Stress-Testing Byzantine Defenses under Data Heterogeneity

Distributed Learning is becoming increasingly adopted as it allows for the collaborative training of large-scale ML tasks on distributed datasets. However, it remains vulnerable to poisoning (Byzantine) attacks, especially when data across clients is non-independent and identically distributed (non-IID). In such settings, benign clients naturally produce diverse gradients, which makes detecting malicious updates more difficult.

While robust aggregation methods have been widely studied, the threat model under non-IID data is comparatively underexplored. This gap may give a false sense of security about current defenses. Notably, the only strong attacks explicitly designed for heterogeneous settings, Min-Max and Min-Sum [4], construct adversarial gradients that blend with honest updates while remaining farther than the most outlying benign client. These attacks exploit natural gradient variance but require full access to honest gradients and solve optimization problems at every iteration, making them impractical. In contrast, ALIE [2] and IPM [5] do not require access to honest clients' gradients, relying only on local information and general assumptions about gradient distributions. This makes them more practical and widely applicable in real-world settings.

In this work, we revisit classic IID-based Byzantine attacks, such as ALIE and IPM, and show that their lack of effectiveness in prior evaluations stems from not being calibrated for non-IID data. These attacks were typically run with conservative perturbation strengths suited for IID scenarios, where gradient similarity forces adversaries to be stealthy. However, in heterogeneous settings, we find that stronger perturbations can go undetected. Our study shows: (1) Under data heterogeneity, both ALIE and IPM can apply significantly stronger perturbations than typically assumed (see Figure 1). (2) Once calibrated, these attacks, especially ALIE, cause severe degradation in test accuracy (e.g., below 24% on CIFAR-10), outperforming even Min-Max and Min-Sum (see table 1). These results demonstrate that state-of-the-art defenses [1, 4, 3] can fail dramatically under realistic adversarial conditions. Our findings highlight the need for threat models and evaluation protocols that better reflect real-world heterogeneity in distributed learning.

Table 1: Top-1 Test Accuracy (%) (mean±std averaged for 3 runs) of different defenses trained for $T = 8000$ under various combinations of β on CIFAR10 for $b = 3$ Byzantine clients out of $n = 17$.

	CIFAR10 ($\beta = 0.3$)				CIFAR10 ($\beta = 0.5$)			
Attack	RFA(buck)	CMLS	CCLIP	trMean(NNM)	RFA(buck)	CMLS	CCLIP	trMean(NNM)
ALIE ($z = 8$)	**23.26 ± 1.39**	**21.71 ± 2.81**	**17.89 ± 5.64**	**21.93 ± 1.49**	**21.29 ± 2.15**	**19.01 ± 0.47**	**24.07 ± 1.43**	**18.90 ± 6.39**
IPM ($\epsilon = 2.5$)	39.30 ± 0.35	44.20 ± 1.16	50.95 ± 1.53	52.29 ± 0.22	43.80 ± 1.39	50.53 ± 0.18	51.01 ± 0.90	56.55 ± 0.74
MinMax	37.99 ± 1.34	42.96 ± 0.68	37.81 ± 0.95	49.60 ± 0.87	44.26 ± 0.69	48.92 ± 2.36	35.33 ± 2.28	53.84 ± 0.07
MinSum	46.00 ± 0.50	43.36 ± 0.89	53.33 ± 0.41	52.67 ± 0.50	51.57 ± 1.14	49.05 ± 0.30	55.58 ± 0.89	59.78 ± 1.36
SF	52.76 ± 1.03	44.12 ± 0.43	63.58 ± 0.47	51.58 ± 2.03	53.37 ± 0.42	50.07 ± 2.24	64.62 ± 1.04	23.93 ± 9.05

References

[1] Allouah et al. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *AISTATS*, 2023.

[2] Baruch et al. A little is enough: Circumventing defenses for distributed learning. In *NeurIPS*, 2019.

[3] Karimireddy et al. Byzantine-robust learning on heterogeneous datasets via bucketing. In *ICLR*, 2022.

[4] Shejwalkar et al. Manipulating the byzantine. In *NDSS*, 2021.

[5] Xie et al. Fall of empires: Breaking byzantine-tolerant SGD by inner product manipulation. In *UAI*, 2020.

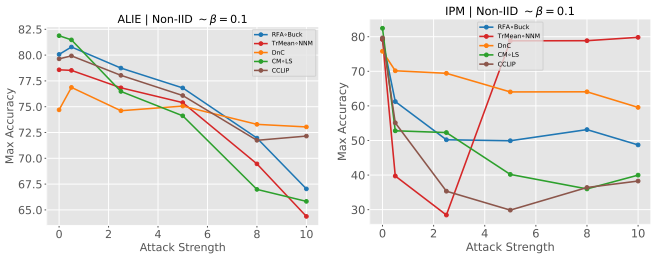

Figure 1: Maximum achieved Test Accuracy for all studied aggregations on varying non-IID data splits under $\delta = 20\%$ Byzantine performing **FMNIST** for varying attack strength.