attack [19] for different cluster sizes. Fig 9(b) represents the reconstruction results from user data for different number of users participating in the aggregation round.



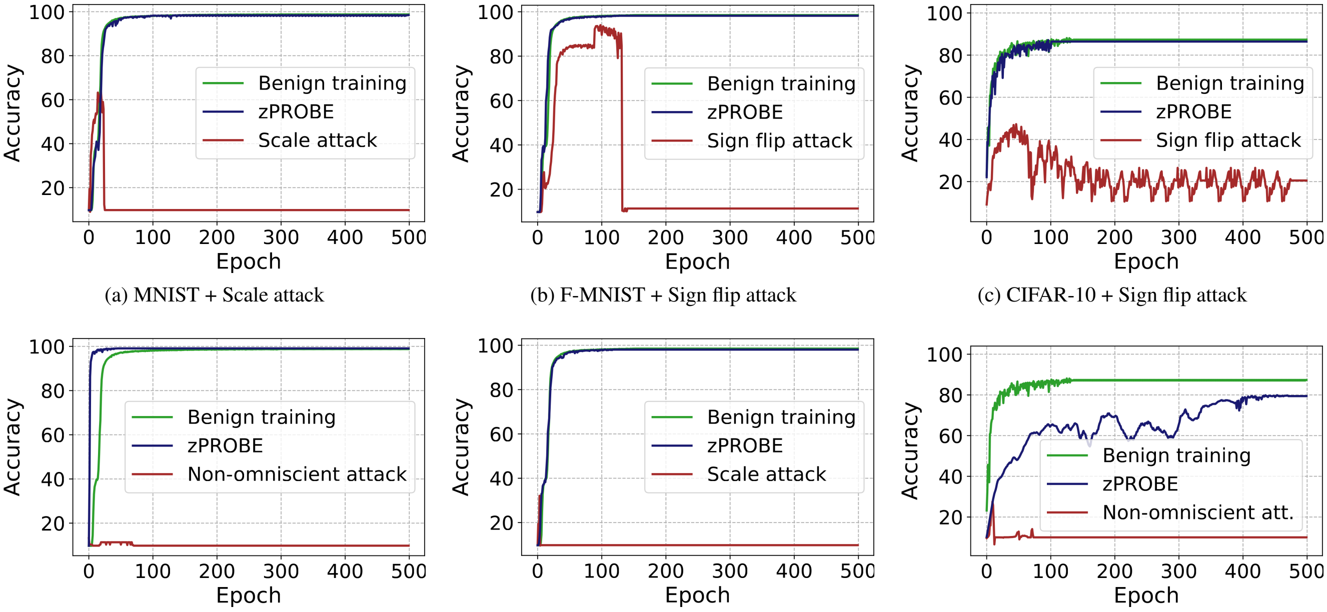
(a)

(b)

Figure 9: Performance of gradient inversion attacks for different cluster sizes.

D. zPROBE Test Accuracy

Fig. 10 shows the test accuracy of zPROBE in face of different variations of Byzantine attacks and datasets. The dataset is distributed evenly (IID) among $n = 50$ clients. The server randomly clusters users into $c = 7$ groups during each training round. We assume malicious users compromise all model updates $|\mathcal{S}_m| = 1$ to maximize the accuracy degradation.



(a) MNIST + Scale attack

(b) F-MNIST + Sign flip attack

(c) CIFAR-10 + Sign flip attack

(d) MNIST + Non-omniscient attack

(e) F-MNIST + Scale attack

(f) CIFAR-10 + Non-omniscient

Figure 10: Test accuracy as a function of FL training epochs for different attacks and benchmarks. Each plot shows the benign training (green), Byzantine training without defense (maroon), and Byzantine training in the presence of `zPROBE` defense.